USING CONTOUR INFORMATION AND SEGMENTATION FOR OBJECT REGISTRATION, MODELING AND RETRIEVAL

by

Tomasz Adamek, M.Sc.

A thesis submitted in partial fulfilment of the requirements
for the degree of Ph.D in Electronic Engineering

Supervisor: Dr. Noel E. O'Connor

School of Electronic Engineering
Dublin City University
June, 2006

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D in Electronic Engineering is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Adamek ID No.: 51161389

Tomasz Adamek

Date: 25/09/2006

ACKNOWLEDGEMENTS

I wish to extend my heartfelt gratitude to the many people who have enabled and supported the work behind this thesis. Special thanks goes to my supervisor, Dr. Noel E. O'Connor, for his valuable guidance and expertise as well as for his kind advice, encouragement, and constant support. Similar thanks are due to Dr. Noel Murphy and Professor Alan Smeaton for their individual advice, support and suggestions at different stages of the work.

I would like to thank all the colleagues at the Centre for Digital Video Processing for the pleasant work atmosphere. I'm indebted to many members of the group for the interesting and fruitful discussions, especially fellow PhD students Orla Duffner, Sorin Sav, and Ciarán Ó Conaire, and also Dr. Hervé Le Borgne.

I would like to thank Christan Ferran Bennström and Dr. Josep R. Casas from the Universitat Politécnica de Catalunya for the interesting discussions which motivated many experiments presented in this thesis. Christan Ferran Bennström devoted significant amount of his time to the creation of the image collection used in this thesis.

Different parts of the research work published in this thesis were financially supported by the Informatics Research Initiative of Enterprise Ireland, by Science Foundation Ireland under grant 03/IN.3/I361., and by the European Commission under contract FP6-001765 aceMedia.

Dr. Miroslaw Bober of the Visual Information Laboratory of Mitsubishi Electric Information Technology Center Europe is acknowledged for generating and providing the CSS results. The Centre for Vision, Speech, and Signal Processing of the University of Surrey is acknowledged for making the database of shapes of marine creatures available, and the Center for Intelligent Information Retrieval from University of Massachusetts is acknowledged for providing the annotated collection of George Washington's manuscripts.

On a personal note, special thanks go to all my friends, especially to the Polish community at Dublin City University.

I wish to extend my gratitude to my parents Henryka and Julian (*Za życie i wychowanie.*) and my aunt Alicja and uncle Bolesław (*Za zawsze serdeczne goszczenie mnie pod swoim dachem.*).

Finally, I would like to thank Ms. Cristina Blanco (*Por su constante apoyo y ánimo, y por recordarme que hay cosas más importantes en esta vida que el ser llamado 'Doctor'. También por proporcionar fotos de sí misma, las cuales mejoran significativamente el valor estético de esta tesis.*).

ABSTRACT

USING CONTOUR INFORMATION AND SEGMENTATION FOR OBJECT REGISTRATION, MODELING AND RETRIEVAL

by Tomasz Adamek M.Sc.

This thesis considers different aspects of the utilization of contour information and syntactic and semantic image segmentation for object registration, modeling and retrieval in the context of content-based indexing and retrieval in large collections of images. Target applications include retrieval in collections of closed silhouettes, holistic word recognition in handwritten historical manuscripts and shape registration. Also, the thesis explores the feasibility of contour-based syntactic features for improving the correspondence of the output of bottom-up segmentation to semantic objects present in the scene and discusses the feasibility of different strategies for image analysis utilizing contour information, e.g. segmentation driven by visual features versus segmentation driven by shape models or semi-automatic in selected application scenarios.

There are three contributions in this thesis. The first contribution considers structure analysis based on the shape and spatial configuration of image regions (so-called syntactic visual features) and their utilization for automatic image segmentation. The second contribution is the study of novel shape features, matching algorithms and similarity measures. Various applications of the proposed solutions are presented throughout the thesis providing the basis for the third contribution which is a discussion of the feasibility of different recognition strategies utilizing contour information. In each case, the performance and generality of the proposed approach has been analyzed based on extensive rigorous experimentation using as large as possible test collections.

LIST OF PUBLICATIONS

T. Adamek, N. E. O'Connor, and A.F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents" in *accepted for publication in Special Issue on Analysis of Historical Documents, Int'l Journal on Document Analysis and Recognition, 2006*

T. Adamek and N. E. O'Connor, "Interactive object contour extraction for shape modeling" in *Proc. 1st Int'l Workshop on Shapes and Semantics, Tokyo*, June 2006.

N. E. O'Connor, E. Cooke, H. Le Borgne, M. Blighe, and T. Adamek, "The AceToolbox: low-level audiovisual feature extraction for retrieval and classification" in *Proc. 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, U.K.*, Dec. 2005.

A. F. Smeaton, H. Le Borgne, N. E. O'Connor, T. Adamek, O. Smyth, and S. De Burca, "Coherent segmentation of video into syntactic regions," in *Proc. 9th Irish Machine Vision and Image Processing Conf. (IMVIP'05), Belfast, Northern Ireland*, Aug. 2005.

T. Adamek, N. E. O'Connor, G. Jones, and N. Murphy, "An integrated approach for object shape registration and modeling," in *Proc. 28th Annual Int'l ACM SIGIR Conference (SIGIR'05), Workshop on Multimedia Information Retrieval, Salvador, Brazil*, Aug. 2005.

T. Adamek, N. E. O'Connor, and N. Murphy, "Multi-scale representation and optimal matching of non-rigid shapes," in *Proc. 4th Int'l Workshop on Content-Based Multimedia Indexing, CBMI'05, Riga, Latvia*, June 2005.

T. Adamek, N. E. O'Connor, and N. Murphy, "Region-based segmentation of images using syntactic visual features," in *Proc. 6th Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05), Montreux, Switzerland*, Apr. 2005.

T. Adamek and N. E. O'Connor, "A multi-scale representation method for non-rigid shapes with a single closed contour," *IEEE Trans. Circuits Syst. Video Technol., special issue on Audio and Video Analysis for Multimedia Interactive Services*, no. 5, May 2004.

T. Adamek and N. E. O'Connor, "Efficient contour-based shape representation and matching," in *Proc. 5th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval (MIR'03), Berkeley, CA*, Nov. 2003.

N. E. O'Connor, S. Sav, T. Adamek, V. Mezaris, I. Kompatsiaris, T. Y. Lui, E. Izquierdo, C. F. Bennstrom, and J. R. Casas, "Region and object segmentation algorithms in the QIMERA segmentation platform" in *Proc. Int'l Workshop on Content-Based Multimedia Indexing (CBMI'03), Rennes, France*, Sep. 2003.

N. E. O'Connor, T. Adamek, S. Sav, N. Murphy, and S. Marlow, "QIMERA: A software platform for video object segmentation and tracking," in *Proc. 4th Workshop on Image Analysis for Multimedia Interactive Services, London, U.K. (WIAMIS'03)*, Apr. 2003.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

THIS chapter briefly discusses the main motivations for the research carried out by the author and reported in this thesis. Also, it briefly outlines the main objectives of the work together with the author's contributions and an overview of the structure of the thesis.

1.1 The New "Digital Age"

In recent years, computer technology has transformed almost every aspect of our life and culture. The advent and popularization of the personal computer in conjunction with the continuous evolution of digital networks and the emergence of multimedia technology, particulary the proliferation of cheap digital media acquisition tools, have enabled a worldwide trend towards a new "digital age". The new models of content production, distribution and consumption have resulted in fast growth of the amount of digital material available.

However, the wealth of information available nowadays has also induced a major problem of information overload. In other words, the new "digital age" challenges us to develop the ability to find helpful and useful information in a vast sea of otherwise useless information. The need for efficient tools to organize, manipulate, search, filter and browse through the huge amounts of digital information is evident from the spectacular success of text-based *Web Search Engines* such as Yahoo or Google [1].

1.2 Challenges Brought by the Explosion of Cheap Digital Media

The emergence of multimedia technology, particulary the explosion of cheap digital media acquisition tools such as scanners and digital cameras etc. as well

as rapid reduction of storage cost, has resulted in accelerated growth of digital audiovisual media collections, both proprietary and freely available on the *World Wide Web*. Applications like manufacturing, medicine, entertainment, education, etc. all make use of immense amounts of audiovisual data. As the amount of information available in visual form continuously increases, the necessity for the development of human-centered tools for the effective processing, storing, managing, browsing and retrieval of audiovisual data becomes evident.

A large amount of visual material is available in the form of still images, either as a photograph representing a real scene or as a graphic containing a non-real image (sketched by a human or synthesized by computer). Images can be stored either in local databases or distributed databases (e.g. the *World Wide Web*) and either embedded in documents or available as stand-alone objects.

The application areas most often listed in the literature where the use, and retrieval in particular, of images nowadays plays a crucial role are: law and crime prevention, medicine, fashion and graphic design, publishing, electronic commerce, architectural and engineering design, and historical research. A detailed study of image users and the possible uses of images was provided by Eakins and Graham in [2]. Due to the volume of visual material in the form of images there is a clear need for image search engines either for private or professional use.

1.3 Retrieval by Content

Currently, techniques for image retrieval (or visual media generally) are one of the most eagerly sought multimedia technologies, with potentially great commercial significance. However visual information systems are radically different from conventional information systems and retrieval of data in such systems requires addressing many novel issues before they can truly become a new fertile area for innovative products.

Specifically, retrieving based on the text usually associated with visual content is no longer sufficient. In fact, in many cases this is not even possible. Often there is no textual information associated with images and manual keyword annotation is labor intensive. Therefore, recently *retrieval by content* has been suggested as an alternative retrieval model for audiovisual media. In *Content-Based Image Retrieval* (CBIR) systems images are indexed (and subsequently browsed and retrieved) based on features extracted directly from their representation rather than by any associated text.

Building modern systems for content-based retrieval of visual data requires consideration of several key issues such as system design, feature extraction, similarity measures, indexing structures, semantic analysis, inferring content

abstraction from low-level features, designing user interfaces, querying models and relevance feedback [3].

Notably, CBIR systems require indices to order the data. However, given a visual signal there is little information readily available for indexing. Therefore in recent years a significant research effort has been dedicated to the problem of extraction of visual descriptors which could be then exploited later for content-based browsing and retrieval. The descriptors may be visual features such as colour, texture, edge direction, shape, scene layout, spatial relationships, or semantic primitives. Also they can be global (one feature describes the whole image) or local (each feature describes an object or a part of the image). Another crucial component of CBIR systems is establishing similarity between visual entities (e.g. between the query and the target).

In order to allow inter-operability between devices and applications attempting to solve various aspects of the query by content problem as well as to provide a common terminology and test material, the *Motion Picture Experts Group* (MPEG) introduced a standard known as MPEG-7 [4, 5]. MPEG-7, formally called the *Multimedia Content Description Interface*, is a standard for describing multimedia content, facilitating sophisticated management, browsing, searching, indexing, filtering, and accessing of that content. Unlike previous ISO standards, it is intended to provide representations of audiovisual (AV) information that aim beyond compression (such as MPEG-1 and MPEG-2 standards) or even object-based functionalities (such as supported by the MPEG-4 standard), and support some degree of interpretation of the AV content. MPEG-7 offers a comprehensive set of standardized audiovisual description tools by specifying four types of normative elements: Descriptors, Description Schemes, a Description Definition Language (DDL), and coding schemes. For a more detailed description of the standard the reader can refer to [4, 5].

This thesis contributes to two key research areas enabling some of the above challenges to be addressed: image segmentation and shape matching. The remainder of this chapter briefly discusses the main motivation for the research carried out by the author in this thesis by discussing the role of image segmentation and shape matching in CBIR. This is followed by an outline of the main objectives of the work together with the author's contributions and an overview of the structure of the thesis.

1.4 The Importance of Image Segmentation and Shape Analysis in CBIR

This section discuses briefly the role of image segmentation and shape matching technologies in addressing some of the major challenges which must be faced

when building CBIR systems.

1.4.1 Image Segmentation

Image segmentation refers to a very broad set of techniques for image partitioning. Typically, the goal is to partition an image using a chosen criterion so that the created partition reflects the structure of the scene at a level suitable for a given application, e.g. homogenous regions or semantic objects. In other words, the term segmentation can be used to refer to any process allowing discovering certain knowledge about the structure of the scene, e.g. shape of the objects present. Therefore, the problem of partitioning an image into a set of homogenous regions or semantic entities is a fundamental enabling technology for understanding scene structure and identifying relevant objects. Identification of areas of an image that correspond to important regions, e.g. semantic objects, is often considered to be the first step of many object-based applications.

The problem of segmentation is fundamental to many challenges encountered in the area of computer vision. This thesis focuses primarily on region and object based segmentation of images in the context of visual content indexing and retrieval applications.

Many of the low-level descriptors, including the ones defined by the MPEG-7 standard, can be used to describe entire images as well as parts of images (e.g. image regions, which ideally would correspond to real objects present in the scene). Utilization of local descriptors, representing images at region or object granularity, allows indexing and retrieval in a manner closer to human perception and scene understanding. Image segmentation is a vital tool for extraction of such localized descriptors of the image content. In other words, the ability of partitioning the image into regions corresponding to objects present in the scene (or at least to parts of the object homogenous according to certain criteria) is central for extraction of local low-level features (e.g. colour, texture, shape) where each feature partially or completely describes an object. However, MPEG-7 does not specify normative tools for image segmentation as this is not necessary for inter-operability. Nevertheless, in practice, image segmentation tools are integral parts of many feature extraction toolboxes – see for example [6]. The author believes that in many cases, the success of MPEG-7, both, commercially and as a tool facilitating research, will depend on the ability to easily produce partitioning of the visual content into regions meaningful in a given application scenario.

It is commonly known that even limited capabilities of filtering of the AV material based on small sets of particulary important semantic objects/concepts can greatly improve performance of CBIR systems, e.g. the user may wish to browse only images containing certain type of objects (e.g. faces) or scenes (e.g.

indoor/outdor) – see for example the Físchlár system [7]. Often, such functionalities require automatic detection, extraction and recognition of semantically meaningful entities from visual data.

Recently it has been shown that textual labels can be assigned (inferred) to homogenous image regions automatically during an off-line training/indexing process [8, 9]. Such labels can be then used for indexing and subsequently for textual querying. Clearly, the ability of automatically segment images based on colour and texture homogeneity criteria is a crucial underpinning technology in such a scenario.

Finally, semi-automatic (often also referred to as supervised) segmentation technologies are becoming sufficiently mature for integration with CBIR opening an interesting possibility of object-based rather than image-based queries. Surprisingly, to the best of authors knowledge so far only a few studies have considered such a possibility [10]. Also, the semi-automatic tools enable the possibility of rapid semi-automatic annotation of images where labels are manually assigned to image parts [11]. Other interesting approaches to CBIR utilizing image segmentation are discussed in the next chapter (section 2.2).

1.4.2 Shape Matching

Shape plays an important role to allow humans to recognize and classify objects. In fact, often shapes represent the abstractions (archetypes) of objects belonging to the same semantic class. Not surprisingly, some user surveys regarding cognition aspects of image retrieval indicate that users are more interested in retrieval by shape than by colour and texture [12]. Therefore shape analysis can play an important role in addressing some of the key problems which must be faced when building CBIR systems such as detection of semantic objects in unseen images, establishing similarity between objects, and extracting indices to index the data.

Shape matching (or in this case often referred to as template matching) plays an important role in the detection and segmentation of semantic objects in unseen images. Once prototypical objects are extracted, shape analysis can make explicit some important information about the object's shape which can be then exploited to recognize that object or distinguish it from other objects. Estimation of similarity between shapes of objects (e.g. between the query and the target objects) is a key technology enabling querying by example. However, estimation of shape similarity can be indirectly important also for textual querying of images, e.g. text recognition in scanned documents by using *Optical Character Recognition* (OCR). Clearly the shape of characters and words is crucial for mapping them into text. One such application is discussed in this thesis in chapter 6 which considers contour matching for word recognition in historically

significant handwritten documents.

Not surprisingly, among several low-level visual features to characterize the AV content MPEG-7 defines two standardized descriptors specifically for 2D shape which are: *contour-based* descriptor relying on the concept of *Curvature Scale Space* (CSS) [13, 14, 15, 16, 17] and *region-based* descriptor based on a 2D complex transform defined with polar coordinates on the unit disk, called *Angular Radial Transform* (ART) [17] – both are discussed in detail in chapter 4. To facilitate the comparison between different methods for shape-based retrieval, a complete evaluation methodology and common test collections were specified. This methodology, among others, is adopted for extensive evaluation of the method proposed by the author in chapter 5. Also, weaknesses of the standardized contour-based descriptor (CSS) are identified and discussed based on several application examples. In particular, it will be shown than the method for estimating similarity between shapes proposed in this thesis produces better results that the CSS-based method in two applications: retrieval in collections of closed silhouettes and holistic word recognition in handwritten historical manuscripts.

Summarizing, image segmentation as well as shape matching are key technologies essential to provide the content-based functionalities required by future multimedia applications. Therefore not surprisingly they have been hot research topics for a large number of groups around the world. Contributions to both of these key research areas are made and reported in this thesis together with a discussion of several applications of the proposed solutions.

1.5 Objectives of this Thesis

This thesis considers different aspects of the utilization of contour information for image segmentation, similarity estimation between objects, and also object registration and modeling in the context of content-based retrieval in large collections of images.

The first objective of this thesis is to place in context the research and provide justification for the investigation carried out by the author. This is achieved by presenting various applications of the proposed solutions throughout the thesis and demonstrating that image segmentation and shape matching are key technologies enabling the content-based functionalities required by future multimedia applications.

The second objective is to outline the current state of the art in image segmentation and shape matching by reviewing the most characteristic categories of techniques in both areas and briefly discussing the most interesting and promising

techniques from each category. This thesis does not attempt to cover all aspects of image segmentation and shape analysis nor does it represent a detailed literature review of the vast amount of work published in these fields. Rather, it restricts itself to a discussion of these technologies in the context of CBIR.

The third, and most important objective is to investigate new approaches to solving the problems related to image segmentation, contour matching and similarity estimation and to discuss their practicability in various applications. One of the main goals is to explore the feasibility of utilizing the spatial configuration of image regions and structural analysis of their contours (so-called "syntactic visual features") for improving the usefulness of the output of bottom-up image segmentation, e.g. for feature extraction, by creating a more meaningful segmentation of the image corresponding more closely to semantic objects present in the scene and improving the perceptual quality of the segmentation. This can be achieved by investigating new ways of utilizing evidence from multiple sources of information, each with its own accuracy and reliability. Another goal is to investigate selected issues of shape analysis, in particular contour representation, matching, and estimation of similarity between silhouettes. The aim is also to demonstrate the usefulness of the proposed approach in a variety of applications.

In each case, the performance and generality of the proposed techniques must be analyzed based on rigorous experimentation using large test collections. Moreover, presentation of selected applications of the proposed solutions should provide a suitable basis for a discussion of the practicability of different recognition strategies utilizing contour information, e.g. bottom-up (automatic segmentation followed by recognition) vs. top-down (segmentation driven by shape models or semi-automatic).

The final objective of this thesis is to indicate directions for further research. Namely, to consider possibilities for further improvement of the proposed solutions and discuss the prospects of using them as a basis for addressing various other challenges in the field.

1.6 The Main Contributions

There are three main contributions in this thesis. The first contribution is a feasibility study of utilizing spatial configuration of regions and structural analysis of their contours (so-called *syntactic visual features* [18]) for improving the correspondence of fully automatic image segmentation to semantic objects present in the scene and for improving the perceptual quality of the segmentation. Several extensions to the well-known *Recursive Shortest Spanning Tree* (RSST) algorithm [19] are proposed among which the most important is a novel framework for the integration of evidence from multiple sources of information, each with

its own accuracy and reliability. The new integration framework is based on the *Theory of Belief* (BeT) [20, 21, 22, 23]. The "strength" of the evidence provided by the geometric properties of regions and their spatial configurations is assessed and compared with the evidence provided solely by the colour homogeneity criterion. Other contributions in this area include a new colour model and colour homogeneity criteria, practical solutions to structure analysis based on shape and spatial configuration of image regions, and a new simple stopping criterion aimed at producing partitions containing the most salient objects present in the scene.

Additionally, the application of the automatic segmentation method to the problem of semi-automatic segmentation is also discussed. An easy to use and intuitive semi-automatic segmentation tool utilizing the automatic approach is described in detail. This example demonstrates that syntactic features can be useful also in the case of supervised scenarios where they can facilitate more intuitive user interactions e.g. by allowing the segmentation process make more "intelligent" decisions whenever the information provided by the user is not sufficient (or ambiguous).

The second contribution is the proposal of a novel rich multi-scale representation for non-rigid shapes with a single closed contour and a study of its properties in the context of efficient silhouette matching and similarity estimation. An initial approach for optimization of the matching in order to achieve higher performance in discriminating between shape classes is also proposed. The efficiency of the proposed silhouette matching approach is demonstrated in a variety of applications.

In particular, the retrieval results are compared to those of the Curvature Scale Space (CSS) [13, 14, 15] approach (adopted by the MPEG-7 standard [16, 17]) and it is shown that the proposed scheme performs better that CSS in two applications: retrieval in collections of closed silhouettes and holistic word recognition in handwritten historical manuscripts. As a matter of fact it is shown that the proposed contour based approach outperforms all techniques for word matching proposed in the literature when tested on a set of 20 pages from the George Washington collection at the Library of Congress [24, 25]. Additionally, it is shown that the matching approach can be extended with some success to the problem of matching multiple contours and therefore facilitating establishment of dense correspondence between multiple silhouettes which can then be employed for contour registration.

Various applications of the proposed techniques are presented throughout the thesis providing the basis for the third contribution which is a discussion of the feasibility of different recognition strategies utilizing contour information.

1.7 Thesis Overview

The first chapters of the thesis consider segmentation of images of real world scenes while the later chapters focus on estimation of similarities between objects' silhouettes. The final chapters present case studies demonstrating the usefulness of the proposed solutions in selected applications.

Chapter 2 presents the state of the art in image segmentation in the CBIR context. The chapter discusses the relevance of image segmentation in the context of content-based image retrieval, outlines the main categories of approaches, briefly describes the grouping cues most commonly utilized in segmentation and reviews the most characteristic approaches to automatic and semi-automatic image segmentation. The chapter also briefly discusses evaluation methodologies for assessment of segmentation quality proposed thus far in the literature and describes in some detail the evaluation method adopted throughout this thesis.

Chapter 3 explores the feasibility of utilizing the spatial configuration of regions and structural analysis of their contours (so-called *syntactic visual features*) for improving the correspondence of the output of bottom-up segmentation by region merging to semantic objects present in the scene and improving its perceptual quality. A new framework for utilizing evidence from multiple sources of information, each with its own accuracy and reliability, for meaningful region merging is proposed. Two practical measures for analyzing the spatial configuration of image regions and structural analysis of their contours are proposed together with several improvements to colour representations used during the region merging process and a stopping criteria aimed at producing partitions containing the most salient objects present in the scene. Finally, it demonstrates the utilization of the approach in a semi-automatic scenario.

Chapter 4 outlines the most important issues related to shape analysis in the context of content-based image retrieval. It focuses on the major challenges related to 2D shape representation, matching and similarity estimation and gives an overview of selected techniques available in the literature in order to provide a context for the research carried out in the remainder of the thesis, particulary in chapter 5.

Chapter 5 looks at the development of a new representation and matching technique devised by the author for non-rigid shapes with a single closed contour. An initial approach for optimization of the matching in order to achieve higher performance in discriminating between shape classes is also presented. The efficiency of the proposed approaches is demonstrated on two collections and the retrieval results are compared to those of the method based on *Curvature Scale Space* (CSS)[13, 14, 15], which has been adopted by the MPEG-7 standard [16, 17].

Chapters 6 and 7 present two case studies, both utilizing the techniques

proposed in the earlier chapters. Chapter 6 examines the application of the silhouette matching method proposed in chapter 5 to holistic word recognition in historically significant handwritten manuscripts. Utilization of both the silhouette matching method presented in chapter 5 and the semi-automatic segmentation approach discussed in chapter 3, in an integrated tool allowing intuitive extraction and registration of contour examples from a set of images for construction of statistical shape models is discussed in chapter 7.

The final chapter summarizes the thesis by reviewing the achieved research objectives and recalling the main conclusions drawn throughout this thesis. Also, it indicates directions for further research by discussing possibilities for further improvement of the proposed techniques and looks at future prospects in using them to address related challenges in the field.

CHAPTER 2

IMAGE SEGMENTATION, APPLICATIONS AND EVALUATION: A REVIEW

THIS chapter discusses the importance of image segmentation in the context of content-based image retrieval, outlines the main categories of approaches, and reviews the most characteristic segmentation methods. The subject of objective evaluation of segmentation quality is also discussed in some detail as it has received relatively little attention in the literature, especially when compared to the large number of publications on the topic of image segmentation itself.

2.1 Introduction

2.1.1 Image Segmentation

The term image segmentation[1], refers to a broad class of processes of partitioning an image into disjoint connected regions which are homogeneous with respect to a certain criteria such as low-level features (e.g. colour or texture), general prior knowledge about the world (e.g. boundary smoothness), high-level knowledge (e.g. semantic models) or even user interactions.

The problem of image segmentation is a fundamental enabling technology for understanding scene structure and identifying relevant objects. Identification of areas of an image that correspond to homogenous or/and important regions, e.g. semantic objects, is often considered to be the first step of many object-based applications such as for example content-based image indexing and retrieval (e.g. using the recently introduced MPEG-7 [26, 4, 5] standard) or region-of-interest coding using the JPEG2000 standard [27]. This thesis focuses primarily

[1]Often exchanged with the term *perceptual grouping*.

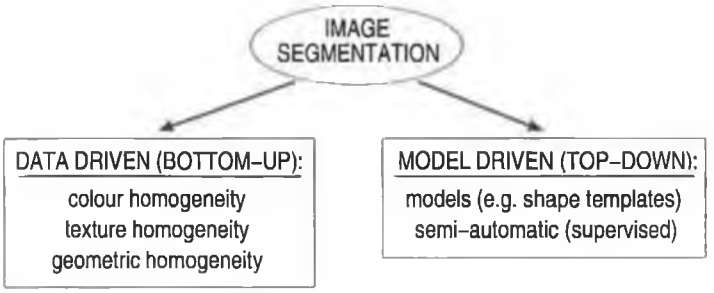


Figure 2.1: Main segmentation strategies.

on region and object based segmentation of images in the context of visual content indexing and retrieval applications.

2.1.2 Taxonomy

A plethora of approaches to image segmentation have been proposed in the literature [28, 29, 30, 31, 32, 33, 34]. Exhaustive surveys can be found in [35, 36, 37, 38, 39]. The techniques available in the literature can be classified according to many different criteria. Many authors divide the methods into two distinct groups: (i) *region-based* approaches relying on the homogeneity of spatially localized features (e.g. colour or texture) and (ii) *boundary-based* methods using mostly gradient information to locate object boundaries. Others classify methods according to the mathematical methodology they employ, e.g. (i) *variational*, (ii) *statistical*, (iii) *graph-based*, and (iv) *morphological*.

Another important criterion, adopted for discussion in the reminder of this thesis, is the information level used for grouping. According to this criterion the methods can be broadly classified as either *visual features-driven* (*bottom-up*) or *model-driven*(*top-down*) [18] – see Figure 2.1. Approaches from the first category attempt to infer meaningful entities (ideally objects) solely from analysis of visual features, e.g. colour or texture homogeneity. Approaches from the second category attempt to segment image into objects whose visual features match best the visual features of the models used. It should be noted that the term *bottom-up* is also often used to refer to methods merging iteratively pixels into more complex entities (e.g. [19]) and the term *top-down* often refers to region splitting techniques (e.g. [40]) or to approaches where segmentation of the down-sampled image is used to initialize the segmentation of the image at finer resolution (e.g. [30]). However in this thesis the terms *bottom-up* and *top-down* are mainly used to refer to the level of information used and not necessarily to the order in which images at different scales are analyzed.

Alternatively, approaches can also be classified as either *object-based* (resulting in semantic objects) or *region-based* (producing regions with homogeneous colour and/or texture). Clearly, there is a close relationship between the last two

criteria, i.e. typically semantically meaningful objects can be produced only by making use of objects' models, prior knowledge about a particular application or user guidance. The approaches *driven by visual features* typically rely on a certain homogeneity criterion, e.g. colour or texture. They are attractive in many applications as they do not require models of individual semantic objects or user interactions. However, they also rarely can be used to extract complete semantic objects as the problem of *visual features-driven* object-based segmentation is under constrained in many applications. In contrast, use of the *model-driven* approaches, although requiring utilization of prior knowledge about the objects to be extracted, leads to well defined detection problems. Typically, in the case of general content, the *visual features-driven* approaches can only partition the input image into homogenous regions although in specific application scenarios can also produce objects, while, the *model-driven* approaches are used mainly to extract the important semantic objects from the irrelevant structures present in the scene. In the case of systems carrying out shape analysis, the segmentation algorithms typically aim at object-based segmentation in order to extract/recover the shape of the relevant objects present in the scene. However, the two approaches (*visual features-driven* and *model-driven*) are not mutually exclusive and they could/should be used both for segmentation and object extraction. For example, in [41] *bottom-up* segmentation was used to obtain a single partitioning of the image and also to build its hierarchical representation in the form of a *Binary Partition Tree* (BPT). BPT was then used to guide the search for the optimum match between a reference contour and the contours of the partition leading to object segmentation and detection.

2.1.3 Chapter Structure

The remainder of this chapter is organized as follows: the next section discuses the relevance of image segmentation in the context of *Content-Based Image Retrieval* (CBIR) focusing primarily on reviewing selected CBIR systems making use of segmentation and describing some recently emerged approaches to image retrieval utilizing automatic segmentation. Section 2.3 briefly discusses cues which can be used for image segmentation. Selected *bottom-up* and *top-down* approaches are reviewed in sections 2.4 and 2.5 respectively. Evaluation methods are discussed briefly in section 2.6. A short discussion of the prospects of image segmentation in the context of CBIR is given in section 2.7 and conclusions are formulated in section 2.8.

2.2 Relevance to Content-Based Image Retrieval

Utilizing segmentation in CBIR systems allows representation of images at the level of homogenous regions, which in an ideal case correspond to semantic

objects or at least to significant parts of the objects. Such region or object granularity allows utilization of local features for indexing and retrieval in a manner closer to human perception and scene understanding.

State of the art CBIR systems make use of image segmentation to extract a separate set of indexing features for each region (or object) [42, 43, 29, 44]. According to [45], the systems utilizing segmentation can be divided into two categories depending on the strategy used for matching regions from the query and the target image: (i) *individual region matching* - where a single region selected by the user from the query image is matched to all regions in the collection [46, 43] or alternatively retrieval results obtained by using several queries with individual regions are merged [29] and (ii) *frame region matching* - where information of all regions composing the images is used [47, 45]. Recently, the possibility of another category of systems utilizing image segmentation emerged when it was shown that keywords can be associated with image regions automatically during an off-line training/indexing process [8]. Such keywords can be then used for indexing and subsequently for textual querying.

The remainder of this section aims at highlighting the importance of image segmentation in CBIR by reviewing selected approaches utilizing some form of image partitioning.

2.2.1 CBIR Systems Utilizing Segmentation

Recent years have witnessed an explosion of image retrieval systems, developed as platforms facilitating research or even as commercial products. Extensive reviews of the major CBIR systems can be found in [2, 48, 3]. This section briefly overviews those systems utilizing some form of image segmentation.

One of the earliest CBIR systems, and also one of the first to provide limited object-based functionalities is the *Query By Image Content* (QBIC) system developed by IBM Almaden Research Center [49, 50, 11]. In this approach simple low-level features (colour, texture and simple shape descriptors) are extracted from entire images or objects semi-automatically segmented in the database population step. The segmentation is based on the active contours technique which aligns the approximate object's outline provided by the user with the closest image edges. Each extracted object can be annotated by textual description. All images are also represented by a reduced binary map of edge points to allow retrieval based on rough sketches. The system allows combination of text-based keyword queries with visual queries, i.e. queries-by-example, by a rough sketches and selected colour and texture patterns.

One of the first image retrieval systems utilizing automatic image segmentation is *Picasso* developed by the Visual Information Processing Lab, University of Florence [51, 52]. At the core of the system is a pyramidal colour segmentation

of images (i.e. at the lowest level, each region consists of only one pixel, while the top of the pyramid corresponds to the the entire image) obtained by iterative merging of adjacent regions. Each level of the pyramid corresponds to a resolution level of the segmentation. The above representation is stored in the form of a multilayered graph in which nodes representing regions are characterized by colour (a binary 187-dimensional colour vector), position of its centroid, size, and shape (elongation and the major axis orientation). Additionally, during database population the objects of interest in each image are bounded with their minimum enclosing rectangle and a Canny edge detector is used within each of these rectangular areas to extract edges. The system allows querying by colour regions (by drawing a region with a specific colour or by tracing the contour of some relevant object in an example image), by texture, or by shape (drawn sketches).

Netra, developed by the Department of Electrical and Computer Engineering, University of California, Santa Barbara, is another CBIR system utilizing image segmentation [53, 46]. In this system images are automatically segmented into regions of homogeneous colour using an *edge flow* segmentation technique [54]. Each region is characterized by colour (represented by a colour codebook of 256 colours), texture (represented by a feature vector containing the normalized mean and standard deviation of a series of Gabor wavelet transforms), shape (represented by three different feature vectors: curvature at each point on the contour, centroid distance function, and Fourier descriptors) and spatial location. The query is formulated by selecting one of the regions from the query image or alternatively, if the example image is not available, directly by specifying the color and spatial location.

iPure is a CBIR system developed by IBM India Research Lab, New Delhi [42]. In this system images are segmented into regions with homogenous colour using the *Mean-Shift* algorithm [55, 56, 33]. Each region is represented by colour (average colour in CIE LUV space), texture (coefficients of the Wold decomposition of the image viewed as a random field), and shape (size, orientation axes, and Fourier descriptors). Additionally, the spatial lay-out is characterized by the centroid, the minimum bounding box, and contiguity. The query is formulated by selecting one or more regions from the example image.

Istorama is an image retrieval system developed by Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece [43]. In this system all images are automatically segmented by a variant of *K-Means* algorithm called KMCC [30]. Each region is characterized by its colour, size and location. Similarly to the Blobword system the user formulates the query by selecting a single region from the query image. Additionally, the user may stress or diminish the importance of a specific feature by adjusting weights associated with each feature.

Blobworld is a CBIR system developed at Computer Science Division, University of California, Berkeley [29]. In this approach images are segmented into regions with uniform colour and texture (the so-called blobs) by a variant of the *Expectation Maximization* (EM) algorithm. Each region is characterized by its colour (represented by a histogram of the colour coordinates in the CIE LAB colour space), texture (characterized by mean contrast and anisotropy over the region), and simple global shape descriptors (size, eccentricity, and orientation). The query-by-example is formulated by selecting one or more regions from one of the query images followed by specification of the importance of the blob itself and also its colour, texture, location, and shape.

An interesting methodology to image retrieval was presented in [44]. In this approach images are automatically segmented by a variant of *K-Means* algorithm called KMCC [30] and the resulting regions are represented by low-level features such as colour, position, size and shape. The main innovation of this methodology is automatic association of these descriptors with qualitative intermediate-level descriptors, which form a simple vocabulary termed *object ontology*. The object ontology allows the qualitative definition of the high-level concepts and their relations in a human-centered fashion and therefore facilitates querying using semantically meaningful concepts. Applicability to generic collections without requiring manual definition of the correspondences between regions and relevant identifiers is ensured by simplicity of the employed ontology. Utilization of the object ontology allows narrowing down the search to a set of potentially relevant images. Finally, a relevance feedback mechanism, based on *Support Vector Machines* and using the low-level descriptors is employed for further refinement of retrieval results.

2.2.2 Automatic Annotation of Image Regions

Some user studies suggest that in practice queries based on global low-level features (e.g. colour histograms, texture) are surprisingly rare while at the same time text associated with images is particulary useful [57]. Interestingly, a number of works [8, 9] have shown that keywords can be associated with image regions automatically during an off-line training/indexing process. Such labels can be then used for indexing and subsequently for textual querying.

One of the most promising approaches so far to predicting words from segmented images was described in [9]. In this approach a statistical model links words and image data, without explicit encoding of correspondence between words and regions. This model combines the aspect model with a soft clustering model. It is assumed that images and co-occurring words are generated by nodes arranged in a tree structure. The joint probability of words and image regions is modelled as being generated by a collection of nodes, each of which has a probability distribution over words and regions. The region probability

distributions are Gaussians over feature vectors and the word probabilities are provided by simple frequency table. A region's features imply a probability of being generated from each node. These probabilities are then used to weight the nodes for word emission. Parameters for the conditional probabilities linking words and regions are estimated from the word-region co-occurrence data using an *Expectation-Maximization* algorithm. While the approach was shown to label only some regions reasonably well (e.g. sky, water, snow, people, fish, planes) it certainly represents an extremely interesting approach to object recognition. In [58] three classes of image segmentation algorithms (*Blobworld* [29], *Normalized Cuts* [28] and *Mean-Shift* [33]) were evaluated in the context of the above annotation approach, e.g. in terms of the word prediction performance. It was found that the *Normalized Cuts* approach provides the best support of all three methods for word predictions closely followed by the *Mean-Shift* segmenter.

2.3 Grouping Cues

This section looks at possible features (grouping cues) which are relevant to the image segmentation problem.

2.3.1 Colour

Colour is the primary low-level feature in practically all image segmentation techniques which can be found in the literature [29, 30, 31, 59, 18]. One of the main issues related to colour is the choice of an appropriate colour space. Typically it is advantageous that the colour space used guarantees a low correlation among components and is *perceptually uniform*, i.e. the numerical distance is proportional to the perceived colour difference. Due to the above requirements many resent publications advocate utilization of CIE LUV and its improved version CIE LAB colour spaces which for similar colours are approximately perceptually uniform [28, 29, 30].

2.3.2 Texture

Texture is arguably the second most important low-level feature after colour in the image partitioning problem. Several approaches to texture characterization have been shown to be useful for segmentation [30] including *multi-orientation filter banks* [60] and the *second-moment matrix* [61, 29].

However, utilization of texture in image segmentation should not be considered simply as a thoughtless extension of the colour feature. There are several issues which must be addressed when making use of texture information in order to achieve satisfactory results for natural images often depicting both textured

and untextured objects. An interesting discussion on utilization of texture information for image segmentation can be found in [60], where the authors discuss issues like adaptive scale selection, i.e. the size of the local window used to compute texture descriptor – see also [29], and the fact (rarely discussed in the literature) that contours may constitute a serious problem for many texture analysis methods, i.e. filters tuned to various orientations and spatial frequencies which are often used for texture analysis produce large responses along the direction of the edge. Finally, it should be noted that colour and texture often provide conflicting evidence (e.g. consider a zebra-like pattern) which must be addressed in the final decision step of the segmentation process. Currently, the most common approach to alleviate such conflicts is a conditional filtering of the colour components, i.e. textured areas are smoothed in order to ensure their colour homogeneity (e.g. a zebra becomes a grey horse (colour feature) with stripes (texture feature)) [29, 30].

2.3.3 General Geometric Cues

Many researchers attempted to incorporate general geometric prior knowledge about the world [62, 63, 64, 65, 66, 67, 18, 68] into the segmentation process. The main idea is to use some generic geometric cues which allow segmentation of images into meaningful entities without building object specific models.

One example are methods based on the deformable model paradigm of *Active Contours* which incorporate prior knowledge about a contour's smoothness and resistance to deformations [64, 65, 66]. Other methods attempt to perform *Perceptual Grouping* which is typically defined as a process that organizes image entities (typically edges or regions) into higher-level structures based on some generic geometric perceptual measures. The majority of such approaches are directly motivated by theories from the field of *Cognitive Psychology* regarding human perception and object recognition[2] such as for example the *principles of Gestalt theory* and the *principle of common-cause*[3].

Geometric perceptual measures commonly used for perceptual grouping (and potentially to image segmentation) are proximity, continuity, closure, compactness, edge parallelism, collinearity, and cocircularity (see examples in a special section on Perceptual Organization in Computer Vision in [67]). Another example of a set of geometric perceptual measures potentially useful for segmentation are the *Syntactic Visual Features*, and in particular a *Quasi-Inclusion Criterion*, advocated by Ferran and Casas [18] for creation of *Binary Partition Trees* (BPT). Syntactic visual features include complex homogeneity (e.g. shape homogeneity), compactness, regularity, inclusion and symmetry.

[2] A brief discussion of the main models of the human visual system and object recognition can be found in appendix A

[3] Based on the idea that many relations in visual information have a low probability to occur at random [69]

Although utilization of geometric properties and/or structural relationships for segmentation is intuitive, integration of such information with other grouping cues (e.g. colour and texture) is a non-trivial task. In many approaches proposed so far such integration is performed in an ad-hoc manner [62, 63, 18, 68]. Moreover, despite the fact that generality of many of the geometric cues is arguable very little work has been done to quantitatively demonstrate their usefulness on large collections, i.e. typically a handful of examples of segmentation is shown.

2.3.4 Application Specific Models

In some applications it is possible to use a prior model of what is expected in the image further reducing ambiguities and leading to fully automatic techniques for object-based segmentation, detection and recognition. In other words, the partitioning process is guided by models of specific objects to be segmented. Intuitively, such models should allow modeling of common real world deformations and variations within object classes. Models commonly used for segmentation and object recognition are shape models [70, 71, 72, 73, 74, 75, 76] and models of colour (or intensity) variations, i.e. designed for synthesizing new images of the modelled object (e.g. *Active Appearance Models* (AAMs) [77, 78] or *Eigenfaces* [79]).

2.3.5 User Interactions

Semi-automatic or *supervised* approaches allow a user, by his/her interactions, define what objects are to be segmented within an image [32]. The result of interaction provides clues as to the nature of the user's requirements which are then used by an automatic process to segment the required object. The user should have the possibility of obtaining exactly the segmentation he/she desires in terms of content and accuracy. The amount of required interactions should not be excessive and should be easily and quickly performed. Also, it is important that the system responds to the users interactions in real-time, ideally instantaneously. Rapid responses are crucial for cases when interactions need to be applied iteratively in order to refine the segmentation result, e.g. for complex objects or objects with adjacent parts exhibiting a certain degree of homogeneity. The strategies to user interaction can be grouped in three classes [80]: *feature-based* [81, 82, 83], *contour-based* [84, 34], and *region-based* [85].

2.4 Selected Methods for Bottom-up Automatic Segmentation: A Review

This section describes selected methods for bottom-up automatic segmentation of heterogenous still images, typically colour, without utilizing any a priori knowledge about the scene depicted in the image.

As already explained, the literature in the area of automatic image segmentation is enormous. Different methodologies were used in the past to tackle the problem of segmentation such as: *histogram thresholding* [86, 87] (i) *clustering* [88, 89, 90, 30, 29, 91, 56, 33], (ii) *mathematical morphology* [92, 93], (iii) *graph theoretic algorithms* [19, 94, 59, 28, 95, 28], (iv) *statistical methods* [96, 97, 98], (vi) *edge based segmentation* [99, 100], (vii) *neural networks* [101] and many others. This section discusses three categories of approaches which arguably are the most significant in the context of content-based image retrieval for general content and which are also relevant to the work presented in this thesis in chapter 3.

2.4.1 Clustering

Clustering methods are one of the earliest approaches to image segmentation [90]. Traditional clustering approaches typically ignore spatial information and operate solely on the feature space (usually colour or texture). The feature samples computed for each pixel are treated as vectors in multi-dimensional space and the objective is to group them into compact, well-separated clusters using a specific distance measure. Once the clustering is completed cluster labels are mapped onto the image plane to produce the final regions. The most widely used methods from this category are *K-Means* algorithm [88] and its fuzzy equivalent *Fuzzy C-Means* [89]. The main drawback of the above approaches are: (i) they require *a priori* knowledge of the number of clusters, (ii) they produce disconnected regions, (iii) they assume approximately spherical clusters in the feature space.

To alleviate the first two drawbacks the *K-Means-with-Connectivity-Constraint* (KMCC) algorithm was proposed in [30]. In this approach, the segmentation is performed in the combined intensity-texture-position feature space in order to produce connected regions that correspond to real-life objects. The initial number of clusters is estimated by applying a variant of the *MaxiMin* algorithm to a set of non-overlapping image blocks. The problem of disconnected regions is alleviated by integrating *K-Means* with a component labelling procedure, i.e. at each iteration the initial clusters (possibly disconnected) are broken down to the minimum number of connected regions using a recursive four-connectivity component labelling algorithm and the centres of the new regions are recalculated.

Other approaches from this category are based on parametric models of feature distributions. The best-known approach from this category is *Blobworld* [29]. In this approach, the joint distribution of colour, texture, and position features is modelled with a mixture of Gaussians. The parameters of this model are then estimated using the *Expectation-maximization* (EM) algorithm and the final segmentation is defined based on the pixels' cluster memberships, i.e. pixels are assigned labels from clusters for which the pixel attains the highest likelihood. The number of mixture components (K) required by the EM algorithm is found by repeating the clustering for a selected range of K where the initial parameters for each EM process are found from a predefined K squared windows in the image. The final partitioning is selected from the set of K solutions according to the *Minimum Description Length* (MDL) principle.

Since the same simple parametric shape for all clusters cannot handle the complexity of real feature space and also mixture models require the number of clusters as a parameter, recently *density estimation-based non-parametric* methods have gained in popularity. In the approaches from this category, cluster centres are automatically identified based on the assumption that dense regions in the feature space correspond to local maxima (modes) in the probability density function. The clusters associated with each mode are delineated based on the local structure of the feature space [91]. One of the most powerful techniques allowing detection of modes of the density is the *Mean-Shift* procedure recently rediscovered in [55, 56, 33] after it was forgotten for more than twenty years. In this procedure, a kernels (windows) initiated at each pixel iteratively move in the direction of the maximum increase in the joint density (i.e. toward region where the majority of feature points reside) until they reach local maxima (peaks) of the density (i.e. stationary points in the estimated density) leading to the identification of the modes of the density (cluster centres). Note that flat plateaus in the density may cause the procedure to stop, therefore modes which are closer than the width of the kernel need to be grouped together. Each pixel is then associated with a significant mode of the density located in its neighbourhood. Although the technique was originally designed for image segmentation by analysis of only colour feature space [55], more recently it was also successfully applied to the joint spatial-range domain [33]. An interesting property of the approach is that it successfully creates large gradient regions which are typically over-segmented by other clustering methods. Its state of the art performance encouraged many researchers to utilize the approach in various applications such as image retrieval [42], object-based video coding for MPEG-4 [102] and shape detection and recognition [103].

2.4.2 Mathematical Morphology

One of the most prominent methods from this category is the *watershed transform* [92] which associates a region to each minimum of the image gradient. The application of the watershed transform involves typically three stages: (i) *pre-processing* - the image is filtered in order to reduce the presence of noise, (ii) *marker extraction* - starting points for the flooding process are extracted, e.g. small homogeneous regions, and (iii) *watershed transform* - the topographic surface of a gradient image is flooded by sources located at the markers. As the markers grow, dams are erected to separate lakes produced by different sources. The final segmentation is defined by the set of dams produced during the flooding process. To reduce the over-segmentation due to many local minima caused by noise, multi-scale gradient operators have been proposed in [93].

2.4.3 Graph Theoretic Algorithms

In *Graph Theoretic* approaches, images are represented as weighted graphs, where nodes correspond to pixels or regions and the edges' weights encode the information about the segmentation, e.g. pairwise homogeneity or edge strength [19, 59, 28, 104]. The segmentation is obtained by partitioning (cutting) the graph so that a criterion is optimized. Several approaches from this category utilize *Region Adjacency Graphs* where nodes represent regions while edges contain region pairwise dissimilarities [105, 59]. The pairwise dissimilarities are computed using a distance between regions' colour features. RAGs can be simplified by successive mergings of neighboring regions, however these operations need to be guided by global information to avoid erroneous solutions.

An interesting representation that emerged from the RAG structure was the *Shortest Spanning Tree* proposed by Morris et al. in [19] and applied to image segmentation and edge detection problems. The approach was later extended to colour images in [94]. A fast version of the *Recursive Shortest Spanning Tree* (RSST) was proposed in [106, 95]. An equivalent region growing algorithm was studied extensively in the work of Salembier [59] and used for creation of *Binary Partition Trees* (BPT), a hierarchical structure that was shown to be useful for filtering, retrieval, segmentation and coding [38, 107].

Although the above approaches are typically computationally efficient, the merging criteria are based on local properties of the graph which may lead to suboptimal solutions to extracting the global impression of a scene. To alleviate this problem a new global graph-theoretic criterion for measuring the goodness of an image partition, called *Normalized Cut*, was proposed in [28]. In this approach, pixels are represented as a weighted undirected graph where all nodes in the graph are connected and the weigh at each node is a function of the pairwise similarity both in spatial and feature space. The graph is partitioned

into two disjoint sets by minimizing the value of the *normalized cut* defined as the cut cost (total weight of the removed edges) as a fraction of the total edge connections to all nodes in the graph. The minimization of this criterion can be efficiently performed based on a generalized eigenvalue problem, e.g. the eigenvectors are used to partition the image into two disjoint regions and if desired the process is continued recursively. In recent years the technique attracted considerable attention in the CBIR community due to its state of the art performance – see for example [58].

2.5 Selected Methods for Top-down Segmentation: A Review

It is common knowledge that fully automatic image segmentation based solely on low-level processing techniques often fails not only due to the complexity of the real world, e.g. noise illuminations, reflections, shadows and colour variability within objects etc., but also due to the fact that the interpretation of the scene is typically context dependent. These limitations of the bottom-up approaches can be alleviated by utilization of prior knowledge of the world or user interactions to additionally constrain the partitioning problem. As already explained in section 2.3 the prior knowledge can be general, e.g. boundary smoothness, or application specific, e.g. models of expected semantic objects. This section briefly reviews selected approaches to supervised segmentation and to segmentation and detection based on shape models.

2.5.1 Semi-Automatic Segmentation

Various strategies to user interaction for extraction of accurate semantic objects from images were proposed in the past. They can be grouped in three classes [80]: *feature-based* [81, 82, 83], *contour-based* [84, 34], and *region-based* [85]. Each category of interactions requires different approaches to the automatic process performing the segmention.

Feature-based Interactions

Feature-based approaches allow the user to draw on the scene, typically by placing *markers* in the form of single point *seeds* or by dragging a mouse over the image and creating a set of labelled *scribbles* for objects to be segmented. Typically, two or more different markers(or scribbles) can have the same label, indicating different parts of the same object in the image.

Probably the best known example of an approach guided by such interactions is *Seeded Region Growing* (SRG) technique [81]. This method is controlled by a small

number of pixels, known as seeds, chosen by the user from K sets. The process evolves inductively from the seeds, labelling pixels which border at least one of the labelled regions. In each step a single unlabelled pixel with the most similar colour to the mean colour of one of the adjacent labelled sets is added to that set. The process is repeated until all pixels are allocated. The approach works well in noisy images but is sensitive to seed initialization (the size and position of seeds may affect the initial average colour and consequently the consecutive iterations) and in some cases results in jagged boundaries. Also the approach seems to be not suitable for cases where the region of interest consists of more than one homogeneous subregions. [81] states that the method may not be applicable to highly textured images.

Another approach to partition the scene into objects based on scribbles was proposed by Chalom and Bove in [82] and further improved by O'Connor [32]. Both approaches [82, 32] model the probability distribution functions (PDF) of each feature (colour, texture and location) parametricaly as a sum of Gaussian PDFs. The segmentation is performed using an *Expectation Maximization* (EM) algorithm. The main drawback of the above methods is the additional requirement of an estimated number of modes for each object and the relatively large number of user interactions required.

An elegant and efficient solution to image partitioning based on scribbles was described in [38] where the input image is automatically pre-segmented into regions and represented as a *Binary Partition Tree* (BPT) to allow rapid responses to users' actions (an approach based on morphological flooding performed using pre-computed tree representation was also presented in [83]). The tree structure is used to encode similarities between regions pre-computed during the automatic segmentation. This structure is then used to rapidly propagate the labels from scribbles provided by the user to large areas (regions) of the image.

Recently, a very interesting approach was proposed in [108] where the globally optimal segmentation is found using *Graph Cuts* approach. The user's scribbles provide hard constraints for segmentation while additional soft constraints incorporate both boundary and region information, i.e. the cost function is defined in terms of boundary and region properties of the segments. The relative importance between the boundary properties term and the region properties term can be controlled by the user. Interestingly the method is valid for N-D images, e.g. 3-D images or even video sequences treated as a single 3D volume.

Contour-based Interactions

Contour-based approaches typically require initialization of the contour model close to the object's borders or placing a seed point on the interior of the region. In some applications however, e.g. medical image analysis [109], the rough

position of the object is known a priori and the initialization of the contour can be achieved automatically.

Typically, the segmentation is based on the deformable model paradigm of *Active Contours* (often called *Snakes* and also referred to as *Variational Methods*), proposed first in the seminal work of Kass et. al. [64]. In these approaches, the models incorporate prior knowledge about a contour's smoothness and resistance to deformations. A regularized estimate of a contour is obtained by applying external forces that attract the model to edges present in the image. It has been shown in [110] that the model can be made less sensitive to the initialization by adding an additional "internal" force imitating inflation of the contour.

The first active contour models were implemented using parametric representations of the evolving curves [64]. One of the main advances in the area was introduction of the *Geodesic Active Contours* [111, 112] models which are parameterization free. In these approaches the segmentation involves solving the energy-based minimization problem by the computation of geodesics or minimal distance curves [112]. The evolving contour is embedded as the zero level-set of a higher dimensional surface (typically signed distance function). The entire surface is then evolved minimizing a metric defined by the curvature and image gradient. Although more computationally expensive than the parametric approaches, the implicit implementation allows convenient automatic handling of changes of contour topology, i.e. splitting and merging, allowing simultaneous detection of several objects. *Geodesic Active Contours* are less sensitive than the parametric counterparts to the initialization requiring only the initial curve to be placed completely inside or outside of the object boundaries.

Recently, motivated primarily by the (region based) Mumford-Shah functional [113], the active contour formulation has been extended to deal with multiple cues [65] (e.g. by including region-based terms) leading to *Active Regions* models which further increase robustness to the initialization and noise. For example, in [114] supervised texture segmentation was obtained by guiding the curve evolution by both: (i) the boundary-based (geometric) term and (ii) region-based (statistical) terms.

Finally, it should be noted that the variational approaches can also be used for fully automatic segmentation. For example, in the approach proposed in [65], the initial region seeds are scattered randomly over the entire image followed by curve evolution aided via Bayes/Minimum Description Length (MDL) criteria. In [66] an unsupervised textured image segmentation method based on geodesic active contours is proposed where the regions and the parameters of the distributions are updated simultaneously thus dynamically learning the model of each region.

2.5.2 Shape Driven Segmentation

Shape Models

While methods based on active contours incorporate a priori knowledge which is relatively general they typically require careful initialization of the model close to the object's borders or placing a seed point on the interior of the region. In some applications it is possible to incorporate much more specific models, further reducing ambiguities and leading to fully automatic techniques for object-based segmentation, detection and recognition. This section focuses on shape models and their applications.

Shape models can be divided into two categories: (i) special-purpose deformable templates carefully constructed ("hand-crafted") for a particular problem [115] and (ii) templates derived semi-automatically via statistical analysis of training examples [116, 117].

Arguably the best known statistical shape model is the *Point Distribution Model* (PDM) (also called *Active Shape Models*(ASMs) or "Smart Snakes") proposed by Cootes and Taylor [70, 117]. The model is derived by examining the statistics of the coordinates of the labelled points over the training set via the *Karhunen-Loeve* transform. The model gives the average position of the points, and a description of the main modes of variation. The approach was later extended to models capable of synthesizing new images of the object of interest, the so-called *Active Appearance Models* (AAMs) [78].

Other statistical approaches to shape modeling include methods proposed by Pentland and Sclaroff [116] (where the contours of prototype objects are represented using *Finite Element* methods and describe variability in terms of vibrational modes), Lades et al. [118] (where shape and grey level information is modelled using elastic mesh and Gabor jets), Mardia et al. [119] and Grenander et al. [120] (the use of which in image segmentation, however, appears difficult [78]).

Utilization of Statistical Shape Models for Image Segmentation

There are several different approaches which could be taken to make use of statistical shape models in the segmentation process [70, 117, 72, 73, 74, 75, 76, 71]. However, all involve an optimization step finding the set of parameters which best match the model to the image (or hypothetical group of regions from the image).

The most straightforward approach is to fit different instances of the expected object to the image. For example, in the approach proposed by Cootes et. al. [70, 117] different sets of parameters are used to generate instances of the

shape from the PDM model which are then projected into the image. The set of parameters which minimizes a cost of fitting the instance to the image measure is chosen as the best set to interpret the object. Note that the above use of statistical models shares some similarities with approaches matching a single shape to an unseen image – see for example [121].

Recently, a considerable number of approaches have been proposed for integration of statistical shape models into the variational segmentation process [72, 73, 74, 75, 76] where the evolved contour is constrained not only by the smoothness and the fidelity of the segmented image but also by the compatibility with the expected shape. The main difficulty in such approaches is the need to account for possible pose transformations between the model and the object present in the segmented image. An example of such an approach was proposed in [72] where shape information is incorporated into the evolution process of *Geodesic Active Contours*. An initial curve is embedded as the zero level set of a higher dimensional surface (in this case signed distance function). The training set consists of a set of surfaces and the shape model is build over the distribution of surfaces similar to the approach from [70] for constructing PDMs. At each step of the evolution, the maximum a posteriori (MAP) position and shape of the object is estimated based on the model and the image gradient. Finally, an active contour is evolved based on image gradients, curvature and towards a maximum *a posteriori* estimate of shape and pose.

Finally, statistical shape models can be used to aid the region grouping process. One such approach was proposed by Sclaroff and Liu [71]. The method consists of two stages: (i) over-segmentation using the *mean-shift* colour region segmentation algorithm and (ii) region merging via grouping and hypothesis selection guided by deformable shape models. In the second stage, various combinations of candidate adjacent region groupings are tested. The candidate regions are selected based on colour characteristics. The number of merging hypothesis is limited by evaluation of edge strengths between the initial regions. Each grouping hypothesis is matched with the model by deforming the model in order to minimize a cost function consisting of three terms: (i) a colour compatibility term, (ii) an area overlap between model and grouping hypothesis, and (iii) a deformation term. The final partitioning of the whole image is obtained by enforcing global consistency, i.e. a global cost function consisting of two terms: (i) sum of costs of individual grouping hypotheses and (ii) number of groupings, minimized using the *Highest Confidence First* algorithm (HCF). The system was shown to perform well on images containing fruits, leafs, fish, and blood cells. Although the technique can only segment objects consisting of one colour, utilization of the shape model allowed correct segmentation in the presence of shadows, poor lighting conditions and object overlapping.

2.6 Evaluation

Meaningful assessment of performance is crucial for the development of any algorithm. However, the subject of objective evaluation of segmentation quality has received relatively little attention in the literature [122, 123] compared to the large number of publications on the topic of image segmentation itself. This state of affairs is surprising since clearly discussion on a consistent and an objective evaluation criterion is crucial for comparison of algorithms and could help answer fundamental questions regarding the feasibility of incorporating segmentation in the context of a particular application.

The remainder of this chapter discuses difficulties that must be faced in comparing segmentation results, describes briefly evaluation methods proposed in the literature, and explains the methodology adopted by the author to evaluate the approach discussed in the next chapter.

For the sake of brevity, the discussion focuses on evaluation of region-based image segmentation, however, many issues presented here are relevant to other segmentation scenarios.

2.6.1 Evaluation Strategies

The current practise in the field of region-based segmentation is to justify new advances by providing a set of qualitative results, i.e. by presenting selected segmentation examples. Although such an approach may be sufficient to illustrate the effects of low-level processing, e.g. edge detection, it does not provide satisfactory demonstration of the generality of segmentation methods. This is especially important in the context of CBIR systems operating with general content.

Another alternative is subjective ad-hoc assessment by a representative group of human viewers. Subjectivity should be minimized by respecting strict evaluation conditions, e.g. [123] suggests to follow recommendations for video quality evaluation developed by ITU [124, 125]. However, such a process requires a significant number of human evaluators and is very time-consuming. Because of its high cost, such evaluation is extremely rare. The author is familiar with only one such study presented in [98], where subjective experiments provided the basis for better understanding of human perception of segmentation artifacts.

From another point of view, segmentation is almost never the final objective in any application. Typically (e.g. in CBIR), it is just one step towards scene understanding. In fact, the main difficulty in segmentation evaluation lies in establishing a criterion adequate to a given application [123]. Therefore, it should be possible to perform meaningful objective assessment of a grouping

algorithm by measuring the overall performance of the system in which it is integrated. However, in the case of CBIR systems, such an approach is often impractical due to its complexity. Usually the performance of the system depends on combinations of parameters of the system and the segmentation algorithm itself. Therefore, it is difficult to guarantee optimal evaluation conditions. Moreover, in recent years, it has become apparent that the key element of such systems are user interactions, and relevance feedback [10] in particular. Unfortunately it is practically impossible to take such factors into account in extensive evaluation experiments such as those required for development of a new segmentation approach.

Since the development carried out in the next chapter requires multiple series of extensive evaluation experiments, all three above strategies had to be discarded. Instead, objective automatic evaluation methods based on measures capable of adequate assessment of segmentation quality are investigated.

There are two main classes of automatic assessment methods: *standalone* and *relative*. The first one is typically performed only in cases where no reference segmentation is available. It exploits available knowledge about expected properties of desirable segmentations for a particular application. The main difficulty lies in establishing an evaluation measure suitable for the type of content and application addressed. Since the results of such evaluation do not necessarily correspond to human perception of the segmentation quality [126, 127], this scenario is not considered in this thesis. However, it should be noted that despite the above limitation, the problem of standalone evaluation of segmentation quality is central to the problem of segmentation itself. As a matter of fact, there is certain analogy between features proposed for standalone evaluation in [126] and the geometric properties of regions proposed for incorporation within the segmentation framework in the next chapter.

Relative evaluation methods are based on similarity measures used to quantify differences between the evaluated and reference segmentation often referred to as *ground-truth*. This type of evaluation is significantly more reliable than the standalone method, however large collections of test data with a ground-truth are rarely available. It should also be recognized that relying on a single reference segmentation mask for each image may be questionable, especially in the context of CBIR. The quality of partitioning is query and user dependent and possible queries are typically not known prior to segmentation. A ground-truth collection consisting of single masks for each image does not take into account such ambiguities in the partitioning of the scene. This issue is discussed in more detail in section 2.6.4 which describes the creation of the ground-truth for the collection utilized in all experiments presented in the next chapter.

Despite the above limitation, relative evaluation methods represent the best trade-off between reliability of evaluation (value of the information provided)

and the practicability of extensive evaluation experiments required for the next chapter. Therefore, the discussion presented in the following sections is limited only to relative evaluation methods.

2.6.2 Relative Evaluation Methods

One of the main difficulties in objective evaluation is finding an adequate measure capable of assessing segmentation quality aligned with human perception of segmentation quality – see the discussions in [123] and in [98]. Since none of the methods proposed in the literature are currently well-established in the research community, this section gives a brief review of available approaches most relevant to experiments in this thesis and provides arguments for the adopted solution. For a more comprehensive review of the existing methods the reader should refer to [127, 98].

One of the earliest set of evaluation methods suitable for video segmentation were proposed in [128, 129]. They evaluate both spatial and temporal accuracies for scenes composed of two objects (foreground & background). In [128] spatial accuracy is measured in terms of misclassified pixels. Weighting of such misclassified pixels according to their distance to the reference object was introduced in [129].

A more advanced methodology was proposed in [123] where a comprehensive set of relevant features to be compared was identified and objective quality metrics combining these features were suggested. The proposed classes of spatial features included: (i) *shape fidelity* - computed similarly to spatial accuracy measure from [129], (ii) *geometrical similarity* - based on similarities of simple geometrical features like size, position, elongation and compactness, (iii) *edge content* - based on similarities between edge contents of the reference and estimated object, (iv) *statistical data similarity*, e.g. similarity of the brightness and "redness" values of reference and estimated objects. Although the above list of features is quite comprehensive, the usefulness of combining all different measures into one single evaluation criterion is questionable. Combining various features using ad-hoc weights seems impractical since in many cases analysis of such combined results may still require manual inspection of the result in order to fully understand which features contributed to the combined error (and in what ways). In this thesis, simpler, but easier to interpret measures are preferred.

A very interesting recent study of the impact of topology changes, in particular *added regions* and *holes*, was presented in [98]. Subjective tests have been carried out to measure the annoyance of these artifacts when presented alone or in combination. The experiments indicated that, independently from content, annoyance values corresponding to *added regions* quickly reaches saturation

level for larger sizes. It also confirmed the common hypothesis that *holes* contribute more annoyance compared to *added regions* of the same size. Another important output of the above experiments was demonstration that various types of topological errors contribute differently to the overall annoyance. Moreover, the proportions in which they contribute to the annoyance are content dependent. Therefore, more studies of this problem have to be carried out on this subject before combination of multiple errors and such differentiations can be meaningfully incorporated in automatic evaluation metrics for general content.

A relatively simpler, yet well aligned with human intuition, method was proposed in [127]. It was specifically designed for evaluation of image segmentation in the context of CBIR systems. It is primarily based on the approach from [129] for evaluation of spatial accuracy, but extended to still image segmentation where both the evaluated segmentation and the ground-truth mask might contain several regions. Special consideration was devoted not only to the accuracy of boundary localization but also to over- and under-segmentation. To tackle such issues, the evaluation starts by establishing exclusive correspondence between regions from reference and evaluated masks. Then, three different types of errors are taken into account: (i) accuracy errors for the associated pairs of regions from reference and evaluated masks, (ii) errors due to under-segmentation computed based on non-associated regions from the reference mask, and finally (iii) errors due to over-segmentation computed from non-associated regions from the evaluated mask. Although the method is quite simple, a set of convincing evaluation results was provided in [127] indicating that the method correlates well with subjective evaluation. The above method is adopted in this thesis due to the small number of parameters involved and ease of interpretation. A formal description of the method and justifications for parameter values chosen for all experiments are provided in section 2.6.3.

Finally it should be noted that humans are very sensitive to errors corrupting objects with high semantic value. All the above studies regarding evaluation acknowledge the importance of object's relevancy, e.g. such relevance measure could indicate the likeliness that a particular object will be further manipulated and used [130]. However, due to the difficulties with providing such relevancy information within the reference segmentation currently it is rarely taken into account during evaluation.

2.6.3 Adopted Evaluation Method

This section gives the formal description of the evaluation method proposed in [127] and discusses the author's implementation of the above approach.

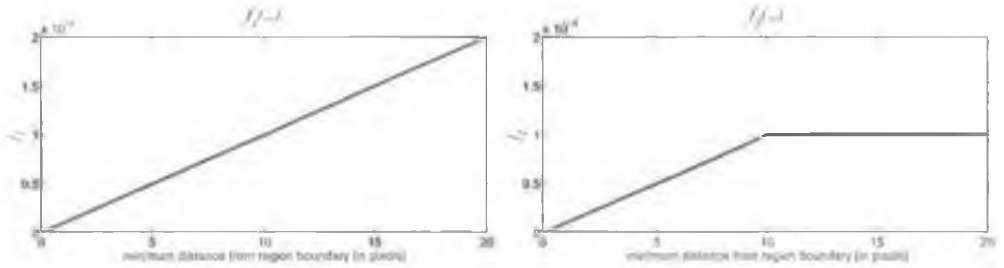


Figure 2.2: Functions used for weighting the contributions of misclassified pixels based on their distance from the reference region to which they belong: $f_1(.,.)$ is used for false negatives (pixels belonging to r_q but not in s_k), whereas $f_2(.,.)$ is used for false positives (pixels assigned to s_k but not in r_q) [129].

The Original Evaluation Approach from [127]

Let $\mathcal{S} = s_1, s_2, \ldots, s_K$ be the segmentation mask to be evaluated comprising K regions s_k, and $\mathcal{R} = r_1, r_2, \ldots, r_Q$ be the reference mask comprising Q reference regions. Let also a region be defined as a set of pixels p. Prior to calculation of segmentation errors, the correspondence between reference and evaluated regions is established. The pairing strategy involves exclusive associations of regions r_q from the reference mask with regions s_k from the evaluated mask on the basis of region overlapping. In particular, maximization of $r_q \bigcap s_k$ was suggested in [127]. It should be noted that due to over and under-segmentation of certain parts of the image not all regions from the evaluated and the reference mask are associated.

Once correspondences are established, the spatial errors can be computed. Let $\mathcal{A} = (r_q, s_k)$ denote the set of region pairs identified using the pairing procedure, and let $\mathcal{N}_R$, $\mathcal{N}_S$ denote the sets of non-associated regions of masks $\mathcal{R}$ and $\mathcal{S}$ respectively. For every pair (r_q, s_k) from set $\mathcal{A}$ the spatial error E_q is evaluated using the criterion from [129]:

$$E_q = \sum_{p \in (r_q - r_q \cap s_k)} f_1(p, r_q) + \sum_{p \in (s_k - r_q \cap s_k)} f_2(p, r_q), \qquad \forall (r_q, s_k) \in \mathcal{A} \tag{2.1}$$

where $f_1(.,.)$, $f_2(.,.)$ are weighting functions, introduced in [129], to take into account the distance of misclassified pixels from the reference region to which they belong. The functions are depicted in Figure 2.2.

The error E_q contributed by every non-associated region r_q from the reference mask is calculated using the following formula:

$$E_q = \sum_{p \in r_q} f_1(p, r_q), \qquad \forall r_q \in \mathcal{N}_R \tag{2.2}$$

Analogically, the error contributed by every non-associated region s_k from the

mask under evaluation is calculated as:

$$E_k = \alpha \sum_{p \in s_k} f_1(p, r_q), \qquad \forall s_k \in \mathcal{N}_S \tag{2.3}$$

where, α is a scaling factor allowing different penalty for over-segmentation and, following the suggestion in [127], will be set to two in all experiments.

The final objective evaluation of segmentation accuracy is computed as the sum of error measures for all associated pairs and all non-associated regions from both reference and evaluated masks:

$$E = \sum_{\forall (r_q, s_k) \in \mathcal{A}} E_q + \sum_{\forall r_q \in \mathcal{N}_R} E_q + \alpha \sum_{\forall s_k \in \mathcal{N}_S} E_k \tag{2.4}$$

Author's Implementation of the Evaluation Tool

Although the use of weighting functions f_1 and f_2 is well justified it also substantially increases the computational cost of the evaluation. Establishing the contribution of every misclassified pixel requires a search for the closest distance to the boundary of the reference region to which it belongs. Moreover, the complexity does not scale gracefully with the size of the evaluated image, e.g. it may takes around 1 minute to evaluate a CIF size (352x288) mask on a standard PC with Pentium IV processor. Since many experiments from the next chapter require a multiple series of evaluation experiments, e.g. for different combinations of parameters, a simplified fast version of the above evaluation method, implemented by the author, is used instead.

In the simplified approach, all misclassified pixels contribute equally[4] to the error irrespective of their distance to the boundary of the reference region. To enable comparison of results for images with different sizes the segmentation accuracy is normalized by the size of the image: $E_n = E/N$ where N denotes the number of pixels in the image. The above version requires only a single scan of the evaluated mask, leading to rapid execution, and despite the simplification, still provides meaningful assessment of the results.

Also, the description from [127] leaves freedom regarding the pairing strategy. In fact, assuming exclusivity of association and using maximization of $r_q \bigcap s_k$ as the only pairing criterion may lead to pairing of large regions having only a few common pixels, if a region better matching them has been already associated with another. Therefore, in the author's implementation an additional pairing requirement $(r_q \bigcap s_k)/(r_q \bigcup s_k) > 0.5$ is imposed to prevent such associations. In other words, the evaluated and the reference regions can be paired only if their intersection is at least half the size of their union. This restriction results in a stricter evaluation criterion however was found to better correspond to human

[4] Both weighting functions are replaced by the unit function.

perception of the segmentation quality.

2.6.4 Ground-Truth Creation

Although a handful of methods allowing satisfactory relative quality evaluation of segmentation results were proposed in the past, practically no image collections with a ground truth are publicly available at the time of carrying out the research presented in this thesis.

To the best of author's knowledge, no study using a ground truth with a significant number of images of real scenes partitioned into homogenous regions is currently available in the literature. For example, in [123] mainly MPEG-4 sequences with reference masks containing objects rather then homogenous regions were used. Three MPEG sequences and one from an European IST project *Art.live* were used in [98]. Finally, the evaluation method from [127] was tested with synthetic images created using the reference textures of the VisTex database [131] and only seven, rather simple, images from the Corel gallery [132].

Since none of the above collections was sufficient or suitable for the experiments reported in the next chapter the author committed significant effort into collaboration with a colleague working in another university to create an appropriate ground-truth for a small collection of images. The collection contains 100 images from the Corel gallery [132] and 20 photographs from private collections taken using various digital cameras (including mobile phone cameras).

Reference masks were created manually in a collaborative effort using a tool developed by the author. A screenshot of the tool is shown in Figure 2.3.

Typical creation of a new reference region involves choosing a label for the new region, drawing its boundary on the original image and then filling its interior. To ensure high accuracy of the masks, the area around the cursor is automatically zoomed. Furthermore, the mask can be edited in a *Boundary Mode* in which only region boundaries superimposed on the original image are displayed allowing accurate localization of the borders on the original image and fine corrections. Additionally, the tool allows definition of objects by specification of their constituent homogenous regions into semantic objects with a hope that such semantic information may be useful in the author's future research in this field. The typical time required to manually partition an image was between 5-10 minutes, depending on user drawing skills and complexity of the scene. Although this time seems acceptable, in fact manual segmentation of several images proved tiresome. This provided very strong motivation for tools allowing semi-automatic segmentation. An example of such tool is discussed in the next chapter.

The biggest difficulty in the creation of the ground-truth proved to be the specifi-

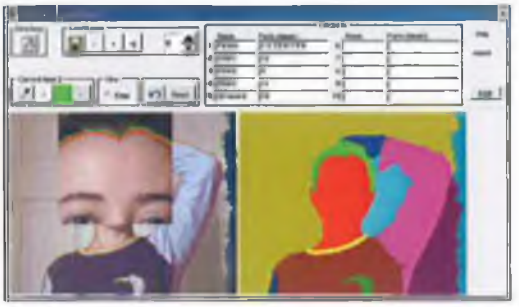


Figure 2.3: GUI of the tool for manual creation of ground-truth segmentation.

cation of rules ensuring that manual annotators provide consistent segmentation of images into homogenous regions. In fact, to the best of author's knowledge, there has been no formal description of the expected output of region-based image segmentation useful in the context of CBIR proposed in the literature. For example Mezaris et.al. simply claim that their segmentation algorithm should *"produce connected regions that correspond to the real-life objects shown in the image"* and *"provide fast segmentation with high perceptual segmentation quality"* in order to be useful in the context of content-based image and retrieval [30] .

In the case of the ground-truth described here four annotators were provided with a description of the desired segmentation rather than with a formal definition. The properties of the desired segmentation were specified as following:

- The image should be divided into as large as possible connected regions with homogenous colour or texture,
- Lighting effects should be ignored, e.g. shadows, reflections, etc.

Unfortunately, the above description leaves annotators with the problem of deciding what terms like "colour homogeneity" and "large regions" mean in the context of a particular image. In fact, the lack of a formal definition of desired results represents a major challenge for the utilization of bottom-up segmentation in applications involving general content.

However, the main objective of the next chapter is to evaluate the potential impact of geometrical information in improving automatic segmentation and investigate the problem of integration of evidence from multiple sources of information in a region merging process. Despite possible ground-truth inconsistencies, relative evaluation can still provide useful information on how well particular segmentation results correlate with the majority of the manually created masks. In other words, such ground-truth can be used to obtain comparative results, e.g. that a particular method produces results closer to the majority of masks in the ground-truth. Nevertheless clearly they cannot provide the answer to how useful a given tool is in the context of CBIR systems or any other applications.

The author is also conscious that the size of the dataset used in terms of

number of images seems be far from optimal. However, it should be also noted that the ground-truth contains 1089 reference regions which should still ensure statistically relevant results. Although, the evaluation results obtained using the above collection should be interpreted with a certain reserve, the author believes that the approach used represents a significant departure from solely presenting selected qualitative which is currently a common practice in the literature.

The full dataset together with the ground-truth segmentation is presented in appendix B.

Finally, it should be also recognized, that very recently the problem of the assessment of image and video segmentation started to attract deserved attention within the research community resulting in a several articles devoted solely to this problem [133, 134, 135, 136, 137, 138].

2.7 Discussion

Unfortunately, in many applications, image segmentation has proven to be very challenging and remained to a large extent an unsolved problem, despite being the subject of study for several decades. Although, it is well recognized that in certain scenarios utilizing segmentation can be very successful, many researchers consider reliable unsupervised general-purpose image segmentation as a hopeless goal. For instance, it is commonly accepted that the content of interest is a user dependant quality and will therefore always depend on some degree of user interaction. Consequently, the problem of inferring content of interest from generic visual data without any human intervention is ill defined. This thesis attempts to objectively discuss the prospects of utilization of image segmentation in the context of CBIR systems focusing primarily on the abovementioned issues.

Many of the above aspects of image segmentation are discussed throughout the thesis in the context of several applications. Word recognition in handwritten historical documents presented in chapter 6 utilizes a simple segmentation by thresholding approach. Chapter 3 focuses primarily on the feasibility of utilizing contour-based syntactic features for improving the correspondence of the output of automatic segmentation of images representing real world scenes to semantic objects present in these scenes but discuses also some aspects of the semi-automatic segmentation approach proposed originally in [38, 107].

2.8 Conclusion

This chapter outlined the most important aspects of image segmentation in the context of content-based image retrieval. It aimed at discussing the main

categories of approaches and giving an overview of a small selection of the most commonly used techniques and/or those particulary relevant to the research carried out by the author in chapters 3 and 7. The subject of objective evaluation of segmentation quality is also discussed in some detail providing a basis for the quantitative experiments carried out by the author in chapter 3.

CHAPTER 3

REGION-BASED SEGMENTATION OF IMAGES USING SYNTACTIC VISUAL FEATURES

THIS chapter presents a new method for segmentation of images into large regions that reflect the real world objects present in the scene. It explores the feasibility of utilizing spatial configuration of regions and their geometric properties (the so-called *syntactic features* [18]) for improving the correspondence of image segmentation produced by the well-known *Recursive Shortest Spanning Tree* (RSST) algorithm [19] to semantic objects present in the scene and therefore increasing its usefulness in various applications.

The main challenge is to asses the "strength" of the additional evidence for region merging provided by the spatial configurations of regions and their geometric properties and compare it with the evidence provided solely by the colour homogeneity criterion. Several extensions to the RSST algorithm [19] are proposed among which the most important is a novel framework for integration of evidence from multiple sources of information with the region merging process. The framework is based on the *Theory of Belief* (BeT) [23] and allows integration of sources providing evidence with different accuracy and reliability. Other extensions include a new colour model and colour homogeneity criteria, practical solutions to analysis of region's geometric properties and their spatial configuration, and a new simple stopping criterion aimed at producing partitions containing the most salient objects present in the scene.

Extensive experimentation is used to asses the generality of the proposed modifications. It is demonstrated that syntactic features can provide additional basis for region merging criteria which prevent formation of regions spanning more than one semantic object, thereby significantly improving the perceptual

quality of the segmentation.

The second part of this chapter describes a fast, easy to use and intuitive semi-automatic segmentation tool utilizing the automatic algorithm proposed in the first part.

3.1 Introduction

3.1.1 Motivation

The problem of partitioning an image into a set of homogenous regions or semantic entities is a fundamental enabling technology for understanding scene structure and identifying relevant objects. Identification of areas of an image or a video sequence that correspond to important regions is the first step of many object-based applications. Unfortunately, the problem of segmentation remains to a large extent unsolved, despite being the subject of study for several decades. Although, it is well recognized that in certain scenarios utilizing segmentation can be very successful, many researchers consider reliable unsupervised general-purpose image segmentation as a hopeless goal. The research in this chapter was motivated by the author's belief that in a given application (e.g. CBIR) all possible sources of grouping evidence should be investigated and the quality of segmentation objectively evaluated in extensive experiments. The author shares the opinion presented in [29], that segmentation, while imperfect, is an essential first step for scene understanding and that a combined architecture for segmentation and recognition is needed. For example, sensible integration of segmentation features in an image indexing context should lead to more efficient object-based indexing solutions [139].

As described in chapter 2 a large number of approaches to image or video segmentation have been proposed in the literature [140, 38, 107, 28, 33, 29, 30, 31, 18]. They can be broadly classified as either *Bottom-Up* (visual feature-based) or *Top-Down* (model-based) approaches. This chapter focuses on automatic *Bottom-Up* segmentation, which does not require developing models of individual objects, and is potentially useful in the context of CBIR systems.

Many existing approaches aim to create large regions using (relatively simple) homogeneity criteria based on colour, texture or motion. However, the application of such approaches is often limited as they fail to create meaningful partitions due to either the complexity of the scene or difficult lighting conditions. In [18], Ferran and Casas propose a new source of evidence for region merging which they term *Syntactic Features*. *Syntactic features* could be potentially important to a wide range of segmentation applications. They represent geometric properties of regions and their spatial configurations. Examples of such features include homogeneity, compactness, regularity, inclusion or symmetry.

They advocate the use of the above features in bottom-up approaches as a way of partitioning images into more meaningful entities without assuming any application dependent semantic models. In [18], they carried out a proof of concept of this idea using a *quasi-inclusion criterion*. Other approaches attempting to integrate geometric properties of regions include [64, 62, 65, 66, 67, 18, 68] – see also section 2.3.3 for brief overview.

Although the idea of using geometric properties within a region merging process to create more meaningful partitions is not new, the approaches currently available in the literature integrate the geometric features in a rather ad-hoc manner [62, 18, 68] typically developed to suit only particular geometric properties and not easily extendible to other features. Also, to the best of the author's knowledge, currently there is no objective evaluation available in the literature to support the usefulness of any such features.

Syntactic features can serve as an important source of information that helps to close the gap between low level features and a semantic interpretation of the scene. Of course, probably they do not provide evidence "strong enough" to reliably group regions with very different colour, texture and motion into complex semantic objects. However, it can be argued that they can significantly improve the performance of automatic bottom-up segmentation techniques bringing them closer to the semantic level by allowing creation of large meaningful regions enhancing their usefulness in CBIR systems. In other words, creation of larger, more meaningful regions should facilitate extraction of more meaningful local features useful for content based indexing and retrieval.

Furthermore, syntactic and semantic sources of information are not mutually exclusive, and if available they could, and probably should, be used together to complement the segmentation procedure. Syntactic features can be useful even in the case of supervised scenarios where they can facilitate more intuitive user interactions e.g. by allowing the segmentation process make more "intelligent" decisions whenever the information provided by the user is not sufficient.

This chapter focuses on the integration of syntactic features into the image segmentation process and proposes a practical solution for fully automatic, robust and fast segmentation that is potentially useful for object-based multimedia applications. The objective is to automatically segment images into large regions generally corresponding to objects or parts of objects whilst avoiding the creation of regions spanning more than one semantic object. The potential of incorporating syntactic features into the segmentation process to meet this objective and allow satisfactory results with challenging real world images is evaluated by extensive experimentation.

3.1.2 Chapter Structure

The remainder of this chapter is organized as follows: the next section describes in detail the *Recursive Shortest Spanning Tree* (RSST). Then, an overview of the proposed segmentation approach is presented in section 3.3. Next, the training and evaluation methodologies adopted throughout the chapter are described in section 3.4. Section 3.5 discusses and evaluates several improvements to a region's colour representation and the colour homogeneity criteria used during the region merging process. Section 3.6 discusses selected syntactic features, proposes practical solutions to analysis of a region's geometric properties and their spatial configuration and discusses the major difficulties related to the integration of syntactic features within a region merging framework. The novel framework for integration of evidence from multiple sources of information within the region RSST algorithm based on the *Theory of Belief* (BeT) is proposed in section 3.7. A new simple stopping criterion aimed at producing partitions containing the most salient objects present in the scene is proposed and evaluated in section 3.8. Future research is discussed in section 3.9 and the properties of the final complete approach are discussed in section 3.10. Section 3.11 describes a fast, easy to use and intuitive semi-automatic segmentation tool utilizing the automatic algorithm proposed in the first part of the chapter. Conclusions are formulated in section 3.12.

3.2 RSST Segmentation Framework

The *Recursive Shortest Spanning Tree* (RSST) algorithm is selected for the incorporation of syntactic features as it provides a convenient framework for integration of a region's geometric properties, and also due to its speed.

The RSST algorithm was first proposed by Morris et al. in [19] for image segmentation and edge detection problems. It was later extended to colour images in [94]. It was also shown to be useful for image coding [63]. One of its application is object segmentation, extraction and tracking [85]. The COST 211quat project adopted RSST as the grouping method in the *Analysis Model* (AM) [140]. In [141] RSST was utilized for video-object segmentation based on an affine motion model. The most recent work regarding fast implementation of the algorithm can be found in [106, 95]. Similar region merging algorithms were studied extensively in the work of Salembier and Garrido for creation of *Binary Partition Trees* (BPT) [107].

The RSST segmentation starts by mapping the input image into a weighted graph [19], where the regions (initially pixels) form the nodes of the graph and the links between neighboring regions represent the merging cost, computed according a selected homogeneity criterion. Merging is performed iteratively. At

each iteration two regions connected by the least cost link are merged. Merging two regions involves creating a joint representation for the new region and updating its links with its neighbors. The process continues until a certain stopping condition is fulfilled, e.g. the desired number of regions is reached or the value of the least link cost exceeds a predefined threshold.

The homogeneity criterion used to define merging cost and the stopping criterion play a key role in obtaining the desired segmentation[1]. Since at each iteration only two regions connected by the least cost link are merged the merging order is exclusively controlled by the function used to compute the merging cost. Let r_i and r_j be two neighboring regions. The merging cost is based solely on a simple colour homogeneity criterion defined as [142, 140, 85]:

$$C_{orig}(i,j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2^2 \frac{a_i a_j}{a_i + a_j} \tag{3.1}$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the mean colours of r_i and r_j respectively, a_i and a_j denote region sizes, and $\|\cdot\|_2$ denotes the $\mathcal{L}_2$ norm. It should be noted that the second factor encourages the creation of large regions[2] by decreasing the cost of merging for small regions. Alternative merging criteria based on colour can be found in [19, 143, 41, 18].

Various stoping criteria have been used in the past including the desired number of regions [140], thresholding of the least link cost [19, 144] and minimum acceptable value of PSNR between the original and segmented image [41] (reconstructed using mean region intensities). However, none of the above criteria can produce large regions and at the same time ensure that the merging process will stop before merging important semantic boundaries.

The RSST algorithm is fast and capable of producing meaningful over-segmentation (typically more than 100 regions) useful in many applications – see for example [140]. However, the above colour homogeneity criterion is insufficient to produce meaningful results in applications where further simplification of the scene is required, i.e. it is incapable of producing large regions with a high probability of corresponding to semantic notions [38]. Very often the produced regions do not represent the true semantic structure of the scene.

One of the reasons for poor performance is utilization of identical simple homogeneity criterion at all stages of the merging process. In other words, the same merging criterion is used at the level of pixels and at stages where regions contain thousands of pixels.

Summarizing, the RSST is selected for the incorporation of syntactic visual features due to the following reasons:

[1]The stopping criterion may not be required in certain applications, e.g. revealing hierarchical structure of the scene by creating *Binary Partition Tree* (BPT) [107].

[2]In fact, the factor encourages creation of regions similar in size.

- It is based solely on colour homogeneity and, when used to produce segmentations with a small number of regions, often creates regions with geometric properties which are rare in the real world. Utilization of regions' geometric properties promises an intuitive remedy to this problem,
- It provides convenient access to geometric properties of every region at each iteration,
- It is capable of producing meaningful over-segmentation with well localized region contours,
- It is extremely fast, especially when compared with other *graph theoretic* algorithms, i.e. it takes less than 3 seconds on a standard PC (Pentium III 600MHz) to segment an image of CIF size (352x288),
- It is easy to implement.

3.3 Overview of the Proposed Approach

To overcome the limitation of the original RSST algorithm, both, the merging cost measure and the stoping criterion are modified and extended, e.g. by an additional set of features. Additional geometric properties are integrated in the RSST merging framework by employing new measures for calculating the cost of merging two neighboring regions. The new features control the merging order and provide a basis for a reliable stopping criteria. The new merging order is based on evidence provided by different features (colour and geometric properties) and fused using an integration framework based on the *Theory of Belief* (BeT) [22, 23]. The impact of the above modifications on segmentation quality is demonstrated by rigorous experimentation using the evaluation methodology described in section 2.6.3.

The segmentation process is divided into two stages. The initial partition, required for structure analysis, is obtained by the RSST algorithm with the original merging criterion implemented as in [140] since it is capable of producing good results when regions are uniform and small. This stage ensures the low computational cost of the overall algorithm and also avoids the application of syntactic features to small regions with meaningless shape. This initial stage is forced to stop when a predefined number of regions is reached (100 for all experiments presented in this chapter). Then, in the second stage, region representation is extended by region boundary and, depending on a variant of the approach, by a new colour model. Subsequently, homogeneity criteria are re-defined based on colour and syntactic features and the merging process continues until a certain stopping criterion is fulfilled. The new merging order is based on evidence provided by different features (colour and geometric properties) fused using the

proposed integration framework. It should be stressed that all extensions proposed in the following sections consider mainly the second stage of merging when regions already have meaningful shape.

3.4 Methodology Used for Training and Testing

All solutions proposed in this chapter are tuned and evaluated using the spatial accuracy error measure and the collection of images with ground-truth discussed in sections 2.6.3 and 2.6.4 respectively.

3.4.1 Training and Test Corpus

Recall from section 2.6.3 that the collection of manually segmented images contains 100 images from the Corel dataset and 20 images selected from various sources such as keyframes from well known MPEG-4 test sequences, digital camera, and mobile phone camera segmented by four annotators. The full dataset together with the ground-truth segmentation is presented in appendix B.

The images from the Corel dataset are typically well composed and contain objects with distinctive colour while the images from other sources are often taken with difficult lighting conditions and contain objects with similar colour. Although the images from the Corel collection appear somehow to be easier to segment due to the fact that different objects have distinctive colour, they also present the ultimate challenge to any approach utilizing features other than colour. In other words the aim is to integrate the additional information, e.g. region's geometric properties, in such a way that the segmentation of the images with difficult lighting conditions is improved but also at the same time the good performance in the case of images containing objects with distinctive uniform colour is retained.

Taking into account the relatively small number of images available the proposed extensions are evaluated by two-fold cross validation. The dataset is divided into two subsets (let us denote them as subset A and B), each containing 60 images, i.e. 50 images from the Corel dataset and 10 from the other sources. Each time one of them is used for parameter tuning, the other subset is used as the test set. The final result is computed as the average error from the two trials. Also, all images showing selected segmentation results are generated during the cross-validation. However, for clarity of presentation, whenever a particular parameter is discussed its influence on segmentation results is demonstrated on the entire image collection, e.g. all plots are generated using the entire collection.

3.4.2 Development Methodology

For clarity, various merging costs and stopping criteria are developed and tested separately. First, new merging costs controlling the merging order are proposed in sections 3.5, 3.6, and 3.7. They are evaluated by stopping the region merging process based on the number of regions available from the ground-truth and an assessment of the resulting segmentation mask in terms of spatial accuracy error. This helps to simplify the integration of various features within the merging cost and preliminary evaluation of their generality and usefulness and at the same time to avoid the complex influence of a stopping criterion on the results. Once useful features and their meaningful integration for controlling the region merging order are identified, various stopping criteria are discussed in section 3.8.

3.5 Colour Homogeneity

Colour homogeneity is a basic criterion for partitioning in practically all image segmentation approaches which can be found in the literature [29, 30, 31, 59, 18]. This section discuses some modifications to the original colour homogeneity criterion used within the RSST algorithm [140]. The main aim is to ensure that the most appropriate colour homogeneity criterion is employed before integrating other, potentially less influential, features.

3.5.1 Colour Spaces

Many resent publications in the area of colour segmentation advocate utilization of the so called perceptually uniform colour spaces: CIE LUV and its improved version CIE LAB [28, 29, 30]. In the remainder of this chapter, colour is represented using the CIE LUV colour space, which is often considered as a good trade-off between the approximation of perceptual uniformity and the cost of conversion from commonly used RGB or YUV colour spaces.

3.5.2 Optimized Colour Homogeneity Criterion

Despite its simple form, the merging cost measure defined in equation 3.1 realizes relatively complex merging rules. It gives merging preference to small regions, even if the colour is significantly different, by "punishing" merging of large regions. This illustrates the implicit assumption that all homogenous regions (or objects) present in the scene should be of equal sizes. In other words, when segmenting an image with uniform colour, at each iteration, all regions will have approximately equal size.

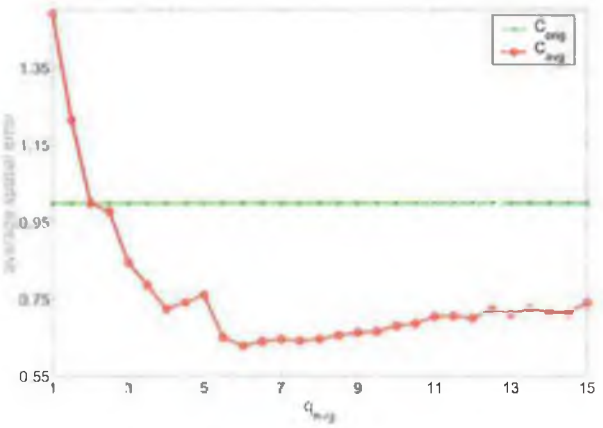


Figure 3.1: Influence of parameter q_{avg} on the spatial segmentation error.

Intuitively, the strong preferences of merging of small regions even for regions with significant colour differences is one of the main reasons for the poor performance of the RSST algorithm in the final merging stages. Moreover, it is common knowledge, that even uniform colour spaces do not guarantee that distances between very different colours agree with human perception of uniformity.

For example, lets imagine two links between pairs of regions, both with equal costs according to formula 3.1, but the first one between two large regions with quite similar colour and the second one between two smaller regions but with more distinct colour. We would like to introduce a modification to the equation 3.1 so that the contribution of the colour distance to the overall merging cost would be greater comparing the contribution induced by their sizes and therefore resulting in a smaller merging cost for the first link compared to the second.

It is proposed here that a more suitable balance between colour distance and the size dependent scaling factor can be obtained using a new merging cost defined as follows:

$$C_{avg}(i,j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2^{q_{avg}} \cdot \frac{a_i a_j}{a_i + a_j} \tag{3.2}$$

where parameter q_{avg} is found experimentally using the training collection and evaluated later on the test collection.

Figure 3.1 shows the evaluation results for the approach with the modified merging cost measure C_{avg} utilized in the second merging stage for different values of parameter q_{avg} obtained for the entire image collection using the optimal number of regions as the stoping criterion. It can be seen that values of parameter q_{avg} greater than 4.0 lead to a significant decrease of average spatial accuracy error. Interestingly, for $q_{avg} = 4.0$ the above modification is somehow similar to the merging criterion proposed recently in [144] for segmentation of

luminance images[3]. In our case the lowest segmentation error is obtained by $q_{avg} = 6$ but the difference may be specific to the collection, evaluation method or stopping criterion used.

The significant improvement by this trivial modification is also confirmed by cross-validation where the average spatial segmentation error is reduced from 1.00 to 0.64. It should also be stressed that the improvement is consistent over the majority of image classes and was confirmed by manual inspection of the results. To provide as illustrative results as possible, an equal number of examples showing cases where the modification leads to a significant improvement (the first 5 rows) and examples where the modification resulted in larger spatial segmentation error (the last five rows) than the original RSST approach are shown in Figure 3.4 (3rd and 4th column). It can be seen that unlike the original approach the modified version of the merging criterion allows creation of large regions which easily explain the significant reduction of average spatial segmentation error.

3.5.3 Extended Colour Representation

In the initial merging stage average values of colour components are sufficient in representing a region's colour. However, as segmentation progresses and regions grow, average values may become unsuitable for effective characterization of their colour. To ensure that this simplification does not compromise the merging order, an extended alternative colour model is proposed and evaluated. A fine and compact representation of region's colour motivated by the so-called *Adaptive Distribution of Colour Shades* (ADCS) proposed in [31] is used during the second merging stage. In this model, each region contains a list of pairs of colour/population[4] which can more precisely represent its complex colour variations.

For example, imagine an extreme case where a chessboard like object is initially pre-segmented into white and black squares. The ADCS representation of the region representing the whole chessboard is capable of accurately represent its colour, e.g. by two pairs of colour/population: one for black and one for white colour, while an average colour simply would not allow distinguishing it from a completely grey object.

The ADCS is utilized in the proposed approach in the following way. After the initial segmentation each region's colour representation (average values of colour components) is converted into the ADCS representation. Such conversion

[3]The solution proposed in [144] was motivated as being the geometrical mean of the criterion from equation 3.1 (when applied to luminance component only) and a criterion computed as the squared difference between regions' mean luminance values.

[4]Where population refers to the ratio between the number of pixels with such colour and the total size of the region.

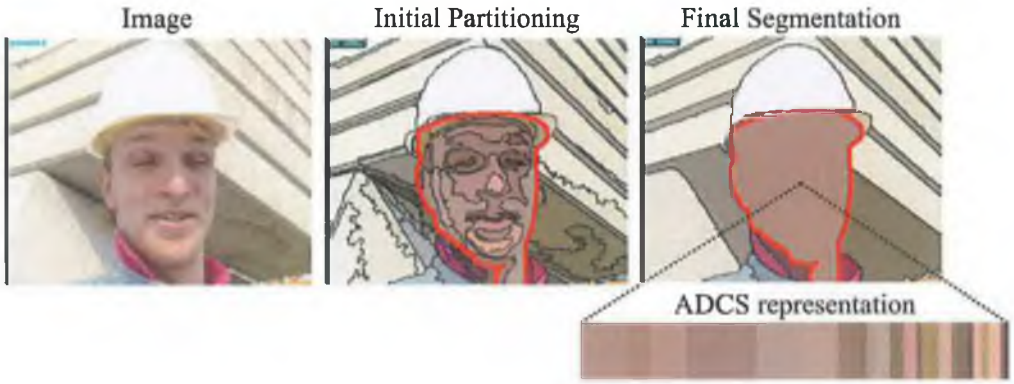


Figure 3.2: Example of ADCS representation.

is trivial as at this stage a region's ADCS list contain only one colour/population pair representing their average colour and size. Subsequently, in the second merging stage, whenever two regions are merged, the joint representation is created by concatenating lists from both regions. In other words, at the end of the segmentation process each region's colour is represented by an ADCS list of colour/population pairs, each corresponding to one of its composing initial regions and representing its average colour and ratio between its size and the size of the final region. This extended colour representation has the advantage of low requirements for both storage and for updating the merging cost. Also, no prior assumption needs to be made about the underlying colour distribution. Figure 3.2 shows an example of the extended colour representation for a large face region.

It has been shown in [31] that two such ADCS representations can be efficiently compared using *Quadratic Distance* [145]. Therefore, the colour difference between regions r_i and r_j is calculated as the quadratic distance $d^2_{quad}(\mathbf{I},\mathbf{J})$ between their corresponding colour distributions characterized by two ADCS representations $\mathbf{I} = (c^I_1, p^I_1), \ldots, (c^I_v, p^I_v), \ldots, (c^I_V, p^I_V)$ and $\mathbf{J} = (c^J_1, p^J_1), \ldots,$ $(c^J_w, p^J_w), \ldots, (c^J_W, p^J_W)$ consisting of V and W pairs of colour/population respectively:

$$d^2_{quad}(\mathbf{I},\mathbf{J}) = \mathbf{I}^T\mathbf{A}^I\mathbf{I} + \mathbf{J}^T\mathbf{A}^J\mathbf{J} - 2\mathbf{I}^T\mathbf{A}^{IJ}\mathbf{J} \tag{3.3}$$

where matrices $\mathbf{A}^I$, $\mathbf{A}^J$ and $\mathbf{A}^{IJ}$ express colour similarities between, respectively, $\mathbf{I}'s$ colour pairs with themselves, those of $\mathbf{J}$ with themselves and those of $\mathbf{I}$ with those of $\mathbf{J}$. The colour similarities are computed based on the Euclidean distance, e.g. matrix $\mathbf{A}^{IJ} = [a_{c^I_v c^J_w}]$ is calculated as:

$$a_{c^I_v c^J_w} = 1 - \frac{d_{c^I_v c^J_w}}{d_{max}} \tag{3.4}$$

where $d_{c^I_v c^J_w}$ is the Euclidean distance between two colours c^I_w and c^J_v represented in the LUV colour space and d_{max} denotes the maximum of this distance in the

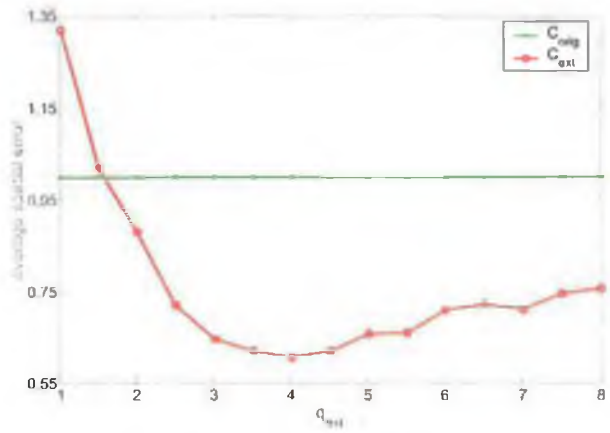


Figure 3.3: Influence of parameter q_{ext} on the spatial segmentation error.

colour space.

The cost of computing the quadratic distance depends strongly on the number of colour shades representing each region ($O(N_{\mathrm{I}}N_{\mathrm{J}})$ complexity). However, the computation can be speeded-up by pre-computing the first two terms of (3.3) when each representation is created (i.e. after merging) such that only the last cross term needs to be evaluated when representations are compared. In the case of the proposed segmentation approach, the additional cost of computing quadratic distance is limited by the number of regions obtained in the initial stage and is practically negligible. Nevertheless, if necessary, the number of colours representing each region can be further reduced by an additional clustering performed each time a new region representation is created or by global colour quantization as in [31].

By analogy to equation 3.2 the new merging cost based on the extended colour representation is defined as:

$$C_{ext}(i,j) = \left[d^2_{quad}(\mathbf{I},\mathbf{J})\right]^{q_{ext}} \cdot \frac{a_i a_j}{a_i + a_j} \tag{3.5}$$

Figure 3.3 shows the average spatial segmentation error obtained for the entire collection with the extended colour representation for different values of parameter q_{ext}. It can be seen that the lowest average spatial segmentation error is obtained for values of q_{ext} close to 4.0. The improvement is also confirmed by the cross-validation where the average segmentation error is reduced from 1.00 (obtained by the original RSST approach [140]) to 0.63. However, it should be also noted that although the extended representation improves the results compared to the original RSST algorithm the improvement is very close to the one obtained by the simple modification of the original homogeneity criterion (equation 3.2) proposed in the previous section – see Figure 3.1 (5th column). The hypothesis that the influence of the extended colour representation is more pronounced when additional homogeneity criteria, other then colour, are incor-

porated is evaluated in section 3.7.5.

In the current approach only spatial colour homogeneity is considered. Integration of texture homogeneity, e.g. by adopting the approach from [30], incorporation of the so-called *Boundary Melting* approach which favors merging of regions with low magnitude of colour gradient along their common boundary [105] and utilization of shape homogeneity, i.e. shape similarity (a potentially very important syntactic feature) will be part of the author's future research in this area. Since integration of multiple feature within the RSST merging algorithm is a challenging problem the remainder of this chapter considers instead a new flexible integration framework and demonstrates its application to the difficult problem of integration of colour homogeneity with a region's geometric properties into single region merging criterion.

3.6 Syntactic Visual Features

As already discussed in the previous chapter several approaches have attempted to incorporate some kind of region geometric properties into image segmentation. The work described here is primarily motivated by the proof of concept of using the so-called *Syntactic Visual Features*, and in particular a *Quasi-Inclusion Criterion*, for creation of *Binary Partition Tree* (BPT) advocated by Ferran and Casas [18].

Syntactic visual features include homogeneity (e.g. shape homogeneity), compactness, regularity, inclusion and symmetry [18]. This section discuses a subset of the above properties together with issues related to their measurement and some general issues related to their integration within the RSST framework.

The employment of syntactic features for automatic segmentation appears to be well motivated. Since most "real world" objects are regular, i.e. since their shape complexity tends to be rather low, analysis of regions complexity could provide additional evidence for merging. Also, it is well known that objects tend to be compact, i.e. they exhibit adjacency of their constituent parts [30] therefore analysis of adjacency could also provide additional evidence for merging. Furthermore, objects may contain holes but more often small regions totally included inside larger regions correspond simply to less significant parts of the larger semantic object. Therefore, merging the included region would lead to simplification of the analyzed scene without merging important semantic boundaries. Also symmetry of objects or their parts and similarities between parts of the same object appear to be rather common in "real world" scenes and therefore could provide evidence for more meaningful merging.

Although compactness, regularity and inclusion are often discussed separately [18]

Figure 3.4: Segmentations obtained by different colour homogeneity criteria (the required number of regions is known). The 3rd, 4th and 5th columns show results obtained by the original RSST approach (C_{orig}), the modified merging criterion (C_{avg}) and the merging criterion based on the extended colour representation (C_{ext}) respectively.

in fact there is a strong overlap between these concepts. For example, it is rather difficult to provide a definite example of an object which has complex shape and at the same time is compact. Also, one could consider inclusion as an extreme case of adjacency between regions. In fact it appears impossible to propose completely independent measures of regularity, compactness and inclusion when considering pairs of regions. Such strong overlaps between geometric properties contributes additional difficulty to their integration into the segmentation process.

The following sections propose two measures for quantification of the quality of merging based on geometrical properties and spatial configuration of pairs of neighboring regions (namely adjacency and shape complexity) and discuss in detail their overlaps and potential role in region merging.

3.6.1 Adjacency

It is a well known fact that "real world" objects tend to be compact, i.e. they exhibit adjacency of their constituent parts. In the case of the RSST algorithm, adjacency of parts is imposed intrinsically during link initialization where regions are considered adjacent when they have at least one adjacent pixel. It is proposed here that this binary notion of adjacency be extended by defining an adjacency measure:

$$C_{adj}(i,j) = 1.0 - \frac{l_{ij}}{\min\{l_i, l_j\}} \tag{3.6}$$

where l_{ij} is the length of the common boundary between regions and l_i and l_j are their perimeter lengths.

Although simple, the above measure provides quite rich information about geometrical relations between both regions. Values of $C_{adj}(i,j)$ close to zero indicate almost total inclusion (one region is almost completely surrounded by another) [18] whereas values close to one point to weak adjacency and therefore provide strong evidence against merging. It should be also noted that low values of $C_{adj}(i,j)$ can result from high local complexity (jaggedness) of the common border between the two neighboring regions. Therefore using low values of $C_{adj}(i,j)$ as a good indication for merging may result in preferences for merging regions for which the common border is jagged.

It should also be mentioned that a measure of the so called *quasi-inclusion* for improving the order of region merging based on the concept of convex-hull was proposed in [18]. Quasi-inclusion may play very similar role to the adjacency measure defined here. Since quasi-inclusion should not be influenced by the contour jaggedness it could be very interesting to compare their performances.

As already discussed, total inclusion of a small region inside another should provide strong evidence supporting merging the two regions. In the proposed

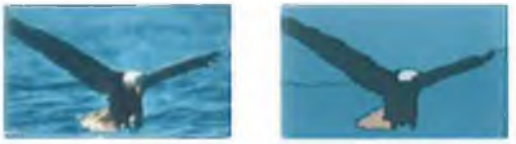


Figure 3.5: Over-segmentation of occluded background caused by favouring of merging of adjacent regions [68].

approach merging of included regions is encouraged by incorporation of the adjacency measure and merging of small regions is favoured by the size dependent factor in the colour homogeneity measures. The author's observations indicate that merging of small included regions is sufficiently encouraged by these two measures.

However, favoring the creation of compact (and simple) regions may also discourage formation of large regions corresponding to occluded background – see for example Figure 3.5. It is clear that in this particular case the adjacency information should be ignored due to strong colour similarity. Examples like the above provide strong motivation for research of an advanced way of fusing the information which would allow modeling of doubt and taking into account the reliability of different sources of evidence.

3.6.2 Regularity (low complexity)

Further evidence for region merging is provided by the fact that the shape complexity of most objects tends to be rather low. Both global and local shape complexities can be explored in this regard. However, for the reasons described below, only global shape complexities are used in this work.

The term "complexity", although often used in the context of shape analysis, is not well defined. Probably the most advanced work towards better understanding of perception of shape complexity can be found in [146] where the complexity measure is based on entropy measures of global distance and local angle, and a measure of shape randomness[5]. To maintain the simplicity of the set of features tested here a more straightforward measure of complexity is tested instead. Depending on the performance of this simple approach, incorporation of more advanced measures of complexity may be part of the author's future research in this field.

Global Contour Complexity

Intuitively, more complex shapes have a longer perimeter in relation to their area compared to simple shapes. Therefore global shape complexity x_i of a region r_i can be measured as the ratio between its perimeter length l_i and the square root

[5]Unfortunately, the final measure requires regularization parameters set in an ad-hoc manner.

of its area a_i. In fact the above ratio is often used in the literature as a measure of compactness and it is also dependent to a certain degree of the local contour complexity.

Average changes of global shape complexity for a pair of neighbouring regions r_i and r_j caused by their merging is measured in the proposed scheme as:

$$C_{cpx}(i,j) = \frac{x_{ij}}{\left[\frac{a_i x_i + a_j x_j}{a_i + a_j}\right]} \tag{3.7}$$

where x_{ij} denotes shape complexity of a hypothetical region formed by merging r_i and r_j. The denominator in the above formula can be seen as an estimation of the average complexity of both regions before merging. Low values of $C_{cpx}(i,j)$ provide evidence for merging the two regions e.g. values below 1.0 indicate that the complexity of the joint region is lower than the average of complexities of the two original regions.

However it should be also noted that there is a certain overlap with the adjacency measure. The shape of the merged region r_{ij} is more likely to be simpler than the shape of the constituent regions r_i and r_j if they share a longer common border.

Local Contour Complexity

Objects with very jagged contours are rather rare in the "real world". Moreover the RSST algorithm tends to produce jagged contours at the border separating regions which should be merged [38, 62]. Therefore excessive jaggedness could be used for detection of such "artificial" borders and provide evidence for merging them.

However, contours with corners with high value of curvature are also created at the meeting point of two different convex objects interpenetrating each other. This phenomenon is often referred in the literature as the *Principle of Transversality* [147]. An attempt to distinguish such salient corners created by interpenetrating objects from the corners belonging to a jagged boundary artificially created by the merging algorithm and to utilize such perceptual information within the RSST-based merging was presented in [62, 63]. Disappointingly the performance of the approach was demonstrated only on a single image (Lenna).

Although in the proposed approach, detection of jaggedness of the common border using two different methods (one based on the Moravec operator [148] and another based on contour curvature) and explicit utilization of such information within the proposed framework led initially to plausible results for many images (see the author's initial approach in [68]), extensive experiments revealed also that it can cause some erroneous merges. Therefore several problems have to be addressed before the local complexity of region's contours can be explicitly utilized for reliable region merging: (i) choice of correct scale at which the lo-

cal complexity is measured, (ii) it is not clear whether isolated corners with high curvature or a measure of the overall jaggedness of the common border provides more useful evidence for merging, (iii) a reliable way of distinguishing between artificially created jagged borders from the high curvature points resulted by interpenetrating objects.

The above problems together with the fact that both measures C_{adj} and C_{cpx} proposed in this thesis already capture well the jaggedness of the common borders motivated the fact that the local complexity of contours is not explicitly used in this work.

3.7 Integration of Evidence from Different Sources

Meaningful integration of evidence from different sources into a single merging cost has proven to be a non-trivial task. Intuitively, colour homogeneity should provide much more reliable evidence than any other geometric properties, i.e. the new features described in the previous sections should offer only an auxiliary evidence. Therefore it is necessary to take into account the reliability of different sources of information when integrating them within the region merging algorithm. Moreover, certain geometric properties in some cases cannot be measured and therefore cannot provide any evidence supporting the region merging hypothesis, e.g. changes in shape complexity caused by merging two regions are meaningless if the hypothetical region created by the merging shares a significant part of its contour with the entire image as explained later in section 3.7.4. In other words, the integration framework should be able to combine the evidence for merging or not from various sources but also should be able to accept that certain measurements may not be precise (doubtful) or even "unknown" in some cases. Both considerations provide strong motivation for adopting the *Belief Theory* [20, 21], which offers a convenient model for handling imprecise and uncertain information, and particulary its revised version proposed recently by Smets [22, 23] called *Transferable Belief Models* (TBM), as the main framework for integration of multiple features within the proposed region merging algorithm.

3.7.1 Belief Theory

Traditionally, uncertain data was dealt with the use of probability theory, which in some cases has proven to be inadequate [149]. More recently, other models have been proposed for handling imprecise knowledge, e.g. *Theory of Fuzzy Sets* [150], *Possibility Theory* [151], or uncertain information, e.g. *Belief Theory* [20, 21, 22, 23]). This chapter examines the possibility of using *Belief Theory* to combine evidence provided by colour homogeneity and the syntactic features

for improving the performance of the RSST-based segmentation.

Belief Theory(BeT) (or *Theory of Belief Functions*) was introduced by Dempster [20] and Shafer [21] and later revised by Smets [22, 23] who proposed *Transferable Belief Models* (TBM). The BeT is adopted in this work as it represents a powerful tool for combining measures of evidence. The main features of Dempster-Shafer's theory are:

- Convenient management of uncertainties by explicitly modelling the doubt;
- The ability of taking into account reliability of information sources;
- Distinct representation of ignorance, imprecision and conflicting information;
- Convenient incorporation of statistical and/or expert knowledge.

In other words, BeT appears more general and more flexible than the commonly used probability model and allows representation of states of knowledge and opinions that probability models cannot [152].

Typically Dempster-Shafer's theory is used for information fusion in decision making problems, e.g. medical diagnosis or recognition/classification [153, 154]. Here, it is used to compute merging cost utilizing evidence provided by multiple sources such as colour homogeneity and the syntactic features proposed in the earlier sections.

3.7.2 Dempster-Shafer's Theory: Background

Frame of Discernment

Assume a definition of a finite set of N exclusive and exhaustive hypotheses Ω called a frame of discernment such as:

$$\Omega = \{H_1, \ldots, H_N\} \tag{3.8}$$

Basic Belief Assignment

Also, the power set 2^Ω of Ω be composed of 2^N propositions A of Ω, such as:

$$2^\Omega = \{\emptyset, \{H_1\}, \ldots, \{H_N\}, \{H_1 \cup H_2\}, \{H_1 \cup H_3\}, \ldots, \Omega\} \tag{3.9}$$

Finally, let us define a set of U independent evidence sources $S_1, \ldots, S_U$. A piece of evidence brought by a source S_u on a proposition A is modelled by the belief structure m_u called the *Basic Belief Assignment* (BBA) defined as:

$$m_u : 2^\Omega \longrightarrow [0, 1] \tag{3.10}$$

with $m_u(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m_u(A) = 1$.

In other words, a model m_u associates a belief mass to each of the symbols of 2^Ω for each value of the evidence from S_u .

Combining Information Sources

Evidence theory provides a convenient framework for fusing or combining several sources of evidence. A commonly used operator is the orthogonal sum also called the Dempster's rule of combination [22, 155]. According to this operator, which is commutative and associative, the combined belief structure $m^\oplus$ is defined by:

$$m^\oplus = m_1 \oplus m_2 \oplus \ldots \oplus m_U \tag{3.11}$$

where for two sources of information S_u and S_v the combined belief structure $m^\oplus$ is defined as:

$$\forall A \subseteq \Omega \qquad m^\oplus(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_u(B) \cdot m_v(C) \tag{3.12}$$

where K can be interpreted as a measure of the conflict between the sources to be combined ($m^\oplus(\emptyset)$) and is defined as:

$$K = \sum_{B \cap C = \emptyset} m_u(B) \cdot m_v(C) \tag{3.13}$$

Although, due to the assumption $\sum_{A \subseteq \Omega} m^\oplus(A) = 1$, the normalization of $m^\oplus(A)$ by $(1 - K)$ seems natural, it has been also criticized by some authors [156]. It has been shown in [155] that the above normalization corresponds to assimilation of the *closed-world* postulate (the evidence may support only propositions from Ω), while combining sources without normalization corresponds to assimilation of the *open-world* assumption (the evidence might support propositions from outside Ω). The utilization of the normalization is further discussed in the next section in the context of the region merging problem.

Taking Into Account Source Reliability

When source S_u is considered not completely reliable the confidence in this source can be attenuated by a factor $\alpha_u \in [0, 1]$. This procedure is often referred to as *discounting* [21]. The new attenuated structure m_u^α is then defined as:

$$m_u^{\alpha_u}(A) = \alpha_u m_u(A) \qquad \forall A \neq \Omega \tag{3.14}$$

and

$$m_u^{\alpha_u}(\Omega) = 1 - \alpha_u + \alpha_u m_u(\Omega) \tag{3.15}$$

3.7.3 Application of Dempster-Shafer's Theory to the Region Merging Problem

This section describes the major aspects of the application of BeT to the problem of integration of multiple features within the region merging process: (i) definition of the frame of discernment, (ii) general form of the belief structure used to model the way a piece of evidence is brought by a source on a proposition, (iii) taking into account source reliability, (iv) combination of beliefs from multiple sources and (v) a new merging cost formula based on the combination of beliefs.

Frame of Discernment

Considering the region merging problem, let the frame of discernment Ω represent a set of two hypothesis, namely $MERGE$ and $DONTMERGE$, which are exhaustive and exclusive:

$$\Omega = \{MERGE, DONTMERGE\} \tag{3.16}$$

Since in case of the region merging problem the *closed-world* postulate appears more intuitive than the *open-world* postulate[6] the *closed-world* assumption and consequently the assimilation of the normalization in equation 3.12 are adopted in the proposed framework.

Basic Belief Assignment

The power set 2^{Ω} of Ω is composed of 4 propositions A:

$$2^{\Omega} = \{\emptyset, \{MERGE\}, \{DONTMERGE\}, \{MERGE \cup DONTMERGE\}\} \tag{3.17}$$

The proposition $\{MERGE \cup DONTMERGE\}$ is also referred to in this thesis as doubt $\{DOUBT\}$.

Let us define a set of U independent[7] evidence sources $S_1, \ldots, S_U$ where S_1 always denotes one of the colour homogeneity criterion and $S_2, \ldots, S_U$ denote a set of other properties associated with a link between two regions. A piece of evidence (measurement) C_u brought by a source S_u on a proposition A is modelled by the belief structure m_u formally defined by formula 3.10. In other words, for each measure C_u (e.g. a colour homogeneity or a syntactic feature) model m_u maps each value of C_u into a belief on a proposition (singleton or composed hypothesis) of 2^{Ω} .

[6] In the proposed framework the evidence should support propositions only from Ω.

[7] The assumption of independency of sources is further discussed in section 3.7.5 and section 3.9.

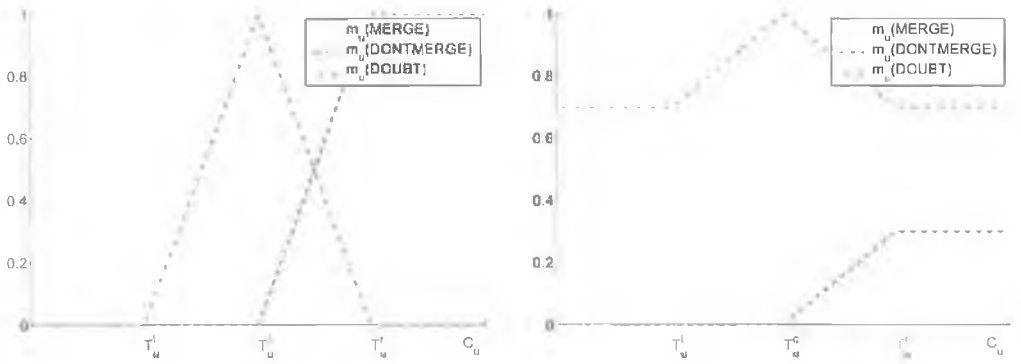


Figure 3.6: General form of BBA functions ($DOUBT = MERGE \cup DONTMERGE$): (a) - BBA for a reliable source of information, (b) - Discounted BBA for a source of limited reliability, i.e. increased doubt.

Theoretically the belief structures (BBA) could be obtained based on both statistical or expert knowledge. However, automatic derivation of BBA based solely on the training data is a nontrivial problem and is currently investigated by several researchers [153]. Therefore in this work the models are based on both expert knowledge and statistical knowledge obtained from the training collection similar to the methodologies proposed in [154] for training of static posture recognition system and in [152] for training of facial emotions classifiers.

The general form of such models is shown in Figure 3.6a. Let us denote the neutral value of measure C_u as T_u^c, i.e. the value of C_u which does not provide evidence either for merging or not merging. Therefore, in the proposed belief structure for value of measure C_u equal to T_u^c the entire mass of belief is associated with $DOUBT$ ($MERGE \cup DONTMERGE$) while the masses of belief associated with propositions $MERGE$ or $DONTMERGE$ are equal to zero. As the value of C_u decreases(increases) from T_u^c to $T_u^l(T_u^r)$ the mass of belief associated with $DOUBT$ decreases while the mass of belief associated with $MERGE(DONTMERGE)$ increases. Finally, for values of C_u below(above) $T_u^l(T_u^r)$ the entire mass of belief is associated with $MERGE(DONTMERGE)$ and the mass of belief associated with $DOUBT$ is equal to zero.

Taking Into Account Source Reliability

Figure 3.6b shows an example of the BBA from Figure 3.6a discounted by factor $\alpha_u = 0.3$. It should be observed that discounting simply transfers the belief from propositions $MERGE$ and $DONTMERGE$ to $DOUBT$ ($MERGE \cup DONTMERGE$).

Clearly, the difficulty then lies in the correct choice of the thresholds T_u^l, T_u^c, T_u^r and discounting factor α_u for each source of information. All the above

parameters are estimated using the training collection as discussed later in section 3.7.4.

Combining Information Sources

Although the general form of belief structures (BBA) is rather intuitive the result of combining such BBA for two or more information sources using the Dempster's combination rule from equations 3.11 and 3.12 may not be immediately apparent. Therefore, let us consider an example where two belief structures m_u and m_v are combined using the above rule. According to equation 3.12 the combined belief structure $m^{\oplus}$ is computed as follows:

$$\begin{aligned} m^{\oplus}(MERGE) = \frac{1}{1-K}\big(&m_u(MERGE)\cdot m_v(MERGE) + \\ &+ m_u(MERGE)\cdot m_v(DOUBT) + \\ &+ m_u(DOUBT)\cdot m_v(MERGE)\big) \end{aligned} \tag{3.18}$$

$$\begin{aligned} m^{\oplus}(DONTMERGE) = \frac{1}{1-K}\big(&m_u(DONTMERGE)\cdot m_v(DONTMERGE) + \\ &+ m_u(DONTMERGE)\cdot m_v(DOUBT) + \\ &+ m_u(DOUBT)\cdot m_v(DONTMERGE)\big) \end{aligned} \tag{3.19}$$

$$m^{\oplus}(DOUBT) = \frac{1}{1-K}\big(m_u(DOUBT)\cdot m_v(DOUBT)\big) \tag{3.20}$$

$$\begin{aligned} K = &m_u(MERGE)\cdot m_v(DONTMERGE) + \\ &+ m_u(DONTMERGE)\cdot m_v(MERGE) \end{aligned} \tag{3.21}$$

Merging Cost

In the proposed approach the new cost of merging of two neighbouring regions r_i and r_j based on evidence from one or a combination of sources is defined using an empirically derived formula:

$$C_{Total}(i,j) = m^{\oplus}_{(i,j)}(DONTMERGE) - m^{\oplus}_{(i,j)}(MERGE) \tag{3.22}$$

Table 3.1: Combination of two BBA using the Dempster's combination rule. Each table entry indicates hypothesis supported by a certain conjunction of belief masses from the individual BBA.

m_v \ m_u	MERGE	DOUBT	DONTMERGE
MERGE	MERGE	MERGE	$\emptyset$ (K)
DOUBT	MERGE	DOUBT	DONTMERGE
DONTMERGE	$\emptyset$ (K)	DONTMERGE	DONTMERGE

where $m^{\oplus}_{(i,j)}(DONTMERGE)$ and $m^{\oplus}_{(i,j)}(MERGE)$ are combined beliefs that measurements obtained for the pair r_i and r_j from all integrated sources of evidence support the hypothesis $MERGE$ and $DONTMERGE$ respectively.

The greater (lower) the belief that measurements obtained for the link between r_i and r_j from all combined sources of evidence support the proposition $MERGE$ the lower (greater) the cost of merging. Analogically, the greater (lower) the belief that the measurements belong to class $DONTMERGE$ the greater (lower) the cost of merging. Note that the belief associated with proposition $DOUBT$ ($MERGE \cup DONTMERGE$) does not explicitly contribute to the final cost of merging. The above formula is further clarified in the next section that discusses the design of BBA for each source of information.

3.7.4 Designing Belief Structures for each Source of Information

The design of BBAs for each source of information S_u involves selection of the thresholds T^l_u, T^c_u, T^r_u and the discounting factor α_u all of which are estimated in the proposed approach using the training collection. This section describes the general approach to estimation of the above parameters and discuses in detail the BBA constructed for colour, adjacency and changes of shape complexity.

General Approach to Estimation of T^l, T^c, and T^r

For a source of evidence providing measurement C_u of a certain property of links between region pairs the thresholds T^l_u, T^c_u and T^r_u are found by computing some statistics of C_u from examples of links which "should" or "should not" be merged. The training collection, containing examples of links with an associated set of measurements (i.e. colour homogeneity and geometric properties) and classified as either "link to be merged" or "link which should not be merged" is automatically generated using the training collection of images with ground-truth. It should be stressed that such a collection of links cannot possibly be created directly by a manual annotation. Firstly, such an annotation of all possible mergings between image regions would be prohibitively time consuming, and secondly the exemplar links should be typical of the ones

occurring during the RSST merging.

Instead, the training collection of labelled links is created in the following procedure. Each image from the training collection is segmented using the original RSST algorithm implemented as in [140] except that only links corresponding to merges completely "compatible" with the ground-truth mask are allowed to compete for merging and consequently merged. In other words, each image is segmented by the original RSST algorithm constrained to produce segmentation identical to the ground-truth segmentation mask. When the number of regions becomes lower than the number specified as the beginning of the second segmentation stage (which in all experiments is set to 100 as it has been explained in section 3.3) all relevant properties (i.e. colour homogeneities, geometric properties, and category: either "should be merged" or "should not be merged") of all newly created links are computed and stored in the training collection.

Once the training collection of links is generated the parameters T_u^l, T_u^c and T_u^r can be estimated. First, it should be recalled from Figure 3.6a that for a perfectly reliable source, the parameters T_u^l and T_u^r define the limits of the range of values of measure C_u for which the masses of belief associated with propositions $MERGE$ and $DONTMERGE$ are lower than one, i.e. the mass of belief associated with $DOUBT$ is greater than zero. In other words, for values of C_u lower (greater) than T_u^l (T_u^r) the mass of belief associated with the proposition $MERGE$ ($DONTMERGE$) is equal one. Thus, the parameter T_u^l (T_u^r) can be estimated by finding the minimum (maximum) value of C_u in the training population of links which should (should not) be merged. In fact to avoid bias caused by a single link (e.g. generated from an "accidental" manual segmentation) the above measures are found individually for each image in the training collection and the final values of parameters T_u^l and T_u^r are computed as their average values. Formally, assuming training collection $\mathcal{T}$ containing N images, values of parameters $\hat{T}_u^{l,(I)}$ and $\hat{T}_u^{r,(I)}$ estimated based on image $I \in \mathcal{T}$ are computed as:

$$\hat{T}_u^{l,(I)} = \min_{\forall (i,j) \in \mathcal{B}^{(I)}} C_u(i,j) \tag{3.23}$$

and

$$\hat{T}_u^{r,(I)} = \max_{\forall (i,j) \in \mathcal{G}^{(I)}} C_u(i,j) \tag{3.24}$$

where (i,j) denotes link between a pair of neighbouring regions r_i and r_j. $\mathcal{G}^{(I)}$ and $\mathcal{B}^{(I)}$ denote sets of such links generated from I which "should" and "should not" be merged respectively. Finally $\hat{T}_u^l$ and $\hat{T}_u^r$ are computed as average values of $\hat{T}_u^{l,(I)}$ and $\hat{T}_u^{r,(I)}$ respectively over the entire $\mathcal{T}$:

$$\hat{T}_u^l = \frac{1}{N} \sum_{\forall I \in \mathcal{T}} \hat{T}_u^{l,(I)} \tag{3.25}$$

and

$$\hat{T}_u^r = \frac{1}{N} \sum_{\forall I \in \mathcal{T}} \hat{T}_u^{r,(I)} \tag{3.26}$$

Now let us recall that the parameter T_u^c defines the value of measure C_u for which the mass of belief associated with proposition $DOUBT$ ($MERGE \cup DONTMERGE$) reaches maximum. In this work the value of $\hat{T}_u^c$ is computed simply as mid point of $\hat{T}_u^l$ and $\hat{T}_u^r$.

General Approach to Estimation of Discounting Factor α

The parameter α_u reflecting the reliability of the source S_u is found differently than the previously discussed parameters. First of all, for simplicity the discounting factor corresponding to colour homogeneity α_1 is always set to one since intuitively the evidence provided by colour homogeneity is the most reliable of all sources utilized in this work. For the remaining sources discounting factors are found by combining them with the colour homogeneity criterion using the proposed framework and searching for discounting factor/factors minimizing the average spatial segmentation error on the training collection.

It should be stressed that although the above parameter estimation procedure is intuitive and performs well for all evaluated sources it is also rather crude and an alternative solution should be possible. For example, all parameters could be optimized simultaneously in a procedure similar to the one employed currently only for the estimation of discounting factors. However, even in such a case, the procedure described in this section could be used to provide the initial estimates of each parameter.

BBA for Colour Homogeneity

Two colour homogeneity criteria which performed particulary well in the experiments presented in section 3.5 are further tested with the new integration framework. They are the modified colour homogeneity criterion C_{avg} based on the region's mean colour (formula 3.2) and the colour homogeneity criterion C_{ext} based on the extended colour representation (formula 3.2).

Before discussing the mapping of the abovementioned measures into belief masses it should be first noted that they are not invariant to image size. Therefore for the purpose of integration in the proposed framework both measures are normalized by the size of the entire image A_I: $C'_{avg}(i,j) = \frac{C_{avg}(i,j)}{A_I}$ and $C'_{avg}(i,j) = \frac{C_{ext}(i,j)}{A_I}$. Such normalization does not affect the merging order.

Let us also recall from Figure 3.6 that in the proposed form of the belief structures values smaller than T^l are mapped to identical masses of belief assigned to proposition $MERGE$. To ensure differentiation between small differences in

Table 3.2: Parameters T^l, T^c, T^r, and α for colour homogeneity measures BBA.

colour homogeneity	T^l	T^c	T^r	α
C_{avg}, ($q_{avg} = 6.0$)	0.0	2.1×10^8	4.2×10^8	1.0
C_{ext}, ($q_{ext} = 4.0$)	0.0	9.4×10^{-5}	18.8×10^{-5}	1.0

Table 3.3: Estimated values of discounting factor α_{adj} for use in combination with other features.

	$[C_{avg}]$	$[C_{avg}, C_{cpx}]$	$[C_{ext}]$	$[C_{ext}, C_{cpx}]$
α_{adj}	0.4	0.4	1.0	0.6

colour homogeneity, parameter T^l is set to zero for all colour homogeneity criteria. Interestingly the automatically estimated values of T^l based on the procedure described in the previous section are typically very small and also lead to correct results.

The set of parameters completely defining BBA for both colour homogeneity measures found using the entire available image collection are shown in Table 3.2.

Interestingly, it can be shown that utilizing only the colour homogeneity criterion in the above framework (without additional sources of evidence) results in a merging order identical to the one which would be produced by directly using the colour homogeneity criterion. This can be seen by looking at equation 3.22 defining the new merging cost and the general form of BBA from Figure 3.6. If only the colour information is used the combined belief structure $m^{\oplus}$ reduces simply to just the colour's BBA and the value of $C_{Total}(i,j)$ increases linearly and monotonically for increasing values of colour homogeneity criteria ($C_{avg}(i,j)$ or $C_{ext}(i,j)$) from range $[0, T^r]$. As a consequence, formula 3.22 results in the merging order identical to the one which would be produced by utilizing the colour homogeneity criterion directly. If a single additional source of evidence is used the evidence from this source contributes to the combined belief structure $m^{\oplus}$ depending on the mass of belief associated with $DOUBT$ based on the colour homogeneity criterion. In other words, the additional evidence contributes information mainly when the evidence provided by the colour homogeneity criterion is indecisive, i.e. for values of colour homogeneity criteria close to T^c. Therefore, the proposed merging cost integrating multiple sources of evidence can be seen as generalization/extension of the merging costs based solely on colour.

BBA for Adjacency Measure

Values of parameters T^l_{adj}, T^c_{adj}, and T^r_{adj} found using the entire image collection are 0.0, 0.5, and 1.0 respectively. These values appear sensible since if we recall the definition of C_{adj} from section 3.6.1 values of C_{adj} close to zero indicate quasi-

Table 3.4: Estimated values of discounting factor α_{cplx} for use in combination with other features.

	$[C_{avg}]$	$[C_{avg}, C_{adj}]$	$[C_{ext}]$	$[C_{ext}, C_{adj}]$
α_{cpx}	0.9	1.0	1.0	0.5

inclusion (one region is almost completely surrounded by another) whereas values close to one indicate "weak" adjacency between regions. Estimated values of discounting factor α_{adj} for use in combination with both colour homogeneity measures and measure of change in shape complexity are shown in Table 3.3.

BBA for Measure of Changes in Global Shape Complexity

The measure of average change of global shape complexity C_{cpx} (defined by equation 3.7) requires a more flexible form of BBA. This is due to the fact that reliability of evidence provided by the measure C_{cpx} depends on the length of the contour of the hypothetical region r_{ij} (created by merging two neighbouring regions r_i and r_j) common with the image border. In other words, whenever the hypothetical region r_{ij} shares a significant part of its contour with the image border the measure C_{cpx} does not provide any useful evidence. This fact can be taken into account by additional transfer of the belief from propositions $MERGE$ and $DONTMERGE$ to the proposition $DOUBT$ depending on the adjacency of r_{ij} with the border of the entire image. Such transfer can be performed by applying an additional discounting operation according to the formulas 3.14 and 3.15 for each pair of regions r_i and r_j. In this work the additional discounting factor is defined as:

$$\alpha'_{cpx}(i,j) = 1.0 - \frac{l_{Iij}}{l_{ij}} \tag{3.27}$$

where l_{ij} is the perimeter length of r_{ij} and l_{Iij} denotes the length of the common boundary between r_{ij} and the entire image. Note that whenever $l_{Iij} = l_{ij}$ the total mass of belief is assigned to proposition $DOUBT$. In other words, the above model is equivalent of using a joint BBA for measures C_{cpx} and α'_{cpx}. This example is a good illustration of the ignorance management capabilities of BeT.

All necessary parameters of the BBA are estimated based only on links with $\alpha'_{cpx}(i,j) > 0.5$ to avoid the influence of meaningless examples[8]. Values of parameters T^l_{cpx}, T^c_{cpx}, and T^r_{cpx} found using the entire image collection are 0.7, 1.1, and 1.4 respectively. These values appear sensible since if we recall the definition of C_{cpx} (equation 3.7) values of C_{cpx} below 1.0 indicate that the complexity of the joint region is lower than the average complexity of the two

[8]Examples where the hypothetical region r_{ij} shares a significant part of its contour with the image border.

Table 3.5: Results of cross-validation for different combinations of features.

Integrated Features	Measures	Avg. Spatial Segm. Error		
		Collection A	Collection B	Avg.
original colour homogeneity criterion [140] (colour modelled by mean values of colour components;)	C_{orig} (eq.3.1)	0.95	1.05	1.00
modified colour homogeneity (colour modelled by mean values of colour components; modification of the balance between factors depending on colour difference and region's size;)	C_{avg} (eq.3.2)	0.60	0.68	0.64
modified colour homogeneity and adjacency measure	C_{avg} (eq.3.2) C_{adj} (eq.3.6)	0.53	0.65	0.59
modified colour homogeneity and measure of changes of global shape complexity	C_{avg} (eq.3.2) C_{cpx} (eq.3.7)	0.49	0.58	0.54
modified colour homogeneity, adjacency measure, and measure of changes of global shape complexity (complex dependency between adjacency and changes of global shape complexity)	C_{avg} (eq.3.2) C_{adj} (eq.3.6) C_{cpx} (eq.3.7)	0.48	0.64	0.56
colour homogeneity based on the extended colour representation (the extended colour representation and quadratic distance)	C_{ext} (eq.3.5)	0.59	0.67	0.63
colour homogeneity based on the extended colour representation and adjacency measure	C_{ext} (eq.3.5) C_{adj} (eq.3.6)	0.54	0.57	0.56
colour homogeneity based on the extended colour representation and measure of changes of global shape complexity	C_{ext} (eq.3.5) C_{cpx} (eq.3.7)	0.50	0.56	0.53
colour homogeneity based on the extended colour representation, adjacency measure, and measure of changes of global shape complexity (complex dependency between adjacency and changes of global shape complexity)	C_{ext} (eq.3.5) C_{adj} (eq.3.6) C_{cpx} (eq.3.7)	0.48	0.54	0.51

original regions whereas values above 1.0 indicate increase in shape complexity if the regions are merged. Estimated values of discounting factor α_{cpx} for use in combination with both colour homogeneity measures and adjacency measure are shown in Table 3.4.

3.7.5 Evaluation of the New Framework for Combining Multiple Features

This section presents results of cross-validation obtained by different combinations of features (sources of information). It should be recalled from section 3.4 that only the merging order is evaluated here by using the number of regions available from the ground-truth as the "optimal stopping criterion". Fully automatic stopping criteria are discussed and evaluated in the next section.

Table 3.5 shows performances obtained by different combinations of features during the cross-validation. The first column contains a brief characterization of each feature combination. The second column recalls the formulas used to

compute each homogeneity measure. The performances in terms of average spatial segmentation error obtained during the cross-validation for subsets A and B are shown in the third and fourth columns respectively. Column 5 contains the average error from the two trials. The results for the original RSST algorithm are shown in the first row for reference. All the remaining rows contain results obtained by utilization of a single or multiple features using the proposed integration framework.

All the extensions discussed in previous sections led to some improvement of the segmentation quality. The improvements obtained during the cross-validation are consistent with the improvements promised during the training indicating good generalization. The best result was obtained for the approach integrating colour homogeneity based on the extended colour representation together with the adjacency measure and the measure of shape complexity change which almost halved the average spatial segmentation error compared to the original RSST approach. Also, manual inspection of the results confirms that the proposed extensions significantly improve correspondence of segmentation to semantic notions compared to the original RSST approach.

Although the integration of both geometric properties with the colour homogeneity based on the extended colour representation further reduces the spatial accuracy errors the improvement is not as significant as the one obtained by the relatively simple modifications of the colour homogeneity criteria discussed in section 3.5. Integration of the syntactic features reduced the spatial segmentation error by only around 15% comparing to the best colour homogeneity criteria used alone (C_{ext}). Clearly, this may be attributed to the fact that the results obtained by the proposed modifications of the colour homogeneity criteria on the tested collection are close to saturation and subsequent improvements are much harder to obtain. Manual inspection of the results indicate that the improvement obtained by integration of syntactic features is generally more pronounced in cases of difficult lighting conditions, e.g. shadows across a human face, textured areas and areas containing many small regions with distinctive colour included inside larger objects. Such images constitute only part of the collection used for evaluation. The images from the Corel collection, which constitute most of the collection used here, are probably one of the most challenging image sets publicly available to evaluate the performance of the syntactic features since many of them can be successfully segmented using solely colour homogeneity criteria. Therefore an important property of the proposed approach is that the introduction of syntactic features improves the results wherever possible without degrading the performance in cases where the colour homogeneity alone is sufficient. Nevertheless, the above experiments confirm that syntactic features, although useful, can provide only auxiliary evidence to colour information.

Also, integration of both geometric properties, adjacency and shape complexity,

resulted in a smaller improvement than could be expected from improvements caused by each of them individually. In fact, the performance obtained by integrating only one of the two syntactic features is comparable to the one obtained by integration of both of them. This result is confirmed by manual inspection. This can be explained by the complex relation (dependencies) between these features already discussed in section 3.6. For example, it is possible that utilization of one feature may affect the geometric properties of regions created at each iteration in a way which affects the role of the other feature. Moreover, the formulas 3.11 and 3.12 used for combination of sources are derived based on the assumption that the sources are independent. Combination of correlated sources is currently under investigation by several researchers [157] and potential developments in this area could clearly benefit the proposed approach.

Representative examples of segmented images produced by the approach based solely on the extended colour representation and its various combinations integrating the syntactic features are shown in Figure 3.7. As previously, the first top five examples show cases where the syntactic features result in lower spatial segmentation error than the approach based solely on the extended colour representation whereas the bottom five examples show cases where the syntactic features increase the error.

In fact, close manual inspection revealed that the measure of changes of global shape complexity alone often produce more meaningful results than both syntactic features together except in cases where the initial partition contained erroneous regions with complex shape causing further erroneous merges attempting to reduce the shape complexity or in cases where some objects were over-segmented due to the fact that the regions constituting them had already circular or elliptical shape (e.g. a butterfly with relatively small but circularly shaped spots with distinctive colour).

Moreover, close manual inspection of the results revealed that, although segmentations obtained using the extended colour representation are often close to the one obtained by using the modified criterion based on regions' average colour, in some cases utilization of the extended representation leads to significantly better partitions.

It should also be noted that the colour homogeneity criterion based on a region's average colour (C_{avg}) does not integrate well with the adjacency measure, i.e. this results in very small improvement. In fact, integration of C_{avg} with both syntactic features resulted in worse performance that when integrated solely with the measure of change in global shape complexity (C_{cpx}). This can be explained by the fact that the utilization of the adjacency measure causes many small regions with irregular shape or semi-included but with relatively different colour to be merged at early iterations. Subsequently, the average colour may

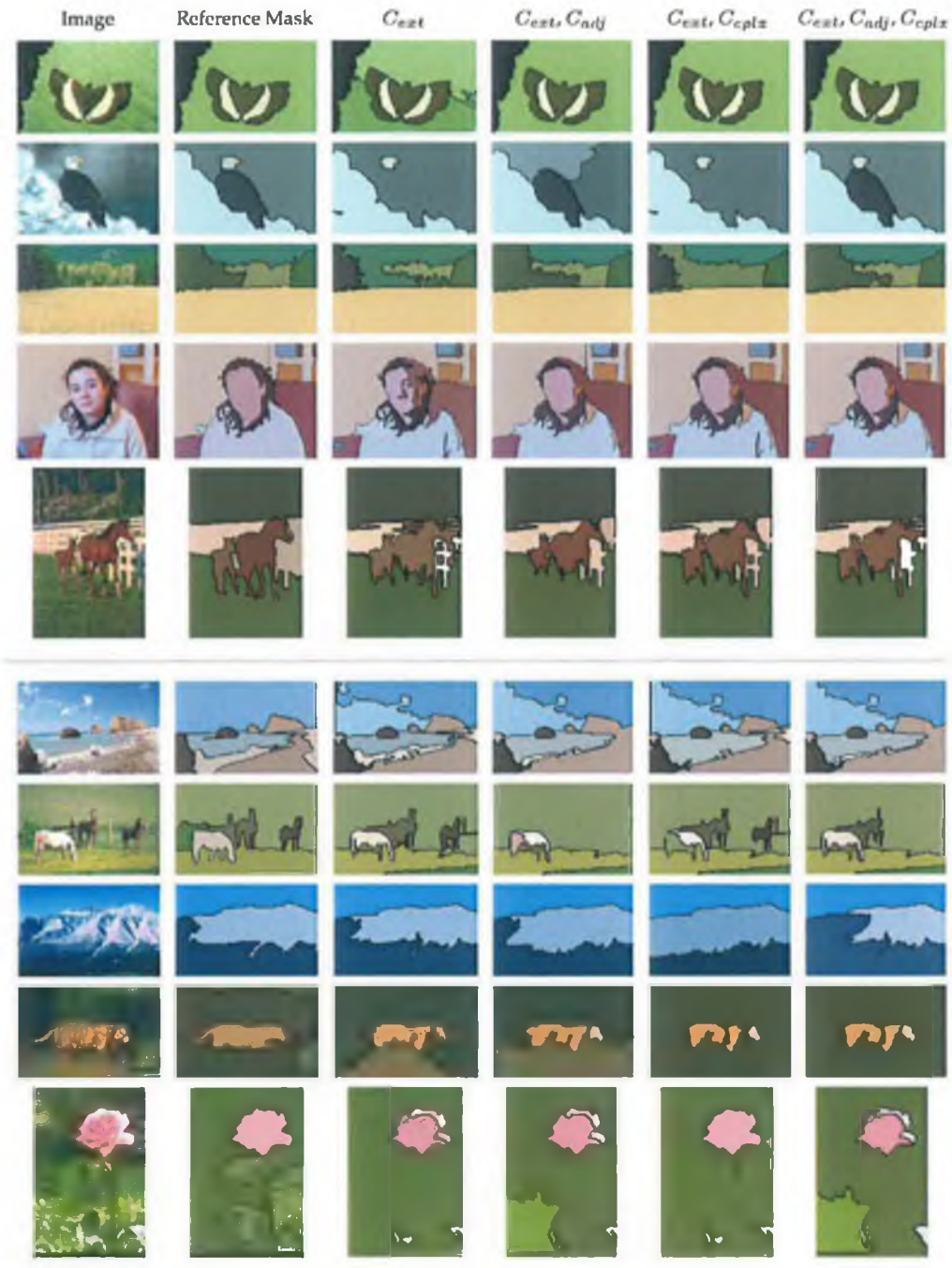


Figure 3.7: Segmentations obtained by merging criteria integrating colour homogeneity based on the extended colour representation with different combinations of syntactic features (the required number of regions is known).

Figure 3.8: Segmentations obtained by merging costs integrating the measure of adjacency and changes in global shape complexity with two different colour homogeneity criteria: the modified colour homogeneity based on regions' average colour (C_{avg}) and the colour homogeneity based on the extended colour representation (C_{ext}). The required number of regions is known.

not be sufficient to represent the colour of such regions resulting in some erroneous merges in consecutive iterations. The above example illustrates the most disappointing limitation of the RSST algorithm which is that it does not explore well the solution space. In other words, in extreme cases more plausible merges performed at early iterations may lead to erroneous merges at the later stages.

Representative examples of segmented images produced by merging costs integrating geometric measures with the two colour homogeneity criteria: C_{avg} and C_{ext} are shown in Figure 3.8. As previously, the first top five examples show cases where the extended colour representation results in lower spatial segmentation error than the approach based on the modified criterion based on regions' average colour whereas the bottom five examples show cases where the extended colour representation increases the error.

Manual inspection of the results revealed that evaluation of merging criteria using partitions with the number of regions equal to the number of regions from the ground-truth may not fully illustrate the effects of the proposed extensions. For example, a modification resulting in a more meaningful partition but producing a significant number of small regions with distinctive colour has a serious disadvantage when evaluated using the proposed methodology. In other words, the high number of small regions may lead to significant under-segmentation of the remaining parts of the image in order to reduce the overall number of regions to the number of regions from the ground-truth. More meaningful comparison of merging criteria could be achieved by analyzing all partitions produced during the merging process and taking into account only the partition with minimum spatial segmentation error.

3.8 Stopping Criterion

Segmentation of scenes into semantic entities based solely on low level features is an ill-posed problem and building systems relying on such single partitions should be avoided whenever possible. Instead the segmentation should produce a hierarchical representation of the image, e.g. BPT. However, currently many applications, including CBIR systems discussed in section 2.2, can utilize only a single partition of the scene. Moreover, even in scenarios where the segmentation is used to produce a hierarchical representation of the image it is often necessary, e.g. due to reasons of efficiency, to identify a single partition within such a representation most likely to contain a meaningful segmentation – see for example [41]. In particular, a stopping criterion could be used to select a range of the most likely meaningful partitions or at least to identify the top nodes in such hierarchical representations which are often meaningless, i.e. correspond to merging different semantic entities.

This section discusses the feasibility of producing such single partitions aligned with the semantically meaningful entities present in the scene by stopping the region merging process. It is assumed that potential applications can benefit from single partitions despite their limited accuracy. First, a very simple but common stopping criterion based on value of *Peak Signal to Noise Ratio* (PSRN) between the original and segmented image (reconstructed using mean region colour) [41] is discussed. This is followed by a proposal of a new approach to stopping the merging process.

3.8.1 Stopping Criterion Based on PSNR

In this case a single partitioning of the image is achieved based on value of PSNR between the original and segmented image (reconstructed using mean region colour) – see for example [41]. First, let us assume that all merges performed during the region merging process are stored in a *Binary Partition Tree* (BPT) together with values of PSNR at each merging iteration. A single partition is then obtained from the BPT by deactivating nodes following the merging sequence until the value of PSNR falls below a pre-defined threshold.

The threshold is typically chosen in an ad-hoc manner to suit a particular application. In this work, only the luminance information is used in order to limit the computational complexity.

Formally PSNR for an $M \times N$ image is computed as:

$$PSNR = 20 \log_{10} \frac{MAX_L}{\sqrt{\frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \Big(L(i,j) - K(i,j)\Big)^2}} \tag{3.28}$$

where L is the original image (luminance information only) and K is the segmented image reconstructed using mean region intensities. MAX_L is the maximum luminance value.

The main drawback of such an approach is that PSNR values do not reflect correctly the similarity between the original image and the segmented one. This is particulary apparent for textured regions or regions with gradual changes of colour. For example it is impossible to chose a single threshold which can result in a meaningful segmentation of both: images containing objects consisting of heavily textured regions and images containing semantically different objects but which have similar colour, e.g. compare the image from Figure 3.10a with the image from Figure 3.10b).

Here we take advantage of the evaluation methodology used in previous sections to demonstrate the influence of different values of the threshold (T_{PSNR}) on the spatial accuracy of the segmentation and compare the approach with the case when the optimal number of regions is used. The stopping method is

evaluated in combination with several selected region merging criteria described in the previous sections.

Figure 3.9 shows the segmentation quality obtained for the entire collection by six homogeneity criteria for different values of the threshold T_{psnr} used to stop the merging process. The results of cross-validation are shown in Table 3.6.

Not surprisingly, the stopping criterion based on PSNR obtained significantly higher spatial segmentation error comparing to the "manual" stopping criterion for all merging criteria except the original RSST algorithm. This is confirmed by visual inspection of the segmentation masks – see the 3rd column from Figure 3.12. Interestingly, Figure 3.9 suggests that different merging criteria may be suitable for different stopping conditions, e.g. over- or under-segmentation. For example, although the original homogeneity criterion (C_{orig}) performs relatively well in cases of under-segmentation (T_{PSNR} below 44dB forces the creation of large regions) it is significantly outperformed by the new merging criteria when over-segmentation is required.

Although for the majority of the evaluated merging criteria the lowest spatial segmentation error was obtained for T_{PSNR} close to $44dB$ manual inspection of the results revealed that for many images this value results in strong under-segmentation. This can be explained by the fact that many of the images in the collection contain a small number of objects, often textured, for which the above value of T_{PSNR} produces results close to optimal.

Interestingly, the lowest spatial segmentation error was obtained by the merging criterion integrating the colour homogeneity based on the extended colour representation and the measure of changes of shape complexity alone. The above merging criterion not only obtained the best performance for the optimal value of $T_{PSNR} = 44dB$, but also outperformed other configurations for higher values of the threshold, possibly being more general for other collections.

The above results suggest that further improvement could be achieved if the merging criteria are optimized simultaneously with the stopping criterion. Also, to fully utilize the improvements proposed in previous sections to merging criteria a new stopping criterion is required.

3.8.2 A New Stopping Criterion

Alternatively to the PSNR-based solution, the merging process could be stopped by comparing the cost of merging at each iteration with a predefined threshold. However the author's experiments (results omitted for brevity) indicate that choosing a single threshold in such cases results either in over-segmentation of some scenes, e.g. with difficult lighting conditions, or under-segmentation, e.g. containing objects with similar colour.

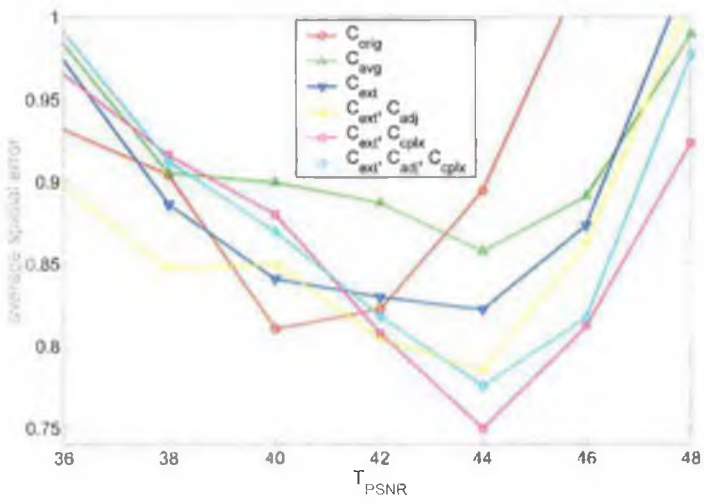


Figure 3.9: Influence of T_{psnr} on the average spatial segmentation error.

This section discusses a new approach to obtaining a single partition of images in which instead of evaluating the goodness of segmentation at a single iteration only, the decision is made based on the evolution of the merging cost accumulated during the complete merging process, i.e. starting from the level of pixels and finishing with the single region corresponding to the entire image. As in the previous approach, the selected partition is obtained from the BPT built during the merging process by deactivating nodes following the merging sequence.

First, let us define an accumulated merging cost measure C_{cum} which measures the total cost of all mergings performed to produce t regions as:

$$C_{cum}(t) = \begin{cases} \sum_{n=t}^{N_I-1} C_{mrg}(n) & \text{if} \quad 1 \leq t < N_I \\ 0.0 & \text{otherwise} \end{cases} \tag{3.29}$$

where $C_{mrg}(n)$ denotes cost of merging of a single pair of regions reducing the number of regions from $n+1$ to n computed using a given merging criterion. N_I denotes the number of regions in the initial partition, i.e. produced in the first merging stage. Only costs of merging performed during the second stage contribute to the value of C_{cum}.

Figure 3.10 shows measure C_{cum} plotted for each iteration t of the merging process for three images, each presenting a different segmentation challenge. The basic idea behind the stopping criterion is to find the number of regions t_s which partitions the curve $C_{cum}(t)$ into two segments in such a way so that with decreasing values of t, values of $C_{cum}(t)$ within segment $[1, t_s]$ increase significantly faster than for the segment $[t_s, N_I]$. Also it should be possible to bias the result towards under- or over-segmentation.

In the proposed approach t_s is found using the method proposed in [158] for bi-level thresholding designed to cope with uni-modal distributions (histogram

Table 3.6: Results of cross-validation of selected merging criteria used with the PSNR-based stopping criterion.

Integrated Features	Measures	Avg. Spatial Segm. Error		
		Collection A	Collection B	Avg.
original colour homogeneity criterion [140] (colour modelled by mean values of colour components;)	C_{orig} (eq.3.1)	0.84	0.84	0.84
modified colour homogeneity (colour modelled by mean values of colour components; modification of the balance between factors depending on colour difference and region's size;)	C_{avg} (eq.3.2)	0.93	0.88	0.91
colour homogeneity based on the extended colour representation (based on the extended colour representation and quadratic distance)	C_{ext} (eq.3.5)	0.90	0.76	0.83
colour homogeneity based on the extended colour representation and the adjacency measure	C_{ext} (eq.3.5) C_{adj} (eq.3.6)	0.82	0.77	0.80
colour homogeneity based on the extended colour representation and measure of changes of global shape complexity	C_{ext} (eq.3.5) C_{cpx} (eq.3.7)	0.74	0.78	0.76
colour homogeneity based on the extended colour representation, adjacency measure, and measure of changes of global shape complexity (complex dependency between adjacency and changes of global shape complexity)	C_{ext} (eq.3.5) C_{adj} (eq.3.6) C_{cpx} (eq.3.7)	0.79	0.78	0.79

based thresholding). The algorithm is based on the assumption that the main peak of the uni-modal distribution has a detectable corner at its base which corresponds to a suitable threshold point. The approach has been found suitable for various problems requiring thresholding such as edge detection, image difference, optic flow, texture difference images, polygonal approximations of curves and even parameter selection for the split and merge image segmentation algorithm [40].

Here, the approach is used to determine the stopping criterion t_s based on the accumulated merging cost measure C_{cum}. Let us assume a hypothetical reference accumulated merging cost measure C_{ref} having a form of a line passing through points $\Big(1, C_{cum}(1)\Big)$ and $\Big(T_{cum}, C_{cum}(T_{cum})\Big)$, where T_{cum} is a parameter whose role is explained later. The threshold point t_s is selected to maximize the perpendicular distance between the reference line and the point $\Big(t_s, C_{cum}(t_s)\Big)$ – see examples in Figure 3.10. It can be shown that this step is equivalent to an application of a single step of the standard recursive subdivision method for determining the polygonal approximation of a curve [159].

Parameter T_{cum} can be used to bias the stopping criterion towards under- or over-segmentation, i.e. smaller values of T_{cum} bias the stopping criterion towards under-segmentation while larger values tend to lead to over-segmentation. However, it should be stressed that it is the shape of $C_{cum}(t)$ which plays the dominant role in selecting t_s.

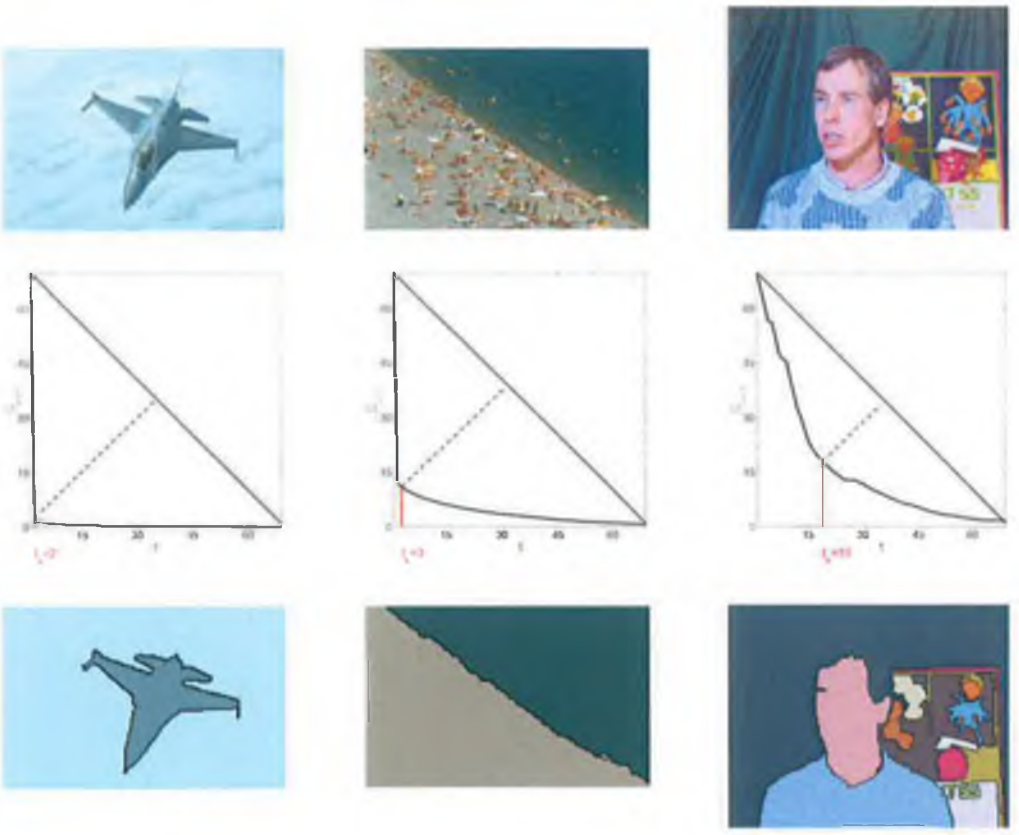


Figure 3.10: Stopping criterion based on accumulated merging cost (C_{cum}) for $T_{cum} = 70$. The merging order is computed using the extended colour representation together with syntactic features. Values of C_{cum} are re-scaled to the range $[0, T_{cum}]$ for visualization purposes.

Figure 3.11 shows the average spatial segmentation error obtained for the entire image collection by six merging criteria for different values of threshold T_{cum}. Table 3.7 shows results of cross-validation.

Clearly, the proposed stopping criterion significantly outperforms the PSNR-based stopping criterion when used with any of the new merging criteria. It should be noted that the best combination of merging criterion (the extended colour representation combined with syntactic features) together with the new stopping criterion leads to significantly lower average spatial segmentation error than the original RSST approach even with the "manual" stopping criterion based on the number of regions in the ground-truth mask.

However, it can be seen from Figure 3.11 that the new stopping criterion is not suitable for use with the merging criterion from the original RSST approach. This suggests dependency of the new stopping criterion on the balance between the regions' colour differences and the size of the regions, i.e. values of q_{avg} and q_{ext} discussed in section 3.5.

Although the merging criterion integrating the syntactic features obtained the best performance its quantitative results are comparable with the merging cri-

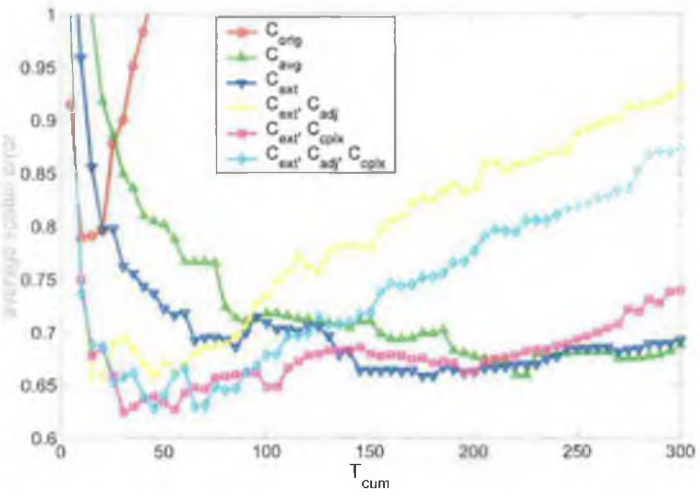


Figure 3.11: Influence of T_{cum} on the average spatial segmentation error.

teria based solely on colour proposed in section 3.5. This could suggest that a stopping criterion which can fully utilize the benefits of syntactic features yet remains to be found. However, further manual inspection of the results for values of T_{cum} different than the ones producing the minimum spatial segmentation error reveals that in fact the syntactic features have strong positive impact on the results, especially for values of $T_{cum} \sim 400$. This is especially true for the merging criterion integrating measure C_{cpx} which greatly limits the number of regions spanning more than one object. The reason this configuration obtained slightly worse results in the qualitative evaluation than the configuration based solely on colour homogeneity criterion for $T_{cum} \sim 400$ is due to the fact that in some cases it also over-segments large regions. However, practically all cases of such over-segmentation are meaningful, i.e. there is a visual difference between the created regions.

Figure 3.12 shows selected segmentation results obtained by the original RSST algorithm with the PSNR based stopping criterion (3rd column) and segmentations produced by the best of the proposed approaches: merging criterion integrating the extended colour representation combined with syntactic features together with the new stopping criterion. As in the previous cases, the first five rows show examples where the new approach improves the results compared to the original RSST and the last five rows show examples where the proposed approach obtained higher spatial segmentation error than the original RSST approach. However, it should be stressed that the latter cases are very rare and in fact even in such cases the partitions appear intuitive. Moreover, it should be stressed that all segmentations for the proposed methods were produced for values of parameters found in the training process which due to the evaluation criterion chosen and the collection used tends to produce under-segmentation. As already explained, for the merging criteria integrating the syntactic features subjectively better results can be obtained by higher values of T_{cum}.

Table 3.7: Results of cross-validation of selected merging criteria with the proposed stopping criterion.

Integrated Features	Measures	Avg. Spatial Segm. Error		
		Collection A	Collection B	Avg.
original colour homogeneity criterion [140] (colour modelled by mean values of colour components;)	C_{orig} (eq.3.1)	0.71	0.91	0.81
modified colour homogeneity (colour modelled by mean values of colour components; modification of the balance between factors depending on colour difference and region's size;)	C_{avg} (eq.3.2)	0.60	0.76	0.68
colour homogeneity based on the extended colour representation (based on the extended colour representation and quadratic distance)	C_{ext} (eq.3.5)	0.60	0.74	0.67
colour homogeneity based on the extended colour representation and adjacency measure	C_{ext} (eq.3.5) C_{adj} (eq.3.6)	0.67	0.66	0.67
colour homogeneity based on the extended colour representation and measure of changes of global shape complexity	C_{ext} (eq.3.5) C_{cpx} (eq.3.7)	0.61	0.67	0.64
colour homogeneity based on the extended colour representation, the adjacency measure, and the measure of changes of global shape complexity (complex dependency between adjacency and changes of global shape complexity)	C_{ext} (eq.3.5) C_{adj} (eq.3.6) C_{cpx} (eq.3.7)	0.61	0.64	0.63

For comparison, the last column from Figure 3.12 shows the results of the Blobworld segmentation algorithm[9]. The Blobworld algorithm has been extensively tested and produced very satisfactory results in the past [29].

Although, in many cases the proposed stopping criterion does not produce perfect segmentation, i.e. ideally aligned with the manual segmentation, it successfully identifies the most visually salient regions in almost all evaluated images. This presents a tremendous opportunity for utilizing such salient regions in CBIR systems. It should be stressed that the method performs well on images presenting very different challenges with fixed value of the parameter T_{cum} and unlike the stopping criterion based on PSNR very erroneous segmentations are extremely rare.

3.9 Future Work

RSST has the advantage of speed and simplicity compared to many other approaches and proved to be convenient for integration of multiple features. However, it has also one inherent limitation which is that it does not exhaustively explore the solution space. Future work could progress in two main directions. One research route could consider solving some of the limitations of the current

[9]Source code obtained from http://elib.cs.berkeley.edu/src/blobworld/

Figure 3.12: Selected segmentation results obtained by the fully automatic versions of the proposed extensions.

approach and the second one could focus at the possibilities of adaptation of the solutions proposed in this work into other segmentation frameworks especially the ones promising a better exploration of the solution space. The new framework for integration of evidence from multiple sources, each with its own accuracy and reliability, into the region merging process opens many possibilities of integration of new features such as texture, additional geometric properties, or in the case of video even motion information.

In particular, possible research routes could include:

- **Utilization of Additional Geometric Features** - The discussed list of geometric properties potentially useful in controlling merging order is by no means complete and could be further extended by for example symmetry analysis, shape homogeneity (e.g. similarity between parts of region's contours), continuity and collinearity [67] (see also section 2.3.3 for other possible generic geometric measures).
- **Utilization of Multi-scale Colour and Texture Features** - The colour homogeneity could be further extended by integration of texture homogeneity, e.g. by adopting the approach from [30], and by incorporation of the so-called *Boundary Melting* approach which favors merging of regions with low magnitude of color gradient along their common boundary [105]. In fact utilization of features computed from pixel's local neighbourhood, e.g. texture or even multi-scale colour information, could provide a remedy to the non-optimal exploration of the solution space in the initial merging stages. In other words, colour or texture information from a pixel's neighbourhood could help avoid some erroneous merges caused by noise and the inability of RSST to find the globally optimal solution. However, utilization of texture or multi-scale colour information could also affect contour localization [31] and increase the computational complexity.
- **Analysis of Feature Dependencies** - The dependencies between geometric features could be further analyzed by using techniques like *Principal Component Analysis* (PCA). Also the most recent findings in the area of generalization of TeB to dependent sources of information, e.g. [157], should be considered for extending the proposed integration framework. Furthermore, optimization of all parameters of the merging criteria simultaneously with the optimization of the stoping criteria could lead to more significant improvement in segmentation quality.
- **Utilization of an Alternative Technique in the Pre-Segmentation Stage** - Using an alternative segmentation method to produce the initial partition could improve its quality and in consequence also the quality of the final segmentation. In fact, the proposed approach, i.e. extensions proposed for the second merging stage, could be utilized as a post-processing step to any segmentation algorithm.

- **Training and Testing of the Approach with Other Collections** - As has been already explained in section 2.6.4 the size of the dataset used for training and testing in this thesis is not optimal. Therefore the findings of this chapter should be confirmed with much larger and more heterogenous image collections. One such collection that has been made recently public is the *Berkeley Segmentation Dataset* [160]. The collection contains 300 images, each segmented by 2-6 annotators. The segmentation masks contain closed contours of each region. Adopting the collection to the evaluation methodology used in this thesis would require conversion of the masks from contour representation into regions. Also it could be extremely interesting to evaluate the approach using the original boundary detection and segmentation benchmark proposed for the Berkeley collection which evaluates the regions boundaries in terms of precision and recall [161].
- **Automatic Assessment of the Segmentation Quality** - Although, as shown in this thesis, automatic image segmentation can be quite successful for many types of images it should be also acknowledged that many images cannot be meaningfully segmented in an automatic fashion. This presents a significant challenge in utilizing such results in CBIR systems. One solution would be to accompany the segmentation results with a confidence measure indicating how likely it is that it contains a meaningful segmentation which could be then taken into account by the CBIR system. Such confidence could be derived based on the measure C_{cum} proposed in this chapter.

3.10 Discussion

Extensive evaluation confirmed that the proposed modifications significantly improve the quality of the segmentation produced by the RSST algorithm. Consistent improvement in spatial accuracy is observed through all classes of tested images confirming the general nature of the proposed solutions. In particular, the most promising of the fully automatic approaches proposed here reduces the average spatial segmentation error by around 20% compared to the original RSST approach with the stopping criterion based on PSNR. The proposed method, in contrast to many existing approaches, including the Blobworld algorithm [29], produces accurate region boundaries. Perhaps the most important property of the proposed algorithm is that it works relatively well on images from different collections, such as standard test images, professional databases, digital photographs and video keyframes from broadcast TV, all *using a fixed set of parameter values*. This, along with the fact that it takes less than 3 seconds on Pentium III 600 MHz to segment an image of CIF size (352x288)[10], makes the proposed algorithm very attractive for object-based applications such as content-

[10]Blobworld ,using mostly Matlab code, requires over 30 minutes for this task.

based image retrieval in large collections or definition of interest regions for content-based coding of still images in the context of the JPEG2000 [27, 30], for example.

Although the introduction of syntactic features reduced the spatial accuracy errors, the improvement is smaller than the improvement obtained by the modification of the colour homogeneity criterion. Therefore, the experiments confirmed that syntactic features, although useful, can provide only evidence auxiliary to colour information. Moreover, the analysis of several syntactic features revealed major challenges in integrating them into a segmentation framework arising from strong overlaps between various geometrical properties.

The experiments also confirmed that automatic bottom-up segmentation of images with general content is a non trivial problem and therefore utilization of a single partitioning should be avoided whenever possible. Instead bottom-up segmentation should be rather used to discover the hierarchical structure of the scene, e.g. for creation of a *Binary Partition Trees* (BPT) which can be used for efficient representation of the structure of the scene [107]. Creation of BPT has been advocated as an important pre-processing step in many applications including region-based compression and region-based feature extraction in the context of MPEG-7 description of content. One such application of BPT is shown in section 3.11 where BPT produced by the automatic approach proposed in the previous sections is used to ensure fast responses of a semi-automatic segmentation tool.

Finally it should be stressed that a significant effort has been devoted in this work to objective evaluation of the proposed modifications. To the best of the author's knowledge this is the most comprehensive evaluation of the bottom-up region merging approach currently available in the literature. However, despite such significant evaluation effort the author acknowledges that some of the results are on the verge of statistical significance, e.g. integration of the syntactic features versus the modified merging criteria. Many approaches to image segmentation proposed in the past have been accompanied by only a handful of examples. Clearly, the research carried out by the author indicates that developing a general solution to image segmentation is an elusive task and generality of any improvement demonstrated only on a handful of examples should be approached with an appropriate reserve.

3.11 Semi-Automatic Segmentation

3.11.1 Motivation

The problem of partitioning an image into a set of semantic entities is a fundamental enabling technology for understanding scene structure and identifying

relevant objects. Unfortunately, as already discussed in this chapter, fully automatic object-based segmentation remains to a large extent an unsolved problem, despite being the subject of study for several decades. For a certain class of multimedia applications it is unavoidable not to include user input in the segmentation process in order to allow semantic meaning to be defined [162]. In the case of automatic techniques the control over desired partitioning is often performed through a trial-and-error-based adjustment of parameters. Clearly, to perform easy segmentation, even by an unskilled user, another form of control which is readily conceptualized is needed. This requirement has lead to development of *supervised* or *semi-automatic* approaches where a user, by his/her interactions, defines what objects are to be segmented within an image [32]. The result of interaction provides clues as to the nature of the user's requirements which are then used by an automatic process to segment the required object. The typical requirements for semi-automatic approaches together with a review of the most characteristic approaches have been already provided in chapter 2 (section 2.5.1)

Many researchers, including the author, believe that in many applications the shape of semantic objects in images can only be accurately recovered by a supervised approach. This makes the problem of semi-automatic segmentation and its viability extremely relevant in the context of this thesis. For example, the shape matching technique presented in chapter 5 assumes that the contours are already extracted. In applications where an automatic partitioning of scenes into semantic entities (*bottom-up* approach) is impossible, including the user in the segmentation process might provide a feasible alternative, especially if the collection is small or in the case of video where the object segmented in one image can be successfully tracked over several frames.

As already discussed in chapter 2, another viable alternative to object segmentation and subsequent recognition is a *top-down* approach where a model of an object is created off-line and subsequently used for recognition in unseen images. Even in this scenario typical shape models require tens of contour examples therefore providing additional motivation for the research on semi-automatic segmentation approach carried out in the remainder of this chapter. Specifically, the tool described in this section has proven to be particularly useful in the context of construction of statistical shape models as described in chapter 7 where it is used for extraction of accurate contour examples from a set of images. The solution presented in the remainder of this chapter takes full advantage of the automatic approach and all proposed advances described in the first part of this chapter.

3.11.2 User Interaction

Various strategies for user interaction for extraction of accurate objects have been already discussed in section 2.5.1. The author believes that one of the most

intuitive schemes for user interactions is the *feature-based* approach where the user is allowed to draw on the scene, typically by placing *markers* in a form of single point *seeds* or by dragging a mouse over the image and creating a set of labelled *scribbles* for objects to be segmented. Two or more different markers (or scribbles) can have the same label, indicating different parts of the same object in the image.

An elegant and efficient solution to image partitioning based on scribbles was described in [38] where the input image is automatically pre-segmented into regions and represented as a *Binary Partition Tree* (BPT) to allow rapid responses to user's actions. The tree structure is used to encode similarities between regions pre-computed during the automatic segmentation. This structure is then used to rapidly propagate the labels from scribbles provided by the user to large areas (regions) of the image. This solution was adapted by the author as it promised high speed efficiency crucial in applications involving user interactions. Also it allowed conveniently taking advantage of the developments, presented in the first part of this chapter of incorporating syntactic features into automatic segmentation.

Since the primary goal of the tool developed by the author is extraction of a single closed contour from a set of images (see chapter 7) the allowed interaction is restricted to only two objects referred in the remainder of this chapter as foreground and background. Both objects can be made up of an arbitrary number of regions, homogenous according to a certain criteria. Only two sets of disconnected scribbles are required: one for foreground and one for background. The author's observations indicate that the above restriction greatly simplifies user interaction and subsequent interpretation of the results. In particular, on a standard PC, foreground and background can be associated with left and right mouse buttons respectively. The scribbles can have a complex shape and cover large areas or can be as small as one pixel.

3.11.3 Implementation Details

The basic idea of the approach can be found originally in [38][11] and some implementation details were also provided in [107]. This section aims at providing a deeper illustration of the central implementation issues and discusses some choices made by the author. In particular, it provides an alternative and simpler solution to bottom-up label propagation, demonstrates some flaws of the strategy suggested in [107] for filling of the entire image space and presents a simple solution to the above problem.

[11] [38] discusses various applications of BPT representation the context of MPEG-4 and MPEG-7 standards.

Creation of BPT

In the author's implementation, the BPT is created using the automatic segmentation approach described in this first part of the chapter. The syntactic approach was preferred over the original RSST algorithm because, as was demonstrated in section 3.7.5, in most cases, it results in a more intuitive/meaningfull merging order. In other words, the incorporation of the syntactic features in the pre-segmentation stage leads to the creation of more meaningful partitions, which is especially important at the higher levels of BPT which are close to the root. One can speculate that this should improve the intuitiveness and efficiency of user interactions e.g. by allowing the segmentation process make more "intelligent" decisions in areas of the scene where there is no explicit information provided by the users interactions.

To ensure that every part of the image can be split, if required by the user, the complete BPT (starting from the level of pixels) is used.

Propagation of Labels - Basic Idea

In the approach described in [38], once the image is pre-segmented and the BPT is created, the user can start interactions by adding "drags" to the original image. Each time a new drag is added the automatic process updates the segmentation mask. The basic idea behind this process is to assign regions coincident with mouse drag to that object. However, a region can be assigned to an object only if it coincides with mouse drags with only one label (only one object) - otherwise, a conflict occurs. The BPT is used to select the largest possible regions (areas) without such conflict (containing scribbles for only one object). Conflicts are identified by propagating labels of the mouse drags from leafs (pixels) towards the root of the BPT (whole image). The algorithm works in three main stages [107]: (i) assignment of labels from scribbles to leafs (pixels), (ii) upwards (bottom-up) propagation of node labels to their parents (remaining non-leaf nodes) if there is no conflict, and (iii) top-down propagation of labels so, in each zone of influence, child nodes have the same label as their parents. Note that, as a result of the above process, a certain number of nodes may remain without labels, judged as being "too different" with respect to the regions defined by the scribbles [38].

The most interesting of the above three stages is the bottom-up propagation. In the implementation suggested in [107] the propagation precedes from the lowest levels[12] of the BPT to the root. For each node from the current level, its associated label is propagated to its parent only if none of the sibling's descendants have been assigned a different label. When a node has two children with two different

[12]Level corresponds to the location of the node within the hierarchy (root is at the top level, its children are one level below, etc.).

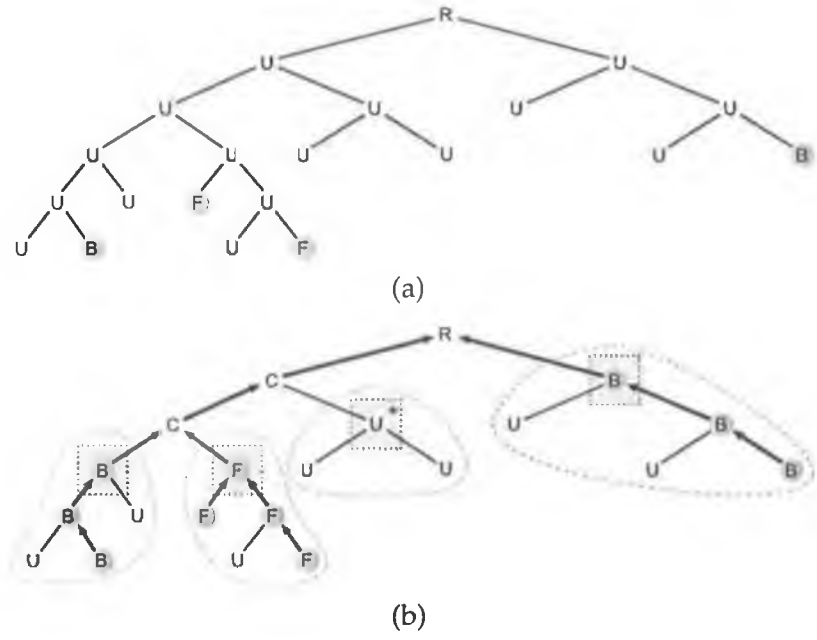


Figure 3.13: Label propagation, i.e. identification of the largest possible regions without conflict: (a) - Initial BPT. Labels (F), (B) and (U) correspond to foreground, background and unknown object respectively. (b) - BPT after label propagation, thicker lines indicate the upward label propagation. Labels (C) denotes conflict. Nodes surrounded by dotted squares (RZI) will be selected for the construction of the initial mask. The node marked with the star is an example of an area (zone) without a label.

labels conflict is detected. This results in the creation of the *zones of influence* (subtrees) for each label formed by nodes without conflict – see example from Figure 3.13. The highest nodes in such zones, referred to in the remainder of this chapter as *root of zone of influence* (RZI), can be identified by searching for non-conflict nodes whose parent is marked with the label indicating conflict. Assigning labels to all pixels in each zone, assuming that its RZI is labelled with a known label, requires top-down propagation starting from a given RZI. The above procedure leads to fast classification of large regions even if the provided scribbles consist of only a few pixels.

Author's Implementation of Label Propagation

The algorithm for bottom-up propagation proposed in [107] visits all nodes of the BPT (except leafs) proceeding level by level and updates their labels. Although, this algorithm may help in understanding the overall approach, it can be significantly simplified thereby reducing computational complexity. The main inconvenience in implementation of the original approach is caused by the requirement to organize nodes into levels. Moreover, the approach does not take advantage of the fact that it is not necessary to visit all nodes as conflicts can occur only on paths between labelled leafs (marked directly by the scribbles)

and the root. Below is the complete algorithm for bottom-up label propagation, developed by the author, which rectifies the above weaknesses.

The procedure begins by initialization of all leafs (pixels) with one of three labels: *Foreground* (F), *Background* (B) or *Unknown* (U) according to the drags made by the user. All remaining nodes are initialized as *Unknown* (U) – see example from Figure 3.13a. The propagation proceeds upwards along paths between labelled leafs (pixels marked by a mouse drag) and the root. Note that leafs are processed sequentially, i.e. for a given labelled leaf the complete path to the root is updated before processing the next leaf. Labels between a given labelled leaf and the root are updated as follows: For a given node (or initially a leaf) n, if its parent node is labelled as unknown (U) the label from the current node is assigned to the parent and then the procedure is repeated for the parent. If the parent node has been already marked (by propagation from another scribbled leaf (pixel)) with the same label as its child then the propagation stops (for this leaf). If a conflict between labels of the current node and its parent is detected then the parent is marked with label *Conflict* (C). Note that since in such a case the propagation towards the root continues, labels (F) or (B) assigned by propagation from another scribbled leaf processed before will be overwritten with label (C). Of course, the conflict propagation will stop if the parent is already marked with (C) (by propagation from other leafs). This makes the final result (result after processing all leafs) independent on the order in which leafs are processed. Below is the pseudo code of the propagation procedure for a single leaf:

INPUT: Leaf l with a known label (F or B)
VARIABLES: r - root
n - current node
p - parent of n
OUTPUT: Updated labels on the path from l to r
1. **START**: $n \rightarrow l$; $p \rightarrow n.parent$;
2. **WHILE** ($p \neq r$ **AND** $n.label \neq p.label$)
{
IF ($p.label = U$) **THEN** $p.label \rightarrow n.label$; **ELSE** $p.label \rightarrow C$;
$n \rightarrow p$; $p \rightarrow n.parent$;
}

The above algorithm is not only more straightforward to implement than the one proposed in [107], but also leads to a reduction in computational complexity. Since the propagation is performed only between scribbled leafs and the root, and typically the number of scribbled pixels is much smaller than the total number of pixels in the image, only a small percentage of all nodes from the BPT have to be updated. For instance, in the most demanding example from Figure 3.16 (CIF size) only 2% percent of all nodes had to be updated. This corresponds to a reduction in execution time for the bottom-up propagation

from 15ms to 1ms on an Intel Pentium III 600MHz when compared with the author's implementation of the original algorithm visiting all nodes. Although both approaches lead to real time execution on a modern PC this may become an important advantage when computational resources are limited, e.g. on handheld devices. An example of such propagation can be seen in Figure 3.14b which demonstrates an attempt at extracting the girl's face from the cluttered background.

Although, the above improvements seem very encouraging, the bottom-up propagation is only one of two main steps in the labelling algorithm. Moreover, it should be remembered that, the computational cost of the proposed solution depends on the number of scribbled pixels and in extreme cases, e.g. whole image covered by scribbles, the computational cost can be equal to or even exceed the cost of the original approach from [107]. However, taking into account that such situations should not occur in practice, the proposed solution is still an attractive alternative.

Labelling the Entire Image

Unfortunately, the BPT-based labelling procedure does not guarantee that all parts of the image will be classified as part of a known object. Some of the subtrees may not contain any leaf with a known label ((F) or (B)) – see the node marked with a star in Figure 3.13b. [38] concludes that since there is insufficient information regarding the unlabelled regions this is in fact an attractive feature of the approach and the unlabelled regions can be classified by the user defining additional markers. It also suggests that filling the entire image space can be useful only for specific applications.

The author's observations indicate that the above conclusion would hold only if the hierarchy (similarities) of regions represented by the BPT always represents the true (or at least expected by the user) semantic structure of the scene. In practice, the tree is constructed with limited homogeneity criteria which are often not sufficient to produce regions with a high probability of corresponding to semantic notions [38]. The main advantage of BPT is that it should represent only the most useful mergings of the regions [107]. However, in practice some of the important mergings may be missing. For example, Figure 3.14c shows an attempt to partition the complex scene from Figure 3.14a into a TV set and surrounding background. According to similarities stored in the BPT, the region corresponding to the girls hair is more similar to the region corresponding to the centre of the TV set than the region corresponding to the bottom of the TV set. The hair and the center of the TV set were merged first and the bottom of the TV set was merged to the already combined hair-TV set region. Since both the hair and the centre of the TV were marked with different labels, the region corresponding to the bottom of the TV set could not be labelled. In other

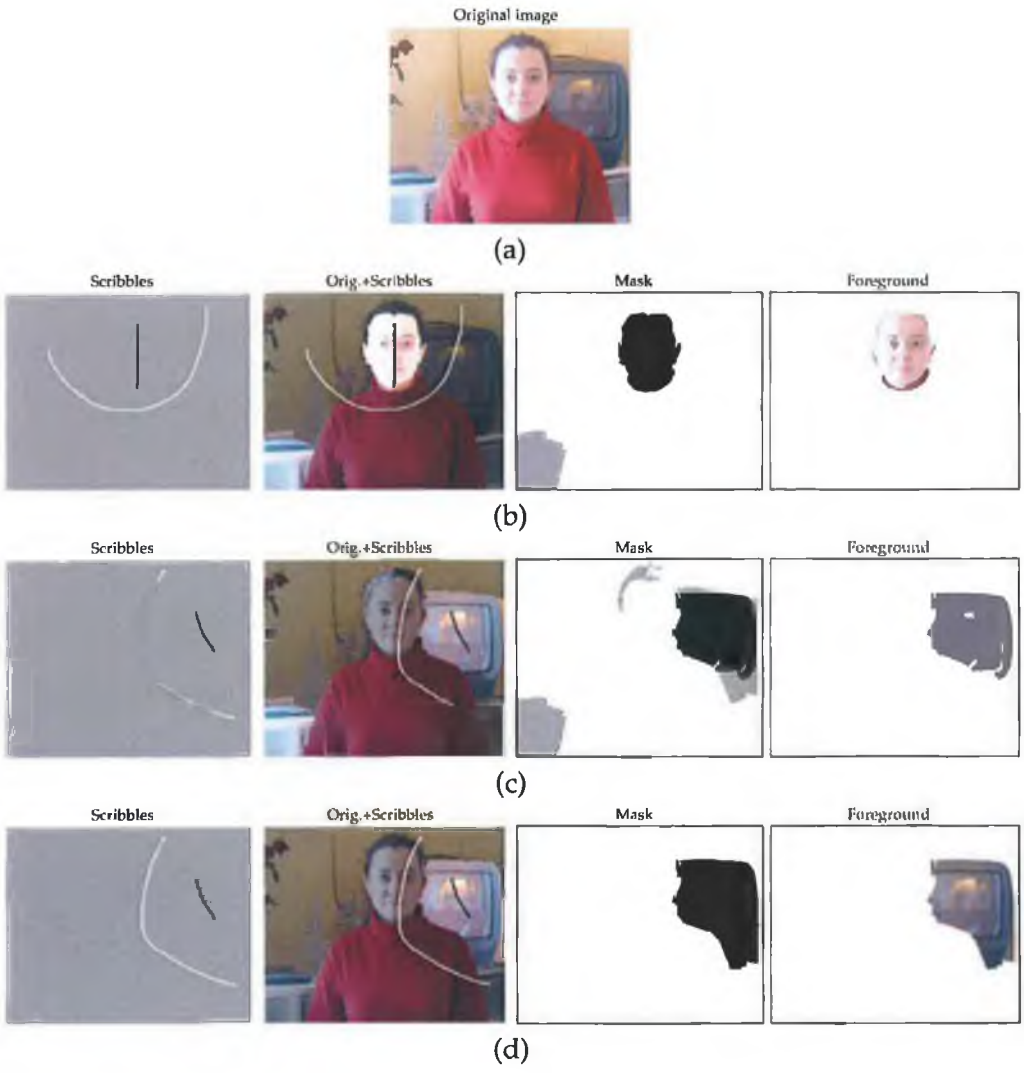


Figure 3.14: Examples of semi-automatic segmentation. Labels (F), (B), and (U) are represented by black, white, and grey colours respectively. (a) - The original image, (b) - Face extraction using only label propagation in BPT, (c) - Attempt at partitioning the scene from (a) into TV-set and background using only label propagation in BPT, (d) - Segmentation based on interactions from example (c) with filling of the entire image.

(a) (b)

Figure 3.15: Example illustrating limitations of the solution for labelling of the entire image proposed in [107]. The example shows a case where there is no labelled descendant of a sibling of an unassigned RZI (region 4) which is at the same time a neighbor of the RZI: (a) - Initial image partitioning; region 4 does not have any pixel adjacent with labelled regions 1 and 2, (b) - BPT.

words, the only similarity information stored in the BPT for this region was with a region labelled with conflict (region representing hair and the centre of the TV set combined). Similarly, many other regions from the example remained unlabelled due to conflicts occurring at lower levels of the tree. Therefore, even if filling the entire image space is important only for specific applications, in practice it may be advantageous to also correct mistakes caused by the above limitations of BPT. This could not only reduce the amount of additional interactions required, but would simplify presentation and in consequence interpretation of the results (especially for an unexperienced user). Of course, if the filling procedure does not perform according to the user's wishes additional interactions might still be required, but they should never exceed the effort required without such a procedure.

[107] suggests a solution for filling the entire space utilizing again BPT. For a given unassigned RZI, the algorithm starts from its sibling and scans all the descendants until one region that is, at the same time, neighbor of the RZI and labelled with a known label, is found. If several regions fulfill this criterion, one can be chosen arbitrary or on specific rules. Clearly, the use of the above approach for correcting mistakes caused by the limitations of BPT discussed earlier may be problematic since it is based on the same BPT. Moreover, it relies on the false assumption that there is at least one labelled descendant of the RZI's sibling which is at the same time a neighbor of the RZI. An example illustrating one such case when this assumption does not hold is depicted in Figure 3.15. Since the above assumption is not always true the filling procedure would have to be executed iteratively, i.e. start from nodes for which the assumption holds and continue until the entire image is labelled. In other words, utilization of an iterative labelling approach seems unavoidable in order to guarantee labelling of the entire image without exceptions.

To avoid such arbitrary decisions and exceptional cases an alternative solution was adopted by the author. The filling is performed by an iterative merging

process similar in concept to the region growing strategy in [81]. However the algorithm operates with RZIs created by the approach described in section 3.11.3 instead of pixels. Because of this the computational cost of the filling is very low. The procedure requires prior top-down propagation of not only labels of RZIs but also identifiers of all zones. The unlabelled nodes (regions) are iteratively assigned to the competing objects, formed by already classified regions. Simple non-parametric models of foreground (O^F) and background (O^B) objects comprise of RZIs with (F) and (B) labels respectively. Classification of an unassigned RZI region is performed by comparison of its similarities to adjacent regions from O^F and O^B. The order of labelling is based on the confidence with which the labels can be assigned. At each iteration, only one region, with the highest confidence, is assigned to an appropriate object. Then, the representation of the extended object is updated and the process continues until all regions are labelled.

In the current implementation, the distance between an unclassified region (RZI) R_i and an object O is computed as the shortest distance between R_i and one of the regions already assigned to O and adjacent to R_i:

$$D(R_i, O) = \min_{R_j \in O} d_r(R_i, R_j) \tag{3.30}$$

where $d_r(R_i, R_j)$ is computed as Euclidean distance between average colours represented in LUV space for two adjacent regions R_i and R_j. If the two regions do not have at least one pair of adjacent pixels $d_r(R_i, R_j)$ is set to ∞. Note that adjacency is not required for regions assigned in the first stage by label propagation therefore disconnected objects are still possible.

The confidence C^F with which an unclassified region R_i can be assigned to the foreground object is computed using the following formula:

$$C^F(R_i) = \begin{cases} \frac{D(R_i,O^B)}{D(R_i,O^F)+D(R_i,O^B)} & \text{if} \begin{pmatrix} (D(R_i,O^F)\neq 0 \vee D(R_i,O^B)\neq 0) \wedge \\ (D(R_i,O^F)\neq \infty \vee D(R_i,O^B)\neq \infty) \end{pmatrix} \\ 0.5 & \text{otherwise} \end{cases} \tag{3.31}$$

C^F has values in the range $[0, 1]$. Values of C^F above(below) 0.5 indicate that R_i should be assigned to the foreground (background) object. Of course, the confidence C^B can be defined analogically leading to the trivial relation $C^B(R_i) = 1 - C^F(R_i)$. Moreover, the confidence C with which R_i can be classified as foreground or background can be defined as $C(R_i) = \max\{C^F(R_i), C^B(R_i)\}$. At each iteration, values of C^F, C^B and C are computed for all unlabelled regions and only one region R_{max} which can be classified with the highest confidence $R_{max} = \arg\max_{\forall R_i \in O^U}\{C(R_i)\}$ is assigned to an appropriate object: foreground if $C^F(R_{max}) > 0.5$ or background otherwise[13]. Then, the confidence values are

[13] O^U denotes a set of all roots of unlabelled subtrees.

updated for all remaining unlabelled regions adjacent to R_{max} according to the new foreground/background models and the process is repeated iteratively until all regions are classified.

An example of the filling of the entire space from the previously considered example is shown in Figure 3.14d. Note, that the procedure utilizes all possible similarities between adjacent regions, not only the ones stored in BPT.

Summarizing this section, the most significant differences between the approach presented in [38] and the solution described in this thesis is the incorporation of syntactic features for the extraction of the BPT, a propagation algorithm optimized to provide maximum speed, and an alternative method for filling the entire space. The final approach combines the speed efficiency of the solution described in [38] with the capabilities of Seeded Region Growing [81] to fill the entire space.

3.11.4 Results

An example of user interactions and corresponding results are shown in Figure 3.16. For easier interpretation of the results the user is presented with a view of the final segmentation mask and a view of the foreground object. Additionally the foreground and background objects are brightened and darkened respectively in the original image.

The above approach has proven to be effective and practical. When testing the automatic approach, the pre-segmentation of a CIF (352x288) image takes under 3 seconds on a standard PC with Pentium III 600 MHz processor. This extra time is negligible[14] taking into account the benefit of practically instantaneous responses of the tool to subsequent user interactions. Rapid responses ensured a good user experience in cases where the interaction has to be applied iteratively in order to refine the segmentation result. In other words, the approach is in fact iterative in nature due to the speed with which the mask can be created following the interactions.

For comparison, the EM-based framework [32] required more extensive interactions to extract the Foreman from Figure 3.16f and produced subjectively poorer quality results. Also the approach based on morphological flooding [83] seems to produce worse quality results for this example.

Strict experimental evaluation of the time required to segment an image by both non-skilled and experienced users is outside the scope of this thesis. However it is safe to claim that a typical time spent by a user obtaining a required segmentation result is between 5-10 seconds. Clearly, the time required depends on user experience, complexity and size of the object under consideration and in

[14] In most cases it is below the time acceptable for opening an image before presenting it to the user.

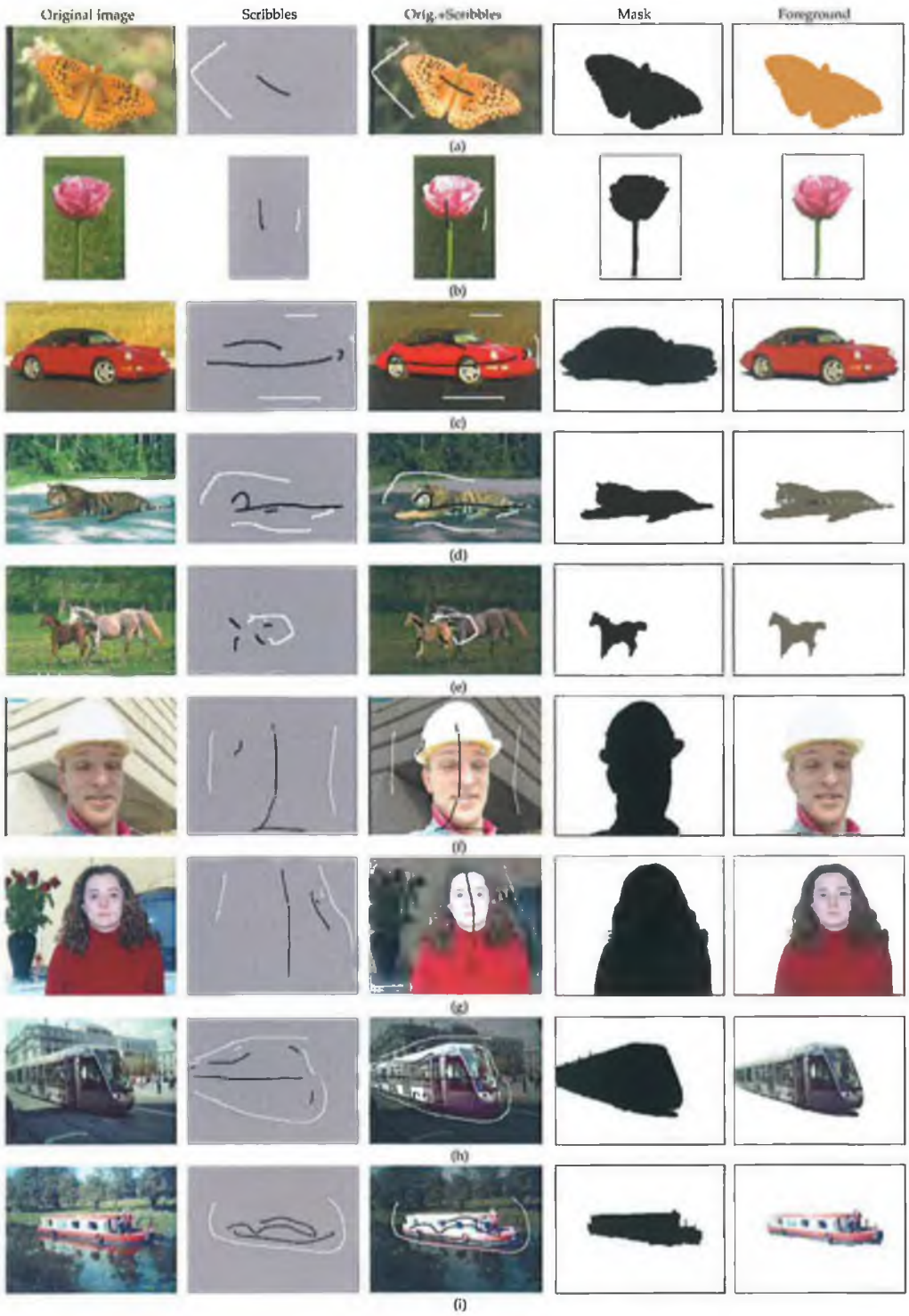


Figure 3.16: Results of semi-automatic segmentation.

some cases on the artifacts introduced by compression of the input image.

3.11.5 Discussion

The described tool has proven to be particulary useful in the context of the construction of statistical shape models which is described in chapter 7 where it is used for extraction of accurate contour examples from a set of images.

The restriction of the approach to two objects only was motivated by the fact that it leads to very intuitive user interactions in many application scenarios where there is only one dominant foreground object present in the scene. However, it should be stressed that, unlike in the case of other approaches, e.g. see [108], adaptation of the presented approach to multiple objects is trivial, i.e. would require mainly changes of the GUI.

The above semi-automatic segmentation approach can be used not only for off-line extraction of object contours, e.g. for construction of shape statistical models, but could be also particulary suited for fast formulation of object-based queries in object oriented retrieval by example systems, like the one proposed in [10]. The author believes that fast and easy to use semi-automatic segmentation tools will play a crucial role in the development of future retrieval-by-example systems. The above approach is particulary suited to such applications as it typically leads to good segmentation results within a few seconds of interactions.

Although the initial segmentation can be obtained very quickly, the author's observation indicate that the tool looses its advantage over other interaction strategies (for example contour adjustment described in [80]) in cases requiring very fine corrections of boundaries, e.g. whenever objects have very similar colour. In such cases scribbles have to be used to define both internal and external contours separating the objects. The most sensible solution seems to be integration of various interaction schemes within one tool allowing the user to freely select the most convenient way of interacting at different stages during the segmentation process, like for example in a tool described in [80]. Moreover, whilst the scribble-based interaction model is intrinsically suited for iterative refinement of the segmentation mask by adding new "drags", it is not equally convenient when the user wants to rectify his/her mistakes caused by a scribble extending to the wrong object. In the case of the tool implemented by the author, this can be achieved only by covering of the mistaken parts of the scribble with the correct type of scribble or by restarting the whole interaction process. The user experience could be further improved by extending the functionality of the GUI by providing means of undoing the last interactions and also selective deleting of whole scribbles.

The subjective feedback from a number of unexperienced users was very

positive. However the effectiveness of the tool should be evaluated by rigorous usability tests. Such rigorous evaluation is outside scope of this thesis and will be a part of author's future research in this area. It would be particularly interesting to quantify the influence of the syntactic extension to automatic segmentation proposed in this chapter on user impression, in terms of number of interactions and accuracy. Also the tool should be compared to other state of the art approaches – see section 2.5.1. Finally, the feasibility of porting the approach to state of the art hardware devices allowing new ways of user interactions, e.g. touch sensitive table, could be also investigated. It would be particulary appealing to compare the tool run on standard PC with mouse against tools run on such futuristic devices.

3.12 Conclusion

Several extensions to the well known RSST algorithm [19] are proposed allowing segmentation of images into large regions that reflect the real world objects present in the scene. The most important is a novel framework for integration of evidence from multiple sources of information, each with different accuracy and reliability, within the region merging process. Other extensions include a new colour model and colour homogeneity criteria, practical solutions to analysis of region's geometric properties and their spatial configuration, and a new simple stopping criterion aiming at producing partitions containing the most salient objects present in the scene.

All the above extensions and their evaluation can be seen as a feasibility study of utilizing spatial configuration of regions and their geometric properties (the so-called *syntactic features* [18]) for improving the quality of image segmentation produced by the RSST algorithm thereby increasing its usefulness to various applications. The "strength" of the evidence for region merging provided by the spatial configurations of regions and their geometric properties is assessed and compared with the evidence provided solely by the colour homogeneity criterion.

To the best of author's knowledge this chapter presents the first comprehensive and practical solution for integration of multiple features, including geometric properties, into a region merging process and demonstrates its performance in extensive experiments. It is shown that syntactic features can provide additional basis for region merging criteria which prevent formation of regions spanning more than one semantic object, thereby improving the quality of the segmentation. The experiments demonstrate that the proposed solutions are generic in nature and allow satisfactory segmentation of real world images without any adjustment to the algorithm's parameters.

The result of the study is a new computationally efficient method for fully auto-

matic image segmentation. The technique can produce both: single partitioning of the image containing the most salient objects present in the scene or hierarchical representation of the scene in the form of a *Binary Partition Tree*.

The chapter also describes utilization of the output of the proposed automatic approach to fast and intuitive semi-automatic segmentation.

CHAPTER 4

SHAPE REPRESENTATION AND MATCHING: A REVIEW

SHAPE analysis is an essential research topic in many areas. It plays an important role in systems for object recognition, matching, registration, and analysis. Unlike any other visual feature, e.g. colour or texture, the shape of an object can usually reveal its identity. Therefore, shape information could become a particulary important and powerful feature when used in object-based similarity search and retrieval. Not surprisingly, users are often more interested in retrieval by shape than by colour and texture [12]. However, retrieval by shape is still considered to be one of the most difficult aspects of content-based image search.

This chapter aims at outlining the most important issues related to shape analysis in the context of content-based image retrieval. In particular, it focuses on the major challenges related to 2D shape representation, matching and similarity estimation and gives an overview of selected techniques available in the literature in order to provide a context for the research carried out in the remainder of this thesis, particulary in chapter 5.

4.1 Introduction

4.1.1 Shape Analysis

Although challenging, shape analysis is an essential research subject in many areas including medicine, biology, physics, and agriculture, computer vision, image understanding, document analysis and many others [163, 17]. Countless vital visual tasks in the fields of computer vision and image understanding include some form of shape analysis. Intuitively, an object's shape often carries the most valuable information about objects present in analyzed visual content. Frequently, a vision system can deduce some important facts about

the visual scene only by exploitation of certain information about the shape of objects present. Shape is clearly an important cue for object recognition since humans can often identify semantic objects solely on the basis of their shapes or composition of primitives. This distinguishes shape from other elementary visual features (e.g. colour or texture), which usually do not reveal an object's identity.

Shape analysis, and representation and similarity estimation in particular, are key technologies allowing indexing and retrieving images based on their visual content. However, retrieval by shape is also considered to be one of the most difficult aspects of content-based image search. Despite three decades of intensive research there are many unsolved problems related to shape-based object detection and recognition, similarity estimation between shapes and utilization of shape information in content-based image retrieval. The most important of these are discussed in the remainder of this chapter.

4.1.2 2D Shapes Versus a 3D World

Although the real world is three-dimensional, reconstruction of 3D scenes from 2D views is often impossible. Moreover, despite recent advances in computer technology, dealing with 3D models is still computationally expensive, frequently to a prohibitive degree [163].

On the other hand, there is very strong empirical evidence provided by the field of cognitive sciences that a substantial part of the information present in an object view is carried by outline 2D shape [164, 165] which often conveys enough information to allow the recognition of the original object. [164] demonstrated experimentally the usefulness of silhouette representations for a variety of visual tasks, ranging from basic-level categorization to finding the best view of an object. Often, 2D silhouettes represent the abstractions (archetypes) of complex 3D objects belonging to the same class. For this reason, shape analysis, and analysis of 2D silhouettes in particular, are considered important components of many systems performing object recognition.

Contour-based matching methods can be very effective in applications where high intra-class shape variability is expected due to either deformations in the object (rigid or non-rigid, e.g. resulted from objects flexibility) or as a result of perspective deformations [16]. Moreover, dealing with 2D silhouettes is cheap from a computational point of view which is crucial when working with large collections.

Therefore, although the real world is three-dimensional the issue of two-dimensional shapes analysis is still of vital importance [163]. Methods for comparing 2D shapes are valuable building blocks in applications requiring object recognition and object-based retrieval from visual collections has been the

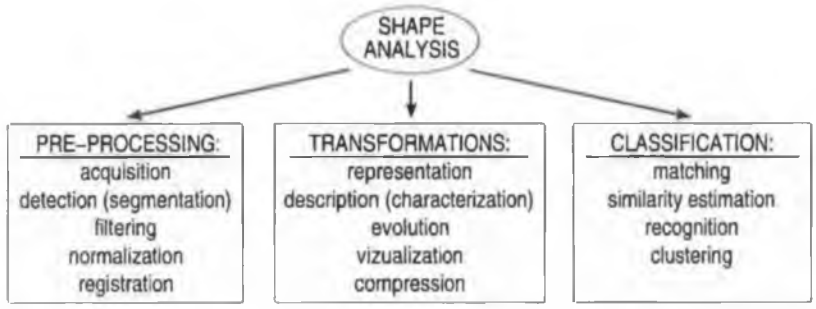


Figure 4.1: Typical problems addressed by shape analysis.

focus of much research.

For these reasons this thesis is primarily concerned with closed planar contours, particularly with aspects related to representation, matching and similarity estimation in cases where the similarity between the two curves is weak.

4.1.3 Chapter Structure

The next section outlines the most important aspects related to retrieval based on shape. Then, section 4.3 discusses a taxonomy of shape representations. Section 4.4 describes in some detail selected methods utilizing shape information embedded into a low dimensional vector space (feature vectors). The most characteristic curve matching methods which establish correspondence between local shape elements prior to similarity estimation are discussed in section 4.5. The review focuses on methods particulary relevant to the research carried out by the author in chapter 5.

4.2 Important Aspects of Retrieval by Shape

4.2.1 Computational Shape Analysis

In general, computational shape analysis and recognition requires addressing several problems which, according to [163], can be broadly divided into three classes: *shape pre-processing*, *shape transformations* and *shape classification* – see Figure 4.1.

From the above list the most important aspects of shape analysis in the context of the retrieval by shape are: (i) shape pre-processing, especially detection and registration, (ii) representation and description, (iii) matching, including establishing correspondence between shapes and estimation of shape similarity and (iv) organization of shapes into search structures [163].

Pre-processing is related to separation of the object of interest from other non-relevant elements of the scene. One of the most challenging, and (in many

applications) critical, of all pre-processing tasks is shape detection which is typically obtained by image segmentation. Due to its importance, selected issues related to image segmentation are outlined in chapter 2 and then further investigated in chapter 3. Shape registration is discussed in chapter 7.

The discussion in this chapter focuses on selected aspects of shape transformations and shape classification, particulary on shape representation, and establishing pairwise correspondence between shapes and similarity estimation.

4.2.2 Shape Representation and Description

One of the most important issues related to shape-based retrieval is extraction of information from shapes which is relevant for a given application. According to [163] this task can be divided into two closely related problems: (i) mapping of the initial shape representation (typically an ordered chain of points) into a representation more appropriate to a specific application, and (ii) extraction of relevant shape features (often referred to as descriptors) for subsequent classification, recognition or analysis of physical properties. In fact, descriptors are often extracted not directly from the initial sequence of points but from some more appropriate representation of the contour. The terms *shape representation* and *shape descriptor* are frequently interchanged in the literature however typically the term *shape descriptor* is used to indicate a compact representation which does not necessarily allow reconstruction of the original shape.

Although the choice of a good representation is task specific the most common criteria which should be considered when selecting a representation (or description) are [166, 147, 167, 17]:

1. *Scope* - This defines the class of shapes that can be described by the method;
2. *Invariance to rigid transformations (invariance to changes of pose)* - Rigid transformations do not change the shape of the object, therefore it is often desirable to use representations which are invariant to these transformations.
3. *Robustness to deformations (stability and sensitivity)* - This describes how sensitive a shape representation (or descriptor) is to changes in shape;
4. *Uniqueness* - This describes whether a one-to-one mapping exists between shapes and shape descriptors; The uniqueness of the representation determines *discriminatory power* of the whole matching technique.
5. *Accessibility (memory and speed efficiency)* - This describes how easy it is to compute a shape representation or descriptor in terms of memory requirements and computational cost;
6. *Scalability* - This means that a representation contains information about the shape at different levels of detail;
7. *Local support* - Only a representation computed using local information can provide an ability to recognize that the shape of a segment of a curve is the

same as the shape of another curve segment.

8. *Biological or psychophysically plausibility;*
9. *Capabilities to reconstruct the original shape;*

4.2.3 Shape Matching and Similarity Estimation

Shape matching refers to a broad set of methods for comparing shapes and is used typically for object recognition [147]. In fact, *shape matching* may consist of three distinctive processes, typically performed consecutively: (i) finding a transformation of one shape into another (often referred to as *shape alignment*), (ii) establishing correspondence between elements of two shape representations, and (iii) measuring the resemblance between shapes.

The complexity of each of the above three processes depends primarily on the shape transformations and deformations expected in a given application and on the representation used in a given approach. For example, finding a rigid transformation of one shape into another may be not necessary if the poses of the shapes are known or if the representation used is translation, scale and rotation invariant.

In the context of content-based image retrieval, one of the main objectives of shape matching is estimation of similarity between two shapes. Measuring the resemblance between shapes is a key tool in interpretation of images, determining an object's identity ("categorization") and in clustering.

Analogically to shape representation the desired properties of the similarity measure are application dependent, however it is possible to identify the most common properties appearing in literature [166, 168, 169, 147, 170, 171, 167, 17]:

1. *Correspondence to human perception/judgment of similarity* - Intuitively, this is the most important characteristic of a similarity measure. However it should also be noted that the perception of similarity is typically context and user dependent.
2. *Respect metric properties* - Many authors suggest that a similarity measure should respect metric properties [168] despite the fact that it is unclear whether the function underlying the human notion of shape similarity obeys metric properties [170, 171]. The main advantage of obeying metric properties is the fact that many computationally efficient methods finding nearest neighbors rely on the metric properties of the utilized comparison function.
3. *Invariance to transformations* - It is typically necessary that the similarity measure is invariant under a chosen group of transformations, e.g. *pose transformations* or *affine transformations*
4. *Robustness to certain type of deformations* - When studying various techniques

for shape analysis it is important to remember that it is not difficult to compare very similar shapes. The difficulties arise when one has to measure the degree of similarity between two shapes which are significantly different [169, 170].

5. *Continuity* - Some research suggested that shape similarity is a continuous phenomenon, i.e. similarity gradually deteriorates as one shape progressively deforms with respect to the other [168]. Although this property appears intuitive, its formal relevancy to human perception is still not obvious. [171] simply uses the term continuity to refer to robustness to various deformations: (i) perturbations, (ii) small cracks, (iii) blur, and (iv) noise and occlusions;
6. In [168] it was suggested that many "small" shape deformations should affect the similarity measure less than one "large" deformation of equal total magnitude and that different deformations should contribute differently to the similarity measure, e.g. the articulation of parts should contribute less than variations in the shape and size of portions of objects;

It should be noted that despite such a relatively exhaustive list of desired properties for similarity measures identified during three decades of intensive research, nobody has really been able to answer the basic question: *"what qualifies similarity of shape?"*.

Given that the above list of desired properties of shape matching contains several contradictory properties it should be apparent that it would be impossible to develop a technique suitable for any kind of problem. Not all methods are suitable for every kind of application, i.e. the choice depends on the shape properties, e.g. intra and inter class variability, and the particular application, e.g. acceptable computational cost or presence of noise [147].

The choice of the proper shape representation and matching method is often a compromise between retrieval effectiveness and computational complexity. The speed criterion is vital for CBIR systems since typically the user expects to have a response within a few seconds after he/she submits the query. Speed requirements often limit the number of techniques which can be used in CBIR systems. In particular, the matching should have low computational cost as it is typically expected to perform on-line. Also, descriptors should be extracted efficiently, especially in the case of large image databases, however, more tolerance is allowed in this case as feature extraction can typically be performed off-line.

Partial Matching in Large Collections of Shapes

Many recently proposed matching approaches can deal explicitly with various degrees of occlusions [172, 3]. Some can even match both open and closed

curves [170, 167]. Clearly, such properties can be very useful in many applications. Partial matching is typically considered more challenging than global matching due to the need to utilize appropriate local features and a higher computational cost of matching. However, in the case of shape-based retrieval in large collections, utilization of techniques performing partial matching has much more serious implications.

In a large collection it is very likely that for a given query many shapes can be found which share many very similar parts but which are globally dissimilar. The main question here is how to "punish" the occlusions in the final similarity measure. The danger is that such partially similar shapes could be ranked as being more similar to the query than other intuitively more similar shapes. An example of such situation can be found in [167] (page 1511) where the authors comment on their results in a following way: "Answer 10 looks dissimilar to the query. However, a closer look reveals that this shape matches the upper part of the query." Therefore, the treatment of occlusions is application dependent. Also, there are many applications that require matching of closed curves and where occlusions simply contribute to the dissimilarity between shapes. This thesis focuses on such cases.

4.3 Taxonomy of Shape Representations

As can be seen from the above discussion there are many important aspects of shape matching to be considered when choosing an approach suitable for a particular application. The large dimensionality of the problem and the vast number of methods available in the literature make consistent and comprehensive overview of the state of the art on shape matching difficult.

This section discusses a taxonomy of shape representations which is the most commonly used in literature to classify shape matching techniques. This taxonomy provides a starting point for the review of the most characteristic shape matching techniques presented in sections 4.4 and 4.5.

4.3.1 Taxonomy [173, 163]

Various taxonomies for shape representation methods were proposed in the past [173, 163]. Figure 4.2 shows the subdivision of shape representations proposed in [163]. The basic criterion for characterizing shape representations (and also matching) is based on whether it is *contour-based* (also known as *boundary-* or *outline-based*) or *region-based* (considers compositions of 2D regions and uses their internal details). According to [163], the *contour-based* approaches can be further subdivided into three classes: (i) *parametric contours* - silhouettes represented as parametric curves by ordered sequences, e.g. contour points,

Figure 4.2: Shape representation taxonomy according to [163].

(ii) *sets of unordered contour points* and (iii) *curve approximations* - boundaries represented by a sequence of geometric primitives fitted to the contour. The *region-based* techniques can be subdivided into: (i) *region decomposition* - regions represented as a set of simple primitives, (ii) *bounding regions* and (iii) *internal features*, e.g. regions represented by their skeletons. According to this taxonomy, a third basic class of shape representations referred to as *transform domain* techniques can be identified. However, it should be noted that this category contains both *contour-* and *region-based* methods since many transforms can applied to both 1D and 2D signals. The *transform domain* category can be subdivided into *linear* and *non-linear* and the *linear* category can be further subdivided into *single-scale* and *multi-scale* representations.

In [173] a somewhat different subdivision of representations was proposed. The two basic categories, *contour-based* and *region-based*, are further divided into *spatial domain* and *transform domain* sub-categories. The subdivision is made on the basis of whether the result of the analysis is numeric or non-numeric. For instance, *scalar transform* techniques produce numbers (scalars or vectors) as results, e.g. *Fourier Transform* [169, 174, 175] or moments [176, 177, 178, 179, 180]. In contrast, a *space-domain* technique may produce another image, e.g. *Medial Axis Transform* (MAT) [181, 182].

4.3.2 Contour-Based Versus Region-Based Techniques

The contour- and region-based methods offer two different notions of similarity, i.e. objects having similar spatial distributions may have very different outline contours and vice versa. Therefore the suitability of the methods from the above categories is strongly application dependent.

Utilization of objects' silhouettes in CBIR systems has strong motivation from studies of human visual systems. The field of cognitive science provides strong empirical evidence that the human visual system pays great attention to abrupt transitions in information, i.e. focuses on edges and tends to ignore uniform regions [183, 164, 165, 163]. Some studies even suggested that 2D outline shape

alone conveys enough information to recognize the original 3D object [164, 165].

On the other hand, in some applications, e.g. when dealing with logos and trademarks, it may be more convenient to represent shapes as compositions of 2D regions with their internal details (e.g. holes) rather than using their silhouettes. Also, region-based representations are somehow less sensitive to non-uniform boundary noise or occlusions than the contour-based approaches. However, many region-based techniques can deal only with rigid transformation since the elastic matching of 2D regions or even sets of points has often prohibitive computational complexity.

This thesis is primarily concerned with closed planar contours – particularly with matching and comparison in cases where the similarity between the two curves is weak.

4.3.3 Transform Domain Versus Spatial Domain Representations

Irrespective of the chosen subdivision convention it should be noted that representations from the *transform domain* category are embedded into a low dimensional vector space (feature vectors) therefore implying very straightforward similarity measures, e.g. Euclidean distance between feature vectors, while methods utilizing representations from the *spatial domain* categories first have to establish correspondence between elements from both shapes and then the final similarity is computed by accumulating local distances between the corresponding elements. In the case of methods from the *transform domain* category, similarity estimation typically has very low computational complexity which makes these approaches particulary attractive in CBIR applications where the similarities between the query and shape from the searched collection have to be computed on-line. However, identifying a set of measures which can completely and unambiguously describe shapes and at the same time are robust to elastic deformations has proven to be elusive [170]. In many applications, especially when elastic deformations or significant occlusions are expected, only methods from the *spatial domain* category can be used due to the utilization of local features. The price to pay for such flexibility is much higher computational cost of similarity estimation compared to the methods utilizing feature vectors. The subdivision into *transform domain* and *spatial domain* categories is the primary classification used for the review of the selected matching approaches presented in sections 4.4 and 4.5.

Closely related to shape representation is the issue of extracting relevant information (shape descriptors) from representations in order to to characterize an object. Shape description methods produce descriptor vectors (known also as feature vectors) from a given shape representation. Choosing a set of features from a shape representation to characterize an object often represents a challeng-

ing problem and the choice of features depends strongly on a specific problem. The following four categories of description methods were identified in [163]: (i) global measurements (see also section 4.4.1), e.g. area, perimeter, number of holes etc., (ii) signal processing (transforms) - where typically only a subset of the complete representation from the new domain is used as a feature vector, e.g. Fourier Descriptors as discussed in section 4.4.2, (iii) decomposition into primitives and (iv) description through data structures, e.g. binary trees.

4.3.4 Shape Modeling

In the past, object detection and recognition was performed solely using a single reference shape or several instances of shapes belonging to a given class used individually. However, as already explained in section 2.5.2, more than a decade ago Cootes and Taylor introduced one of the more influential ideas within the image analysis community, the so-called *Active Shape Model* (ASM) or *Smart Snakes* [70] which are deformable models with global shape constraints learned through observations gathered from multiple examples of shapes from the same class. Since then many new approaches for shape modeling have been proposed [116, 118, 120].

Such models (also often referred to as *statistical shape models*) are widely applicable to many areas of image analysis including: analysis of medical images [184, 185], industrial inspection tools, modeling of faces, identification of the model object in unseen images, e.g. hands and walking people [117, 186].

Clearly, some aspects of statistical shape modelling and utilization of such models for object detection and recognition are closely related to the problem of shape representation and matching discussed in this chapter. However, for clarity of presentation, this chapter focuses primarily on pairwise shape matching. Some issues related to statistical shape modeling and their utilization in top-down image segmentation were already discussed in chapter 2 (section 2.5.2). Construction of ASM is described in some detail in chapter 7.

4.4 The Most Characteristic Methods Utilizing Shape Information Embedded Into a Low Dimensional Vector Space

This section describes three common techniques for measuring the similarity between shapes based on shape representations embedded into a low dimensional vector space (feature vectors). In such a case, the matching process reduces to a very straightforward similarity estimation, e.g. by using Euclidean distance between feature vectors.

4.4.1 Simple Global Geometric Shape Features

Often, coarse estimation of shape similarity, and/or shape classification, can be obtained based on a comparison of simple scalar measures derived from the global shape. Some simple global shape features are perimeter, area, aspect ratio, elongation, eccentricity, circularity, rectangularity, complexity, symmetry, sharpness, number of holes and number of corners.

The obvious advantage of using simple shape descriptors is faster calculation of similarity and general applicability [187]. Although these features are typically easy to compute, allow fast comparison between shapes and are generally applicable to a wide class of shapes [187] they also often result in many false positives [175] and this is unacceptable in many applications.

As a matter of fact, the majority of systems for CBIR dealing with general content currently available in the literature utilize only simple global shape features [11, 52, 29, 44] (see also section 2.2.1). For example, one of the most recent systems such as the *SCHEMA Reference System* [188] and a system described in [10] use simple global shape descriptors which are a subset of the contour-based descriptor standardized in MPEG-7[1].

4.4.2 Fourier Transform

Very often, analyzed shapes do not have holes or internal markings and the associated boundaries can be conveniently represented by a single closed curve parameterized by arc length. In many such cases a useful representation of a curve boundary can obtained using the 1D *Fourier Transform* [169, 174, 175, 173]. The *Fourier Descriptors* (FD) represent silhouettes in the frequency domain as complex coefficients of the Fourier series expansion of some function derived from the boundary (parameterized boundary signatures).

The extraction of FD usually starts by obtaining a 1D representation of the contour (often referred to as a shape signature), which is derived from the coordinate chain, e.g. curvature, centroidal distance (radius) or complex coordinate functions [169]. Then, complex Fourier coefficients are computed by performing a *Discrete Fourier Transform* (DFT) of the shape signature. Typically, only a subset of all coefficients is used for classification purposes. Translation invariance of the FD is achieved by using signature functions which are invariant to translation. Scale normalization is typically obtained by dividing all coefficients by the magnitude of the coefficient corresponding to the size of the shape, e.g. the zero'th coefficient in the case of shape radii or the first coefficient in the case of complex coordinate functions [169]. Rotation invariance is usually

[1] Intended to complement the CSS-based descriptor discussed in section 4.5.5.

achieved by taking only the magnitude information from Fourier coefficients and ignoring the phase. The Fourier transformation captures the general feature of a shape by converting the sensitive direct representation into the frequency domain where the lower frequency coefficients contain information about the general shape, and the higher frequency coefficients contain information about smaller details. The similarity between objects' silhouettes is typically measured using the Euclidean distance between their normalized feature vectors.

Methods utilizing Fourier descriptors are easy to implement and are based on the well developed theory of Fourier analysis [147]. The description is reasonably robust to noise and geometric transformations. Moreover, similarity estimation is very straightforward and has low computational cost which is crucial for allowing real-time querying in large collections. The major limitation of the Fourier descriptors is their rather poor performance in similarity retrieval [3, 17] when dealing with elastic deformations or non-uniform noise. In other words, often the similarity measure based on FD does not correspond to human judgements of similarity. This can be explained by the fact that the frequency terms have no apparent equivalent in human vision and the fact that Fourier coefficients do not provide local spatial information.

4.4.3 Moments

Moments or functions of moments are among the most established and frequently used shape descriptors, mainly due to the fact that they can capture global information about the shape image and do not require closed boundaries. They are particulary attractive in applications dealing with complex shapes that consists of holes in the object or several disjoint regions, e.g. logos and trademarks. They are typically computed from whole 2D regions (gray-level or binarized images), however, they can be also extracted solely from shape boundaries.

The most classical type of moments are *regular moments* which for digital images are defined as:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \tag{4.1}$$

where m_{pq} is the (p, q)-moment of an image function $f(x, y)$.

Translation invariance can be achieved by using *central moments*. The infinite sequence of moments $p, q = 0, 1, ...,$ uniquely determines the shape, and vice versa. Typically, high-order moments, which are usually very noisy, are discarded and only a limited number of low-order moments are included in the feature vector and consequently used for similarity estimation and/or classification [171, 179].

Although central moments were shown to be useful, they are only invariant to translation. A more general form of invariance is given by seven rotation,

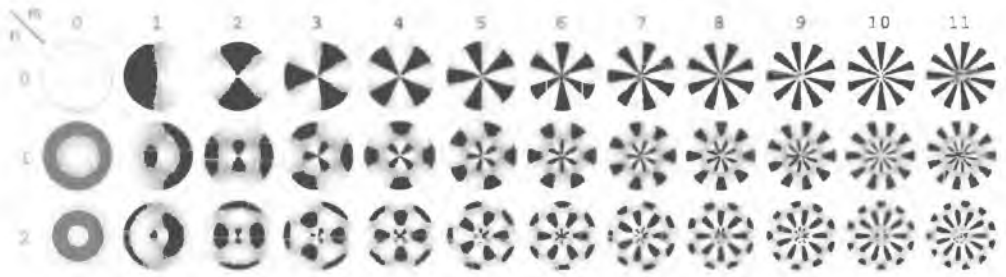


Figure 4.3: Real part of ART 36 basis functions (twelve angular and three radial). Imaginary parts are similar except for quadrature phase difference [3].

translation, and scale invariant moment characteristics introduced by Hu [176] and often referred to as *moment invariants*. These are seven nonlinear functions defined on regular moments derived using the theory of algebraic invariants in rectangular coordinates. Although attractive due to their more general form of invariance, typically only six features (moment invariants) are utilized since generating a bigger number is not a trivial task [177]. This significantly limits their discrimination power.

The physical meaning of moments (or moment invariants) is not clear, except in some low-order cases. The approaches based on them tend to return too many false positives [175], which is a result of their low discriminatory power. In fact, generally moments are sensitive to digitization error, camera non-linearity, non-ideal position of camera and to minor shape deformations, particulary non-linear deformations [178]. Another disadvantage of the methods relying on moments is their high computational cost of extraction since features are computed using the entire region, including interior pixels.

Moreover, as explained in [177], the definition of regular moments has the form of projection of function $f(x, y)$ onto the monomial $x^p y^q$, the basis set which is not orthogonal. This implies a certain degree of redundancy between m_{pq} and makes the recovery of the original image from these moments quite expensive. To overcome the above problems, orthogonal moments were proposed among which the *Zernike Moments* [177] and moments resulting from *Angular Radial Transform* (ART) [17] are by far the most popular. The remainder of this section focuses on the ART as it has been adopted as a region-based shape descriptor in the MPEG-7 standard. In fact, both methods are quite similar with the main difference lying in the basis functions used: Zernike Moments are derived from Zernike polynomials while ART moments are based on cosine functions.

The ART-based descriptor is computed by decomposing the shape image into orthogonal 2D basis functions. The ART is an orthogonal unitary transform whose complex basis functions are defined on a unit disk by a complete set of orthonormal sinusoidal functions expressed in polar coordinates. The ART basis function are separable along the angular and radial directions – see Figure 4.3.

Prior to the extraction of ART coefficients, the shape is normalized with respect to translation and size by "inscribing" into the unit disk. Rotation invariance can be achieved by taking only the magnitudes of the ART coefficients. ART gives a compact and efficient way to express pixel distribution within a 2D object region. The details regarding the format of the descriptor defined by the MPEG-7 standard can be found in [5, 17]. Similarity between shapes is calculated by summing the absolute differences between all descriptor elements (size-normalized ART coefficients) [5].

As pointed out in [17], region-based techniques like ART are generally less sensitive than contour-based methods to extreme distortions due to scaling, but at a price of significantly worse performance in similarity-based retrieval. This can be explained by the fact that there is no evidence for any analogy between ART and the human visual system. Also, the ART coefficients, like all other region-based descriptors, are sensitive to elastic (non-linear) deformations and the initial normalization, e.g. outliers may affect the process of "inscribing" the shape into the unit disk.

4.5 The Most Characteristic Curve Matching Techniques

As already explained in section 4.2.3, shape similarity can be estimated based on shape representations embedded into a low dimensional vector space (feature vectors) or based on distances between local features associated with corresponding elements from the two shapes. This section gives an overview of shape matching techniques which utilize representations from the *spatial domain* category supporting local features. The review focuses on various representations from this category, methods used for solving the correspondence problem between elements of two representations and potential applications of the discussed techniques.

Methods utilizing representations from the *spatial domain* category first have to establish correspondence between elements from both representations, taking into account shape variations and pose transformations, and the final similarity is then computed by accumulating local distances between the associated elements. Due to the utilization of local features and advanced matching procedures such methods can often deal with significant elastic deformations or occlusions. The price to pay for such flexibility is much higher computational cost of matching compared to the methods utilizing descriptors embedded in low dimensional feature vectors.

There are many important aspects of shape matching based on local dissimilarities of elements from both representations such as: (i) identification of basic shape elements (shape segmentation), (ii) choice of local features (attributes) associated with each shape element, and as mentioned already (iii) quantification

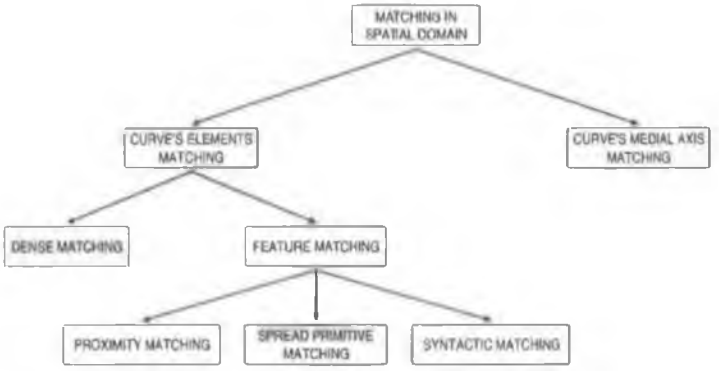


Figure 4.4: Taxonomy of curve matching in spatial domain.

of local and global similarities, (iv) establishing of the correspondence between elements from both representations.

For clarity, this review is primarily based on the classification of matching techniques described in [170] – see Figure 4.4. Many authors [170, 189] make a primary distinction between the majority of the silhouette matching techniques which focus on the curve itself, and approaches based on the curve's *medial axis*. Techniques matching curve elements are then further categorized into either *dense matching* or *feature matching* (sometimes referred also as *transient event matching*). [170] further divides techniques from the latter category into three groups: (i) *proximity matching*, (ii) *spread primitive matching*, and (iii) *syntactic matching*.

It should also be noted that some matching methods discussed in the following sections are applicable, or can be extended, to segmentation and identification of objects in unseen images. This task is closely related to the the so-called top-down segmentation problem discussed in section 2.5.

4.5.1 Medial Axis Transform

A representation that has proven to be relevant in human vision is the *Medial Axis Transform* (MAT) proposed originally by Blum [181, 182]. MAT produces a skeleton and a width value at each point on the skeleton (the so-called *quench function*) [190, 147]. Medial axis are readily computed from contours and provide a topological description of object boundaries which are relatively stable over changes in viewpoint. This approach led Sebastian et. al. [191] to attempt to capture the structure of the shape in the graph structure of the skeleton. The similarity between shapes is computed based on the "edit" distance between shock graphs.

Sebastian et. al claim that their approach based on shock graphs is efficient and robust to various transformations including articulation, deformation of parts and occlusion. The method appears to offer graded similarity measure and may

be combined with appropriate indexing methods. However, matching graphs involves solving a NP-complete isomorphism problem which is computationally very expensive. Therefore, currently, two contradicting views regarding the usefulness of the graph-based approaches can be found in the literature. While [191] claims efficient matching using graph-based approaches and superiority over curve matching, others [170, 192] reported sensitivity of the graph representation to occlusions and small contour changes and also prohibitive matching complexity.

Given the above difficulties with graph representations, a natural alternative is to represent shapes using their curves elements. The following sections discuss the main categories of methods measuring similarity directly using elements from both curves.

4.5.2 Dense Matching

Dense matching is one of the broadest categories of methods measuring similarity directly using elements from both curves. The curves are typically represented using an ordered sequence of equidistant contour points where the number of points is chosen to satisfy a predefined representation error. Dense matching methods perform a dense mapping between the two dense curve representations. The task is often formulated as a parameterization problem with some cost function to be minimized. The final dissimilarity between contours is obtained by summing the cost of local deformations. A common approach is to define the cost function as an "elastic energy" needed to transform one curve to the other [168, 193, 194], where the "elastic energy" is defined in terms of contour curvature. Another common approach is to use the so-called *turning function* [195] instead of curvature. Although many of the approaches from this category allow "elastic matching", e.g. by utilizing *Dynamic Programming* (DP) for finding the optimal mapping between the curves, some assume only rigid deformations [195].

The main advantages of the methods from this category is that they are rather easy to implement and facilitate neatly continuous measures of similarity. However, the author is not aware about any rigorous quantitative evaluation of the retrieval capabilities of any method from this category using a collection containing more than 10 curves. This recent lack of interest in the methods from this category can probably be explained by the common belief that they have high computational complexity compared to feature (event) matching methods [170], and that the methods utilizing curvature are inherently size variant [170].

After analyzing many matching techniques, the author tends to disagree with some of the above opinions. Although the dense matching techniques deal

with dense representations, the algorithms which can be used for establishing the correspondences between curve elements are typically much simpler and have much lower order of complexity than the methods able to deal with sparse representations. Effectively, only a few *feature matching* methods may compete in terms of speed with the *dense matching* methods and only at the cost of utilization of very sparse and therefore often ambiguous representations. Moreover, recently new exact pruning methods for *Dynamic Time Warping* (DTW), which is often utilized by *dense matching* approaches, were proposed in [196] making the dense matching methods even more attractive in terms of fast search in large collections. Moreover, in the case of closed contours, size invariance can be often achieved by size normalization performed prior to the matching, even in cases of modest occlusions.

However, the existing *dense matching* methods do have a serious drawback which is related to the fact that they operate at only one scale, e.g. in [168] the curvature is computed after smoothing the contour by some fixed amount to reduce the influence of the noise. It should be noted that the contour matching method proposed in chapter 5 of this thesis can be categorized as a *dense matching* approach which, unlike other methods from this category, utilizes a rich multi-scale representation.

4.5.3 Proximity Matching

Proximity methods match two sets of unordered key points (feature points) by searching simultaneously for pose alignment and correspondence such that the distances between corresponding key points of each shape are minimized [197, 198, 199, 200]. These methods are attractive in applications where the objects can be assumed to be rigid and the order of points is unavailable. Many techniques from this category can be easily extended to the problem of measuring the similarity between a model and a portion of the image. As a consequence they can be applied for object detection in cluttered scenes without the need for object segmentation [198, 199]. Unfortunately, as is also commonly known, proximity measure is inadequate for weakly similar boundaries, i.e. treating curves as sets of points ignores more qualitative "structural" information [170].

One of the most typical examples of proximity matching methods is the *Hausdorff Distance* [198, 201, 171] which is a measure of resemblance between two arbitrary sets of geometric points superimposed on one another. The Hausdorff distance between two finite point sets $\mathcal{A} = \{a_1, a_2, \ldots a_p\}$ and $\mathcal{B} = \{b_1, b_2, \ldots b_q\}$ is defined as: $H(\mathcal{A}, \mathcal{B}) = \max\{h(\mathcal{A}, \mathcal{B}), h(\mathcal{B}, \mathcal{A})\}$ where the function $h(\mathcal{A}, \mathcal{B})$, often called *Directed Hausdorff Distance*, is computed by finding for each point of $\mathcal{A}$ the distance to the nearest point of $\mathcal{B}$ and taking the largest of such distances as $h(\mathcal{A}, \mathcal{B})$. The Hausdorff distance is sensitive to noise since it can be influenced by a single outlier. A similar measure but without this drawback is the *Partial*

Hausdorff Distance [171]. An important property of the Hausdorff distance is that it can be meaningfully applied to sets of different sizes, i.e. where there is not one to one correspondence between all points. The directed partial Hausdorff distance can even be applied for partial matching [198, 201]. However, as in the case of all proximity matching techniques, computing the Hausdorff distance under pose transformation is very computationally expensive.

An elegant alternative solution from this class was proposed by Shapiro and Brady [197] and later extended by Sclaroff and Pentland [200]. In these approaches points are mapped to an intrinsic invariant coordinate frame. Due to their significance in shape registration their approach is further discussed in section 7.1.2.

4.5.4 Spread Primitive Matching

Spread primitive matching refers to problems where the analyzed curves are divided into shape elements, or primitives (e.g. straight lines) but the information regarding the order of such primitives is unavailable. Many approaches from this category seek the largest set of matches between elements from both shapes inducing compatible coordinate transformation (pose). The methods from this class are relatively robust to noise. However, similarly to proximity matching techniques, they are also inadequate for weakly similar boundaries, i.e. shapes which underwent non-linear ("elastic") transformations.

A common method for finding an objects pose in the abovementioned way is the *Generalized Hough Transform* (GHT) [202]. In this technique, each pair of model and image primitives (such as edges or vertices) defines a range of possible rigid transformations from a model to an image (pose of the model with respect to the image). GHT works by accumulating independent pieces of evidence, coming from each of such pair of primitives, for possible coordinate transformations (pose) in a quantized space of transformation parameters. Alternative approaches from this category represent object primitives as nodes in a graph with the connecting arcs representing their relations. The properties of the primitives are then encoded as the node attributes and the matching problem is formulated as attributed relational graph matching [203].

4.5.5 Syntactical Matching

In applications where the ordering information of curve primitives is available the so-called *syntactical representations* could be used [170], where a curve is represented by an ordered list of shape elements having attributes like length, orientation, bending angle, etc.

Some early approaches from this category relied on pose invariant attributes

of the shape elements to ensure invariance to geometrical transformations of the whole matching algorithm [204, 172]. However they completely ignored the problem of resolution (or scale), e.g. smoothing may result in an entirely different list of curve primitives.

More recently, the resolution problem has been addressed in two distinctive ways. Some approaches utilize a cascade of representations at different scales [13, 14, 189, 3]. In other words, they attempt to match and quantify similarity based on matching transient events occurring at different scales of the multi-scale representations. The most characteristic methods from this category are based on the *Curvature Scale Space* (CSS) introduced first by Witkin [205]. Other approaches, inspired by string comparison algorithms, emulate smoothing during the matching process [170, 167] instead of using pre-computed multi-scale representations. They use edit operations (merging, substitution, deletion and insertion) to transform one string into the other. The operators are designed to handle noise and resolution changes.

The remainder of this section describes the most representative methods from both categories since they are the methods used to benchmark the author's approach. The discussion focuses on potential drawbacks of syntactical matching and aims at providing context for the research carried out by the author and reported in the next chapter.

Curvature Scale Space (CSS)

Scale-space filtering for 1D functions was originally introduced by Witkin [205] and used in several approaches for contour matching [166, 13, 14, 15, 206], modeling [189], and even robust corner detection [207] and fast snake propagation [17].

The *Curvature Scale Space* (CSS) image was first proposed as a representation for planar curves by Mokhtarian and Mackworth [166]. In this approach, a parametric representation of the curve is convolved with a Gaussian function with increasing standard deviation σ. The CSS image is created by tracking the position of inflection points, identified as zero curvature crossing points, across increasing scales and along the contour.

Assume the parameterized contour $C(u) = \Big(x(u), y(u)\Big)$, where u is a variable measured along the contour from a starting point and $x(u)$ and $y(u)$ are periodic. $C(u)$ is convolved with a Gaussian kernel ϕ_σ of width σ:

$$X_\sigma(u) = x(u) \otimes \phi_\sigma(u) = \int_{-\infty}^{+\infty} x(t)\phi_\sigma(u-t)dt, \qquad \phi_\sigma(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}} \quad (4.2)$$

and the same for $y(u)$.

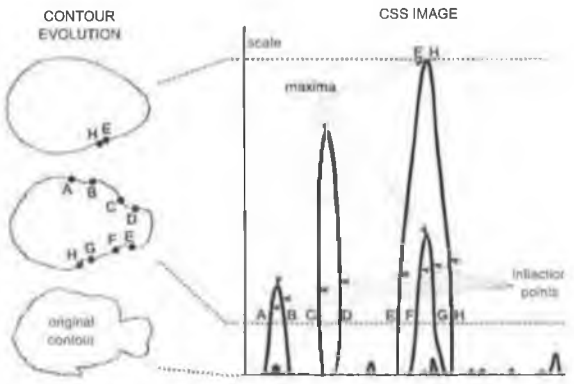


Figure 4.5: Extraction of CSS image for a fish contour [207].

Curvature at different levels of smoothing is computed as:

$$K_\sigma(u) = \frac{X'_\sigma Y''_\sigma - X''_\sigma Y'_\sigma}{((X'_\sigma)^2 + (Y'_\sigma)^2)^{\frac{3}{2}}} \tag{4.3}$$

where X'_σ, Y'_σ and X''_σ, Y''_σ are respectively the first and the second derivatives of X_σ and Y_σ

An extraction of a CSS image for a fish example from [207] is illustrated in Figure 4.5. At a given scale the position of inflection points along the contour can be determined by locating curvature zero-crossings. With increasing value of σ, the resulting contour becomes smoother. The curvature zero-crossings continuously move along the contour until two such zero-crossings meet and annihilate. The number of zero-crossings of the curvature decreases until finally the contour is convex and the curvature is positive. The CSS image of a curve is in a fact binary image computed by recording these positions of the curvature zero-crossings in the $u-\sigma$ plane for continuously increasing σ. The size of each of the arch-shaped contours in the CSS image, often referred to as lobes or peaks, represents the depth and the size of the corresponding feature on the original contour.

The CSS images are invariant under translation. An object's rotation or a change in the starting point of the contour results in a circular shift of its CSS image which can be taken into account during the matching. Scale invariance can be obtained by normalizing CSS images according to the size of the original curves. Noise may result in some small lobes at the lowest scales of the CSS images but the major ones are usually unaffected.

The above properties make CSS images very attractive for extraction of compact descriptors potentially useful for efficient estimation of similarity in the context of CBIR. Therefore, not surprisingly, several approaches used the CSS images for that purpose [166, 13, 14, 189]. By far the most commonly used matching technique utilizing CSS images was proposed by Mokhtarian and Mackworth [13, 14, 15, 206].

Their approach relies on a very compact descriptor containing only the positions of annihilation of inflection point pairs in the $u - \sigma$ plane, i.e. the positions of the zero-crossing point maxima [13, 14, 15, 206]. They can be located in the example from Figure 4.5 at the tops of the lobes. (u, σ) coordinates of these maxima express the position of concavity or convexity along the contour and its significance. Since small lobes are produced by the noise only the most significant ones have to be retained, e.g. not smaller than 10% of the highest lobe. Similarity estimation between two curves requires solving a relatively simple correspondence problem between their two sets of maxima. To ensure invariance to rotation and change of staring point the matching algorithm has to search for the optimal horizontal translation between the two sets of maxima. Mokhtarian and Mackworth propose a fast coarse-to-fine matching strategy. Detailed descriptions of different versions of the matching algorithm can be found in [13, 14, 15, 206]. A description of the most recent and effective variation of the technique can be found in [17].

The above shape descriptor is very compact, allows fast matching, and extensive tests using general shape collections [208] revealed that the method is quite robust with respect to noise, scale and orientation changes of objects[2] Because of these advantages the CSS-based shape descriptor has been included as the contour-based shape descriptor in the ISO/IEC MPEG-7 standard. A detailed description of the MPEG-7 shape descriptors and the most recent matching algorithms can be found in [17].

Despite its extensive advantages, the above description has one serious drawback which is its potentially high degree of ambiguity allowing only coarse comparison between shapes [209]. Specifically, the positions of zero-crossing point maxima for very deep and sharp concavities and for very long shallow concavities may be identical. Moreover, the convex segments of the curve are represented only implicitly by assuming that every concavity must be surrounded by two convexities. As such, it is impossible to use the CSS image, and therefore any descriptors extracted from such image, to distinguish between totally convex curves (i.e. circles, squares, triangles). The last drawback can be somehow resolved by including global descriptors, e.g. eccentricity and circularity.

Finally, some problems with the robustness of CSS images were reported by Ueda and Suzuki in [189]. Figure 4.6 shows an example from [189] where small shape changes result in drastic structural changes in the corresponding CSS image. Clearly, the above limitation could affect the coarse-to-fine matching of CSS maxima proposed by Mokhtarian and Mackworth. Surprisingly, the most recent publications regarding CSS images suggest their stability – see for example [17].

Since the coarse-to-fine strategy may not be adequate for matching of heavily

[2]The capabilities of the method can be checked at "www.surrey.uk".

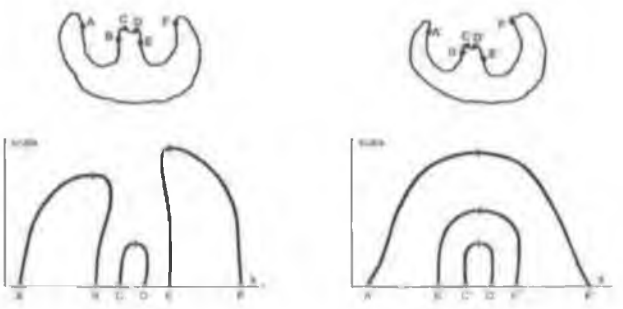


Figure 4.6: Example of radical change in a CSS image caused by a minor change in shape [189].

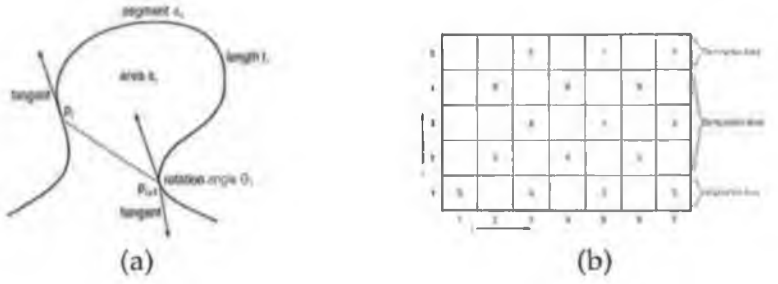


Figure 4.7: Matching approach proposed in [167]: (a) - Geometric quantities for representing a segment, (b) - Dynamic Programming (DP) table.

deformed shapes Ueda and Suzuki proposed a rather complex approach, where inflection points are matched from the lowest scale to the highest [189]. Such a fine-to-coarse convex/concave structure matching algorithm was utilized for automatic acquisition of shape models by generalizing the multi-scale convex/concave structure of a class of shapes [189].

Matching and Retrieval of Contours by Petrakis et. al. [167]

An advanced syntactical contour matching algorithm that emulates smoothing during the matching process rather than using a pre-computed multi-scale representation was recently proposed in [167]. The main feature of the approach is that it operates implicitly at multiple scales by allowing matching of merged sequences of consecutive small segments in one shape with larger segments of another shape, while being invariant to translation, scale, orientation, and starting point selection. The approach can handle both open and closed curves and can deal with noise, shape distortions, and possibly occlusions.

In this approach, the original curves are segmented into convex and concave segments based on inflection points identified based on approximation using local cubic B-splines. Local dissimilarities between segments are computed based on three geometric quantities (segment invariant attributes) defined at the inflection points which are rotation angle, length and area – see example from Figure 4.7a. The method operates by allowing matching of merged

sequences of segments from one shape with larger segments in the other shape. The explicit computation of the scale-space representation is avoided due to the introduction of a mechanism for computing the attributes of the merged segments. Correspondences between similar parts of the two shapes are found using a *Dynamic Programming* (DP) algorithm. The matching starts by building a DP table, where rows and columns correspond to inflection points from both shapes (Figure 4.7b). Starting at a cell at the bottom row and proceeding upwards and to the right, the table is filled with the cost of the partial match containing the segments between the inflection points examined so far. Because convex segments cannot match concave ones, only about half the cells are assigned cost values, in a checkerboard pattern. Merges, where a sequence of segments of one shape matches a single segment of the other shape introduce "jumps" in the traversal of the DP table. Reaching the top row implies a complete match.

The authors of the above approach claim that through merging at different levels of details their algorithm avoids the costly smoothing operations of the scale-space based approaches. However, this leads to the need for searching for the best match at multiple scales during the matching process, implying extremely high matching complexity, e.g. in the case of optimal matching $O(M^3N^2)$, where M and N denote the number of segments in each shape. For instance, a sequential search of a database of 1,100 shapes of marine life species [14] takes several minutes on a standard PC (Pentium PC 1000MHz) [167].

It was shown in [167], that the method outperforms simple techniques such as the one based on moments but produces comparable results to the CSS-based methods (which have typically much lower matching cost) when tested on the database of 1,100 closed contours of marine life species [14]. Interestingly, when tested on smooth shapes, such as the shapes in the GESTURES dataset [210] the method performed slightly better for large answer sets and slightly worse for small answer sets than the method based on Fourier Descriptors.

Moreover, it is not clear from [167] how meaningful a comparison would be between two totally convex curves or weakly similar curves where some segments may not have an intuitive correspondence. Finally, although simpler that the technique proposed in [189] or in [170] it still appears quite complicated and nontrivial to implement, making the comparison with other methods difficult.

An interesting syntactical matching algorithm, sharing many commonalities with the approach by Petrakis et. al. [167], was also proposed by Gdalyahu and Weinshall [170].

4.6 Conclusion

This chapter outlined the most important aspects of retrieval based on shape and discussed major challenges related to 2D shape representation, matching and similarity estimation. It aimed at giving an overview of selected techniques available in the literature in order to provide a context for the research carried out in the remainder of this thesis, particulary in chapter 5.

However, since a complete overview of the vast number of existing shape analysis methods would be impractical, this chapter focuses only on a small selection of the most commonly used and/or those particulary relevant to the research carried out by the author in chapters 5, 6 and 7. A much larger selection of methods has been described by Loncaric in [147], where he identified over thirty shape description and matching methods covering a broad diversity of methods and providing a large set of useful references. State of the art in shape matching research was also discussed by Veltkamp [190, 171]. A very comprehensive discussion on various issues related to shape analysis can be found in [163]. A description of methods adopted by the MPEG-7 standard, the selection process (including description of various methods considered during the standardization) and an interesting set of applications of contour analysis can be found in [17].

Shape analysis and shape matching in particular is an important research subject in the area of CBIR. However, despite three decades of intensive research there are many unsolved aspects of similarity estimation between shapes. Many of the shape matching approaches proposed in the literature are not suited for use in CBIR systems due to factors like high computational cost of similarity estimation, lack of robustness to various deformations, and most importantly inability to provide similarity estimation aligned with human judgment of similarity in a given context. Methods where shape information is embedded into a low dimensional vector space (feature vectors) typically lack robustness against non-rigid deformations. Early dense matching techniques suffer from the inability to utilize multi-scale information and, due to the recent popularity of syntactical approaches, they were never tested with large test collections to fully asses their usefulness/potential in CBIR. The most recent syntactical approaches, although motivated by studies of human perception and dealing with compact structural information, led to utilization of complex matching algorithms which are often difficult to implement and also to ad-hoc solutions for solving ambiguities between compared shapes resulting in similarities poorly aligned with human perception of similarity. The above situation provided motivation for the research reported in the next chapter.

CHAPTER 5

MULTI-SCALE REPRESENTATION AND MATCHING OF NON-RIGID SHAPES WITH A SINGLE CLOSED CONTOUR

THIS chapter presents a new method for efficiently matching non-rigid shapes represented with a single closed contour and for subsequently quantifying dissimilarity between them in a way aligned with human intuition. The approach is primarily intended for indexing and retrieval of objects in large collections based on their 2D shape.

In this new approach, termed the *Multi-Scale Convexity Concavity* (MCC) representation [211], contour convexities and concavities at different scale levels are represented using a two-dimensional matrix. The representation can be visualized as a 2D surface, where "hills" and "valleys" represent contour convexities and concavities that are present at different scales. It is shown that such a rich representation facilitates reliable matching between two shapes and subsequent quantification of differences between them. The optimal correspondence between two shape representations is achieved using dynamic programming and a dissimilarity measure is defined based on this matching. The algorithm is very effective and invariant to several kinds of transformations including some articulations and modest occlusions. The retrieval performance of the approach is illustrated using several collections including the MPEG-7 shape dataset, which is one of the most complete shape databases currently available. Extensive experiments demonstrate that the proposed method is well suited for object indexing and retrieval in large databases. Furthermore, the representation can be used as a starting point to obtain more compact descriptors.

Based on this assumption, the second part of this chapter discusses several improvements of the initial approach. An alternative and more compact form of the MCC representation [211] is proposed, along with an optimization framework designed to further improve matching performance and facilitate the comparison of different versions of the approach by ensuring optimal conditions, and thereby a "level playing field", for the methods being compared.

Attractive properties of the new approach are that it outperforms existing methods in several applications, is robust to various transformations and deformations, has low matching complexity (compared to fine contour matching techniques [170, 167]) and that is relatively easy to implement.

5.1 Introduction

5.1.1 Motivation

The crucial problem in contour matching is the choice of features used to quantify differences between contour segments. Biological research has shown that contour curvature is one important clue exploited by the human visual system in this regard [212]. However, curvature is a local measure and is therefore sensitive to noise and local deformations. Estimating curvature involves calculating contour derivatives, which is difficult for a contour represented in digital form. In fact, the important structures in a shape generally occur at different spatial scales. Therefore, the most powerful shape analysis techniques are those based on multi-scale representations [213, 206, 209, 163]. Whilst multi-scale representations are very useful in analyzing contours, they are also highly redundant – this is the price to be paid in order to make explicit important structural information. Approaches to removing thus redundancy are again motivated by studies of human perception. Many researchers recognized the importance of transient events in human visual perception [212, 205]. In [212] it was demonstrated that corners and high curvature points concentrate more information than straight lines or arcs. Numerous successful shape analysis techniques were motivated by these observations. Various multi-scale contour representations proved to be very useful for detecting such transient events [205, 189, 206, 163, 209].

Methods based on transient events are cognitively motivated, compact and allow fast matching. The main drawback is that they are generally highly ambiguous. As discussed in the previous chapter, *Curvature Scale–Space* (CSS) [206] method (arguably the most popular approach), cannot distinguish between very deep and sharp concavities and very long shallow concavities. In fact, multi-scale structure analysis or event detection introduces an additional processing layer to obtain more meaningful or more compact representation but this additionally complicates the matching process.

Given these, and other observations in chapter 4, one could conclude that a simpler and more robust technique for matching and quantification of similarity between shapes could be obtained by directly using a rich (dense) shape representation extracted by recording certain shape properties computed during a contour evolution. This idea is evaluated in this chapter by introducing a rich (dense) shape description method termed *Multi-scale Convexity Concavity* (MCC) representation, where for each contour point, information about curvature (the amount of convexity/concavity) at different scale levels is represented using a two-dimensional matrix. A *Dynamic Time Warping* (DTW) technique is used to find an optimal refined alignment along the contours upon which the dissimilarity measure is defined. Attractive properties of the new approach are that it outperforms existing methods, is robust to various transformations and deformations, has low matching complexity (comparing to other fine contour matching techniques [170, 167]) and is easy to implement. It should also be noted that MCC extraction has very low computational cost compared to other similar approaches [206, 163] and is therefore more suitable for on-line contour-based recognition, where usually storage requirements do not play the key role.

The proposed multi-scale approach is an alternative solution to the approaches attempting to match events occurring in the scale space. The main motivation for this work was provided by the drawbacks of the scale space technique (CSS) for plane curves proposed by Mokhtarian and Mackworth in [13]. The innovation of the proposed approach is that the multi-scale representation is not used to detect important transient events to obtain a compact shape descriptor, but the amount of convexity/concavity is used directly for quantifying the differences between shapes. It is demonstrated throughout the chapter that using such a rich representation for quantifying differences between shapes can lead to significant improvement in retrieval effectiveness. However the author does not mean to imply that the transient events do not contain significant information about the structure of the contour, but rather that using a rich representation can ensure that none of the important information has been unnecessarily discarded allowing simpler and more robust matching general to all types of closed contours.

Although such a rich descriptor provides excellent retrieval performance, the inherent redundancy between scale levels implies high storage requirements. Therefore, an alternative representation, termed MCC-DCT, that reduces this redundancy by de-correlating the information from different scale levels and allows selective discarding of unimportant information is proposed. Also, the study of the influence of weights with which information from different scale levels is incorporated into the dissimilarity measure on the overall retrieval performance led to the development of an optimization framework to determine optimal proportions between information represented by different contour point features according to a chosen criterion. This framework is presented

in the context of optimizing the new shape representation so that it can be compared both to non-optimized versions of the approach as well as the original MCC approach that, in itself, is difficult to optimize. It is demonstrated that the average retrieval performance can be further improved by using such a framework. Furthermore, this framework could be used to tailor the shape representation to a specific application.

5.1.2 Chapter Structure

The remainder of this chapter is organized as follows: The next section describes the multi-scale shape representation in detail and discusses its properties. Section 5.3 presents an optimal solution for matching two MCC representations and provides a definition of dissimilarity measure. Section 5.4 presents retrieval results, including a comparison with the CSS approach in the author's own simulation of the "CE-Shape" MPEG-7 core experiment. The computational complexity of the proposed approach is discussed in section 5.5. An alternative version of the shape representation and one possible parameter optimization approach is described in section 5.6. Section 5.7 discusses the relationship between the proposed approach and other curve matching techniques followed by a discussion of prospects for further improvements and an outline of possible directions for future research in this area in section 5.8. Finally, conclusions are formulated in section 5.9.

5.2 Multi-scale Convexity Concavity Representation

5.2.1 Definition

Since usage of compact shape descriptors leads to problems with generalization and a trade-off between robustness to deformations and lack of discriminatory power, a new rich multi-scale shape representation termed *Multi-scale Convexity Concavity* (MCC) representation which stores information about the amount of convexity/concavity at different scale levels for each contour point is proposed. The representation takes the form of a 2D matrix where the columns correspond to contour points (contour parameter u) and the rows correspond to the different scale levels σ. Position (u, σ) contains information about the degree of convexity or concavity for the u^{th} contour point at scale level σ.

The simplified boundary contours at different scale levels can be obtained via a curve evolution process similar to that used in [13] to extract CSS images. Assuming a size normalized closed contour C represented by N contour points and parameterized by arc-length u: $C(u) = \big(x(u), y(u)\big)$, where $u \in \langle 0, N \rangle$. The coordinate functions of C are convolved with a Gaussian kernel ϕ_σ of width $\sigma \in \{1, 2...\sigma_{max}\}$:

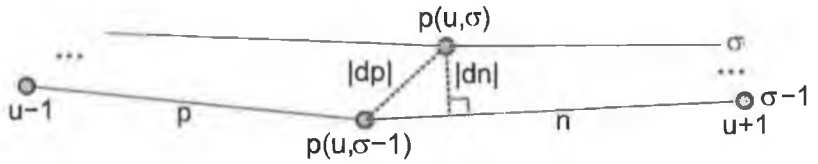


Figure 5.1: Contour displacement at two consecutive scale levels.

$$x_\sigma(u) = \int x(t)\phi_\sigma(u-t)dt, \tag{5.1}$$

$$\phi_\sigma(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2}{2\sigma^2}} \tag{5.2}$$

and similarly for $y(u)$.

The resulting contour, C_σ, becomes smoother with increasing values of σ, until finally the contour is convex. In the above definition σ is constrained to assume integer values only. Possible alternative quantization schemes for σ are briefly discussed in section 5.8.2

The convexity/concavity measure is defined as the displacement of the contour between two consecutive scale levels. If we denote the contour point u at scale level σ as $p(u,\sigma)$, the displacement of the contour between two consecutive scale levels $d(u,\sigma)$ at point $p(u,\sigma)$ can be defined as the Euclidian distance between position of $p(u,\sigma)$ and $p(u,\sigma-1)$:

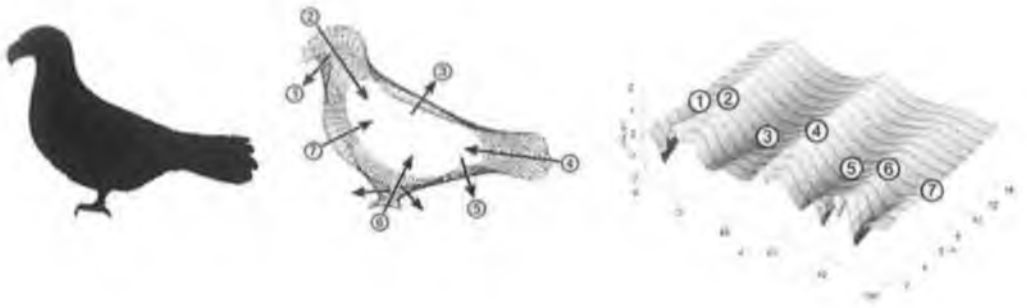
$$d(u,\sigma) = k\sqrt{\left(x_\sigma(u) - x_{\sigma-1}(u)\right)^2 + \left(y_\sigma(u) - y_{\sigma-1}(u)\right)^2} \tag{5.3}$$

where

$$k = \begin{cases} 1 & \text{if } p(u,\sigma) \text{ inside } C_{\sigma-1}, \\ -1 & \text{otherwise} \end{cases} \tag{5.4}$$

The sharper the convexity or concavity, the larger the change in the position of its contour points during filtering, where convex and concave parts are distinguished via a change in sign. The distance $d(u,\sigma)$ has positive (negative) values for u at convex (concave) parts of the contour. If $p(u,\sigma)$ lies inside the contour produced at level $\sigma-1$, this indicates that the contour at $p(u,\sigma)$ is convex ($d(u,\sigma)$ has positive value). If $p(u,\sigma)$ lies outside the contour produced at level $\sigma-1$ this indicates that the contour at $p(u,\sigma)$ is concave ($d(u,\sigma)$ has negative value).

For a small class of shapes, direct implementation of equation 5.3 can result in a loss of continuity for values of $d(u,\sigma)$ at the borders between convexities and concavities. At high scale levels, where the filtered contours are very smooth, some contour points move in a direction roughly parallel to the contour rather than in a direction perpendicular to the contour. In this case, the



(a) (b) (c)

Figure 5.2: Extraction of MCC representation: (a) - original shape *bird19*; (b) - filtering the original contour with different values of σ, arrows indicate direction of displacement; (c) - MCC representation (100 contour points at 14 scale levels);

distance between successive positions of the contour point does not reflect the displacement of the contours between two scale levels - an example is shown in Figure 5.1. This problem can be avoided by taking the displacement measure as the local minimum distance between contour point $p(u, \sigma)$ and neighboring segments of this contour point at scale level $\sigma - 1$. Theoretically, the search for the minimum distance should continue until the first local minimum distance is found. In a practical implementation, it is often sufficient to take the displacement measure as the minimum of two distances $|d_p|$ and $|d_n|$, where $|d_p|$ and $|d_n|$ denote the Euclidian distances between $p(u, \sigma)$ and segments $< u - 1, u >$ and $< u, u + 1 >$ at scale level $\sigma - 1$.

An example of the MCC extraction process is illustrated in Figure 5.2. All concavities and convexities are present at the lowest scale levels. Moving to higher scales the level of detail decreases and only the most significant convexities and concavities are retained. The author's observations indicate that the majority of contours become completely indistinguishable above the 10^{th} scale level. As a result, in all experiments shape representations with scale levels from 1-10 is used. Different schemes for adjusting the range of σ are discussed in section 5.8.2.

Note that the curvature measure could be employed as an alternative convexity/concavity measure as widely used in many other shape analysis methods [13]. However, whilst our experiments show that the use of curvature results in a descriptor with strong discriminatory capabilities, sometimes the matching process can be dominated by large values of curvature extremes. In other words, the above convexity/concavity measure was utilized due to its low computational cost and the ease with which it could be used within the proposed matching algorithm resulting in good alignment with human notion of similarity.

Although the above evolution process and measure of convexity/concavity appear intuitive, their employment is motivated by ease of implementation, even though they do not have a direct counterpart in any physical model. Many alternative, and often better motivated, approaches to measuring convexity/concavity

and to contour evolution could be employed. However, this chapter focuses on demonstrating the advantages of rich multi-scale representations over compact descriptors rather than investigating every possible variation of the contour evolution process and measure of convexity/concavity. Also a methodology for meaningful comparison of different versions of approaches based on such representations is investigated which provides the basis for investigating properties of alternative approaches to extraction of the rich multi-scale representation in the future. The most promising alternative approaches to contour evolution and measuring of convexity/concavity are briefly discussed in section 5.8.1.

The above extraction process implies data redundancy in the representation, since a 1D signal (measure of convexity/concavity along the contour) is represented by a 2D matrix. In [163] this phenomenon is described as unfolding a signal by a 2D transform, for the purpose of making explicit underlying structural information at different scales. Such approaches are commonly used for shape analysis [213, 206, 163], e.g. detection of perceptually significant points across scales. The following sections demonstrate that directly using such a rich representation for quantifying differences between shapes can be very convenient and can lead to significant improvement in retrieval effectiveness across many applications (generality) by providing good balance between robustness and discriminatory power.

5.2.2 Properties of MCC representation

Invariance to Linear Transformations

The MCC representation is invariant to translation (the convexity/concavity measure for a given contour point is calculated solely with the respect to its neighbors) and scaling (due to size normalization performed prior to contour filtering). Both, a rotation of the object and a change in the contours' starting point cause a circular shift of the representation along the u axis which is addressed during the matching process – see section 5.3.

Non-rigid Deformations (Nonlinear Transformations)

Generally it is not difficult to match and quantify similarity between very similar shapes. It is much more challenging to robustly measure similarity in the presence of small geometrical distortions, non-rigid deformations, occlusion and outliers. Also robustness to deformations has to be always considered together with discriminatory power. Moreover, it is important to test the balance between robustness to deformations and discriminatory power on sufficiently large collections. The author's experience indicate that evaluating the discriminatory power on datasets containing less than 200 shapes (which is common practice in

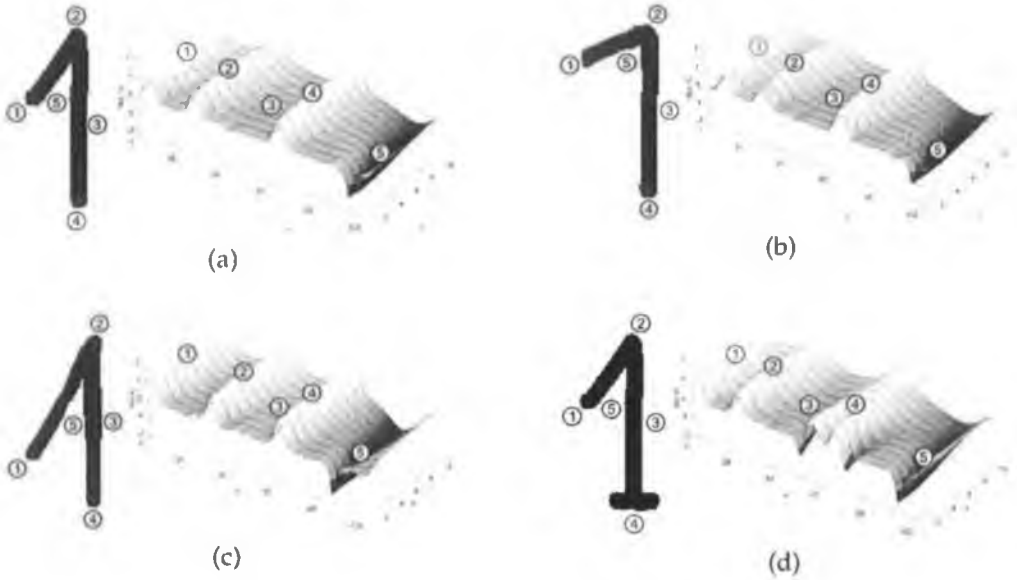


Figure 5.3: Influence of non-rigid deformations on MCC representation: (a) - Original shape, (b) - Top part moved up, (c) - Top part elongated, (d) - Shape with "outliers".

the literature) can be very misleading. Also, it should be noted, that the perfect trade-off between these two properties can differ from application to application.

In the case of the MCC representation, non-rigid deformations might shift the position of contour point features along the u axis, but do not significantly change the values of convexity/concavity. All such differences in the position of the features along the contour can be compensated during the matching process, but in contrast to other approaches [194], are not used as a penalizing factor in the final similarity measure. Only values of convexity/concavity are used for quantifying similarity between shapes in a way which conforms with human intuition.

The influence of shape deformations on the proposed shape representation is illustrated in Figure 5.3 using four different instances of the digit one (shapes in Figure 5.3b, 5.3c and 5.3d are manually modified versions of the shape in Figure 5.3a). For clarity, the most significant parts of the shape have been labelled: three deepest convexities as $\{1\}$, $\{2\}$ and $\{4\}$, one deep concavity as $\{5\}$ and a straight segment as $\{3\}$. The convexities correspond to three maxima in the MCC representation. Similarly, the concavity is represented by a single deep minimum. Shallow plateaus correspond to straight segments between convex parts. Modifying the top of the digit (Figure 5.3b) makes the concavity $\{5\}$ shallower resulting in a slightly shallower minimum in the MCC representation, especially at higher scales. The shape in Figure 5.3c has elongated the top part compared to the original version. In the MCC representation, this results in a larger separation between maxima/minima at all scale levels and the minimum becomes deeper, especially at higher scales. Introducing small outliers, as in

Figure 5.3d, introduces new maxima/minima in the MCC representations at lower scale levels. Similarly, contour noise (not shown) introduces additional convexities and concavities at lower scale levels, but higher scale levels usually remain unaffected due to the implicit filtering process. In all cases, there are small differences in the position of the maxima/minima along the contour – a fact which must be addressed during the shape matching process.

The reader may also notice that digit one can be also represented by a straight vertical line which humans can effortlessly recognize especially when occurring in the neighborhood of other digits. This should serve to remind us about the inherent limitation of any recognition technique based on matching which is the assumption that objects share common appearance and also about the fact that semantic interpretation of shapes often depends on context. Therefore, in many applications, successful recognition may require matching of an unknown shape with multiple shape instances (templates) of the searched object.

Further examples of shapes and their MCC representation are shown in Figure 5.4. Of particular interest are differences between the MCC representations of rigid shapes like *circle*, *square* and *triangle*. It should be noted that the CSS representation of these totally convex shapes would contain no maxima requiring utilization of additional features, e.g. the CSS-based descriptor adopted by the MPEG-7 standard solves this limitation by including global features. Figure 5.4 also illustrates MCC representations of shapes which could be considered elastically bent versions of each other, like *pen* and *horseshoe*.

It should be clear from the above examples that the matching procedure has to take into account differences in the position of the features along the contour in order to be robust to elastic deformations. Discrepancies between convexity/concavity measures of corresponding parts should allow dissimilarities between shapes to be quantified in a way which conforms to human intuition. While there is no theoretical justification for this, this assumption is evaluated experimentally in sections 5.4 and 5.6.5.

5.3 Matching and Dissimilarity Measure

5.3.1 Optimal Correspondence Between Contour Points

In the proposed approach the similarity/dissimilarity measure between two shapes is based on local distances between corresponding points (computed using their multi-scale feature vectors). The correspondences between points are established by finding the globally optimal match between two MCC representations. In other words, the optimal correspondence results in the lowest possible global cost (accumulated from local distances between these corresponding points) and at the same time fulfills a chosen global constraint,

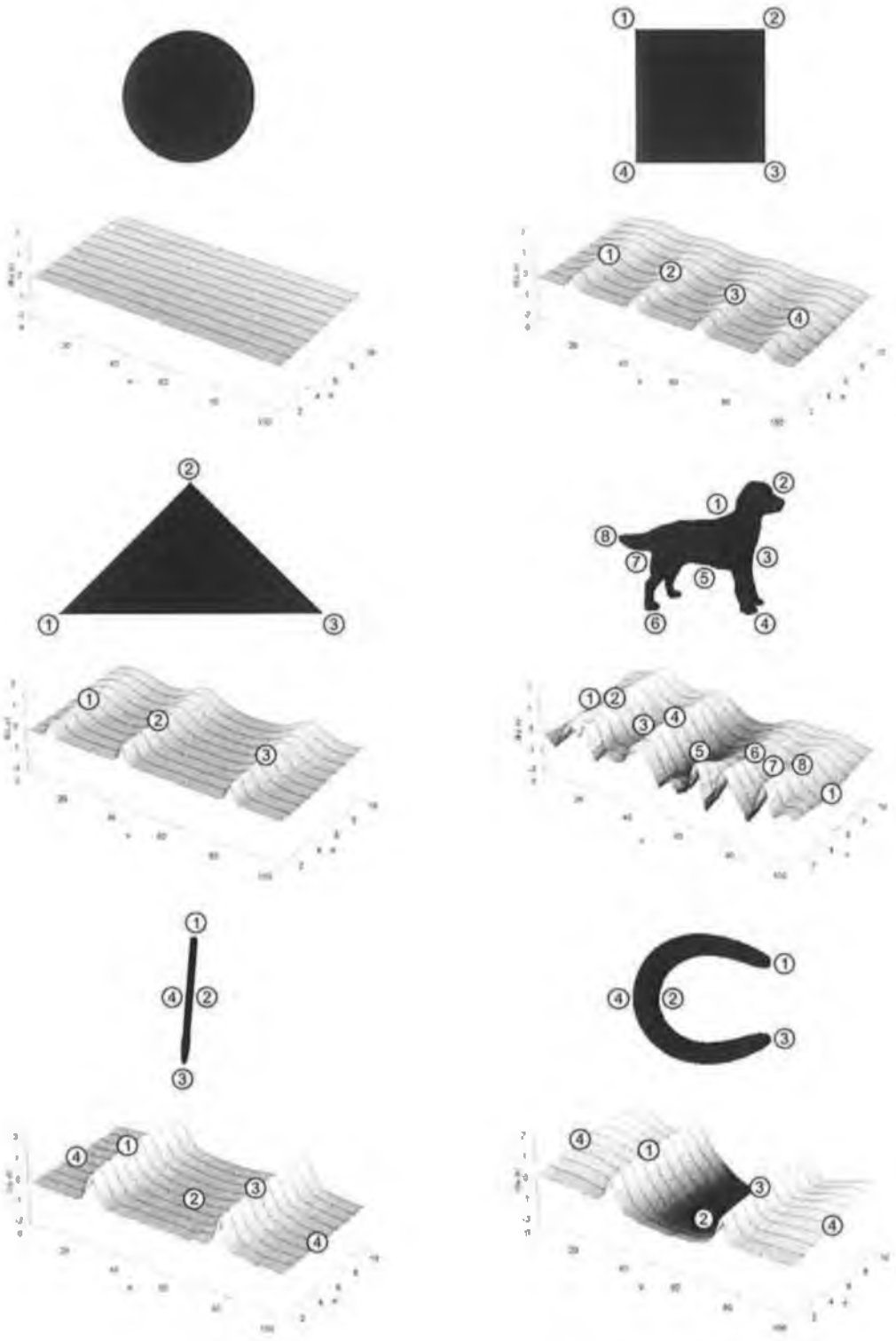


Figure 5.4: Examples of MCC representation.

e.g. maintains the sequential nature of contours or satisfies the condition that all points from one contour need to have at least one corresponding point from the other.

The matching process must determine the optimal circular shift between two representations and take into account small variations in the relative positions of feature vectors along the contour caused by deformations. Occlusions are not detected explicitly, but penalized implicitly by the cost of matching them to the non-occluded contour, i.e. to the part resulting in the best global correspondence. This assumption is motivated by the observation that occlusions or outliers affect the MCC representation only locally, making global alignment still possible.

This simplification allows utilization of a very straightforward *Dynamic Programming* technique for matching similar to one used for speech recognition where it is referred to as *Dynamic Time Warping* (DTW) [214]. DTW has been widely used for speech recognition where several utterances of the same word are likely to have different durations and parts being spoken at different rates. As such, a time alignment is performed to obtain a global distance between two speech patterns. In the case considered here, the task is to find an optimal global alignment along the contour.

When establishing correspondences between two contours, first distances between their contour points are computed using their multi-scale feature vectors and stored in an $N \times N$ distance table allowing their convenient examination. The columns of the table represent contour points of one shape and the rows represent the contour points of the other. Each entry in the table stores a distance between a pair of contour points corresponding to the row and the column of this entry.

Finding the optimal match between two contours corresponds to finding the lowest cost diagonal path through the distance table, i.e. traversing through the distance table from bottom left to top right. If one of the contours is rotated with respect to another or different starting points are chosen for the two contours, the optimal path will be shifted in the distance table. Thus, invariance to rotation and choice of starting point can be obtained by searching for the circular shift resulting in the least cost path. However, in the case of non-zero shift (starting points do not correspond), the optimal path may need to wrap around from the last column to the first column of the distance table. This may be efficiently implemented by repeating the columns of the distance table, resulting in an extended distance table of size $2N \times N$ [204] – see Figure 5.5. If one of the contours is a mirrored version of the other, the optimal path will become perpendicular to the path which would be optimal for a non-mirrored version. Thus, invariance to mirror transformations can be obtained by searching for the optimal path in both directions, i.e. traversing through the distance table from bottom left to top right or from bottom right to top left. The final dissimilarity

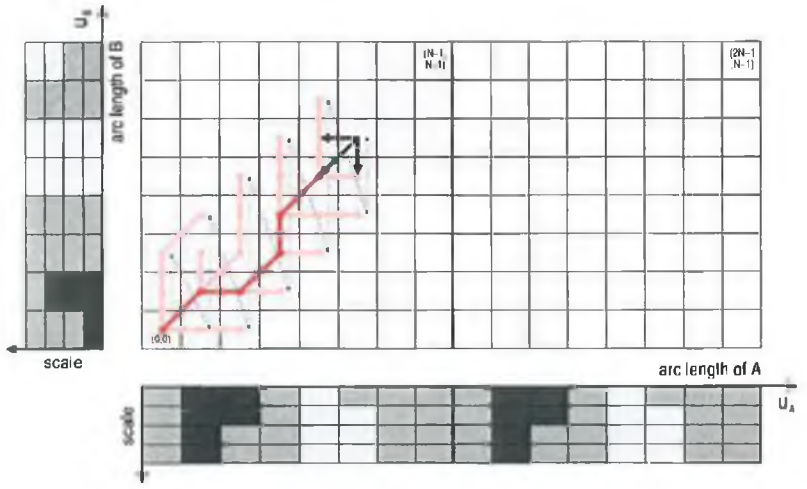


Figure 5.5: Matching two MCC representations using dynamic programming.

between shapes is calculated as the cumulative cost along the least cost path.

5.3.2 Distances Between Contour Points

This section describes in detail the computation of pairwise distances between points of both contours. Define the MCC representation as an $\sigma_{\max} \times N$ matrix $\mathbf{F} = [f_{\sigma u}]$ where $f_{\sigma u}$ denotes the convexity/concavity measure for contour point u at scale σ. Let contours A and B be represented by $\mathbf{F}^A$ and $\mathbf{F}^B$ respectively. The distance between two points u^A and u^B from A and B respectively is defined as the average of absolute differences between their convexity/concavity measure from all scales:

$$d(u^A, u^B) = \frac{1}{\sigma_{\max}} \sum_{\sigma=1}^{\sigma_{\max}} |f^A_{\sigma u^A} - f^B_{\sigma u^B}| \tag{5.5}$$

It should be noted that according to the above definition, all scale levels are incorporated in the distance measure with the same importance (weight). It is possible that using different weights could result in more meaningful matching and dissimilarity measure, leading consequently to improved overall retrieval accuracy. The influence of the weights on retrieval performance is investigated in section 5.6.

5.3.3 Dynamic Programming Formulation

In a dynamic programming formulation, the sequential nature of the contours is maintained, as the path is traced through the distance table, by imposing the condition that paths cannot move backwards along the columns. The path may only proceed upwards and to the right in the table – see Figure 5.5. An additional condition is imposed that every contour point (represented by a column in the input MCC representations) must be used in calculating the matching path. A

single point from one contour being matched to a large part from the other contour (i.e. long horizontal or vertical segments in the least cost path) is avoided. This can be easily achieved by constraining the path geometry by imposing the condition that a point from one contour can only be assigned to a maximum of two points from the other contour.

Furthermore, as is common practise in DTW algorithms [214], the least cost path is restricted to lie in the area close to the ideal straight diagonal line in order to speed up the matching process. The path is allowed to deviate from a straight diagonal path by no more than a chosen deviation threshold of $\pm T_d$ entries. In initial experiments T_d will be set to $0.04N$, which in most cases has no affect on the geometry of the optimal path representing a good trade off between speed and matching performance.

To formally define the dynamic programming formulation, let us initially assume that the starting points in both contours are known and shifted to position 0 along the contours. Therefore the distance between starting points is stored in entry $(0,0)$ of the distance table. The matching algorithm starts at cell $(0,0)$ and proceeds upwards (from the beginning to the end of each column) and to the right (from the first to the last column) through the distance table updating costs of all possible allowable paths. The implementation of such a process requires additional structures to store analyzed paths and costs accumulated along each path. This can be achieved by extending the distance table to a structure commonly referred as a *Dynamic Programming* (DP) table where each entry stores not only the distance between corresponding pair of contour points but also the cost of the least cost path between the entry and entry $(0,0)$ and also the best predecessor in the path. The least cost path through the distance table, corresponding to the best matching between two representations is the total cost accumulated at the top right entry $D(N-1, N-1)$. Contour point correspondence can be achieved by tracing all predecessors of entry $D(N-1, N-1)$ in the least cost path.

The detailed description of the mechanism used for updating (accumulating) the cost along each path is as follows. Let $D_{\mathrm{TOT}}(u^A, u^B)$ to be the total cost accumulated from distances stored along the least cost path between entry $(0,0)$ and (u^A, u^B). The value of D_{TOT} at entry (u^A, u^B) can be evaluated as the sum of the cost of matching contour points corresponding to this entry and the least accumulated cost of one of three (on the left, bottom-left, and bottom) of its predecessors:

$$D_{\mathrm{TOT}}(u^A, u^B) = d(u^A, u^B) + \min \begin{cases} D_{\mathrm{TOT}}(u^A - 1, u^B \quad) \\ D_{\mathrm{TOT}}(u^A - 1, u^B - 1) \\ D_{\mathrm{TOT}}(u^A \quad, u^B - 1) \end{cases} \tag{5.6}$$

where $d(u^A, u^B)$ is the distance between contour points u^A and u^B computed

using the equation 5.5 and $D_{\mathrm{TOT}}(u^A - 1, u^B)$, $D_{\mathrm{TOT}}(u^A - 1, u^B - 1)$ and $D_{\mathrm{TOT}}(u^A, u^B - 1)$ are total distances accumulated along least cost paths between entry $(0, 0)$ and left, bottom-left, and bottom predecessors respectively.

The constraint that a contour point from one contour can only be assigned to a maximum of two contour points from the other contour, means that entry on the left $((u^A - 1, u^B))$ and at the bottom $((u^A, u^B - 1))$ of the current entry can become its predecessors only if their own least cost predecessors were not found to be the bottom $(u^A - 2, u^B)$ and the left $(u^A, u^B - 2)$ entries respectively. This means that the total minimum cost accumulated for the entries on the bottom or to the left of the current entry (u^A, u^B) is defined effectively as:

$$D^{eff}_{TOT}(u^A - 1, u^B) = \begin{cases} D_{TOT}(u_A - 1, u_B) & \text{if } PR(u_A - 1, u_B) \neq (u_A - 2, u_B) \\ \infty & \text{otherwise} \end{cases} \tag{5.7}$$

and

$$D^{eff}_{TOT}(u^A, u^B - 1) = \begin{cases} D_{TOT}(u_A, u_B - 1) & \text{if } PR(u_A, u_B - 1) \neq (u_A, u_B - 2) \\ \infty & \text{otherwise} \end{cases} \tag{5.8}$$

where $PR(u_A - 1, u_B)$ and $PR(u_A, u_B - 1)$ denote the lowest cost predecessors for entries $(u_A - 1, u_B)$ and $(u_A, u_B - 1)$ respectively. Note, that implementation of the last equation requires storing the position of the least cost predecessor for each entry of the DP table even if the objective is solely estimation of dissimilarity between contours without the best correspondences between points.

To summarize, the matching algorithm starts at cell $(0, 0)$ and proceeds upwards and to the right through the DP table and the costs of all possible allowable paths are updated according to equations 5.6, 5.7, and 5.8. The least cost path passing through the distance table, corresponding to the best match between two shape representations, is the total cost accumulated at the top right entry $(D(N - 1, N - 1))$.

5.3.4 The Complete Matching Algorithm

As already explained, both, a rotation of the object and a change in the contours' starting point cause a circular shift of the representation along the u axis. Since the starting points and rotation between contours are in fact generally unknown, it is therefore necessary to perform a search for the circular shift between two MCC representations resulting in the least cost path. Although various methods for fast estimation of the best pair of starting points could be employed, see for example [170], in this thesis a simple exhaustive search is described which is used in several experiments for providing benchmark results.

The exhaustive search algorithm assumes an arbitrary point from one contour as a starting point and investigates all possible points from the other as potential

matches to the selected point. An arbitrary point is selected as a starting point in one contour (to avoid additional shifting it is practical to pick the original starting point) and each point from the other contour is evaluated as the best pairing with this starting point by finding the least cost path through the distance table commencing from the entry corresponding to this pair. This means that the algorithm for finding the least cost path must be repeated N times.

Also, whenever the optimal path starts from a cell other than $(0,0)$ it needs to wrap around from the last column to the first column of the DP table. This is efficiently implemented by repeating the columns of the DP table, resulting in an extended table of size $2N \times N$ [204]. The DTW algorithm can then be run iteratively for $i = 0, 1, ...N$, each time determining the optimal path between entries $(i, 0)$ (bottom row) and $(i+N-1, N-1)$(top row). Invariance to a mirror transformation can be obtained by flipping the columns of the distance table and repeating the search for the optimal path.

The entire algorithm for determining the least cost path through the distance table can be stated as follows[1]:

INPUT: MCC representations of shapes A and B;
OUTPUT: The least cost D_{min} of matching A and B;

1. **SET** $D_{min} = \infty$
2. **FILL** NxN *Distance Table* using Eq. 5.5;
3. **EXTEND** the *Distance Table* to 2NxN by repeating columns;
4. **FOR** $i = 0, 1, ...N$ **DO** *//Search all shifts for the least cost path*
   - **FIND** the lowest cost path between entries $(i, 0)$ and $(i + N - 1, N - 1)$ proceeding upwards and to the right through the *Distance Table* and summing local distances using Eq. 5.6;
   - *//Check if the the new path has lower cost than* D_{min}
   - **IF** $D_{\mathrm{TOT}}(i + N - 1, N - 1) < D_{min}$ **THEN** $D_{min} \rightarrow D_{\mathrm{TOT}}(i + N - 1, N - 1)$; **END IF**;

   END FOR
5. **IF** checking of mirror transformation required **THEN**
   - **FLIP** the columns of the *Distance Table*;
   - **REPEAT** step 4;

   END IF;

The above algorithm finds the minimum cost D_{min} of traversing the distance table for N pairs of starting points (all possible circular shifts) and if necessary in a given application can take into account mirror transformation. It should be stressed that the most computationally expensive step, calculation of pairwise distances between contour points is performed only once and does not have to be repeated for evaluating a mirror transformation or a shift.

[1]For clarity, operations related to tracing of the optimal path whenever the objective is also establishing the best correspondence between contour points are omitted from the pseudo-code.

5.3.5 Shape Dissimilarity

The final dissimilarity between two contours is calculated by normalizing the cost of the optimal path (D_{min}) found by the matching algorithm by the number of contour points N used and a crude estimate of complexities of both shapes computed from their MCC representations:

$$D(A,B) = \frac{D_{min}}{N \cdot \left(x_A + x_B\right)} \tag{5.9}$$

Complexities x_A and x_B are estimated as the averages of dynamic ranges of concavity/convexity measures from all scales:

$$x_A = \frac{1}{\sigma_{\max}} \sum_{\sigma=1}^{\sigma_{\max}} |\max_{u^A}\{f^A_{\sigma u^A}\} - \min_{u^A}\{f^A_{\sigma u^A}\}| \tag{5.10}$$

and similarly for x_B.

The introduction of the second normalization is motivated by the author's observations that humans are generally more sensitive to contour deformations when the complexity of the contour is lower. Thus, dissimilarities between low complexity contours should be additionally penalized.

It should be stressed that there is no extra penalty for stretching and shrinking of contour parts during the matching process. Rather, a globally optimal match between two contours is found using a rich multi-scale feature for each contour point and dissimilarity is based solely on distances between corresponding points. This is contrary to the majority of other contour matching approaches based on dynamic programming where different penalty factors for stretching and shrinking, usually chosen in an ad-hoc fashion [170], drive the matching process.

5.3.6 Matching Examples

Examples of typical matching are shown in Figure 5.6. Figure 5.6a shows matching of two shapes from the MPEG-7 dataset: *stef09* and *stef13*. It can be seen, that despite significant non-rigid deformations between these shapes, the optimal path deviates only minimally from the straight diagonal line (roughly part {6} of the shapes), resulting in intuitive correspondences between contour points. This example confirms that the approach exhibits robustness to non-rigid deformations. Figure 5.6b shows an example of matching in the presence of outliers (or occlusions) where shape *horse14* from the MPEG-7 dataset is matched with its modified version representing the horse with a rider. The fragment of the optimal path corresponding to matching of unaltered parts of the modified shape to the original shape is practically straight indicating that

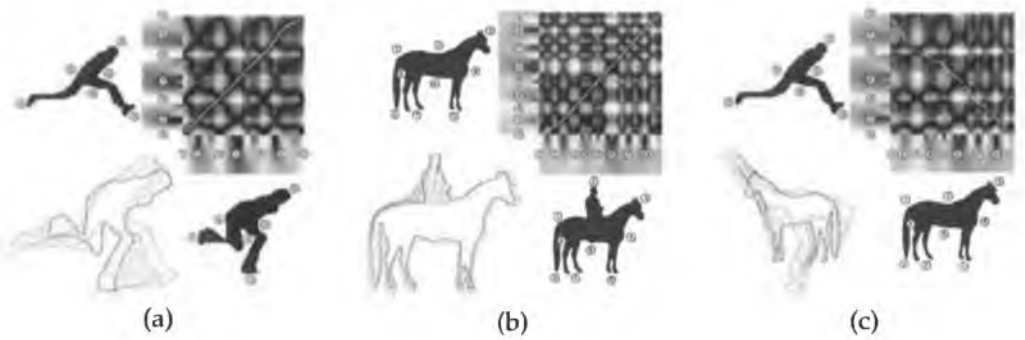


Figure 5.6: Matching examples: (a) - Matching of two shapes *stef09* and *stef13* from MPEG-7 collection, (b) - Matching of *horse14* from MPEG-7 collection and its manually modified version to illustrate matching in presence of outliers or occlusions, (c) - Matching of *stef09* with *horse14*.

the representation of these parts remained unaffected even in the presence of the significant outlier. The deviation of the path from the straight diagonal line (segments labelled as {1} and {2}) is related to matching of the rider from one contour to the horse's back from the other contour. The cost of matching the rider to the back of the horse is used to quantify the dissimilarities between two shapes. Matching two very dissimilar shapes is illustrated in Figure 5.6c. Despite significant differences between both shapes, the matching algorithm is able to find an intuitive match. The best match is found for a mirrored version of shape *horse14*. Of course, even such globally optimal matching should result in a large dissimilarity measure between these shapes.

The above examples illustrate the capabilities of the matching algorithm to compensate for small deviations in the position of the convexities/concavities along the contour and the fact that an intuitive correspondence can be found even in the presence of significant outliers.

5.4 Results

This section demonstrates the usefulness of the proposed method to shape-based retrieval in large datasets. All presented experiments are performed using MCC representations with 100 equally spaced contour points and scale levels from 1-10.

5.4.1 Retrieval of Marine Creatures

Illustrative retrieval results obtained using the proposed approach on a database of 1100 shapes of marine creatures [206], are illustrated in Figure 5.7. Although there is no ground-truth classification associated with the above collection it can be safely stated that the retrieval results correspond closely to human perception of similarity. Moreover the results compare favourably with results obtained by

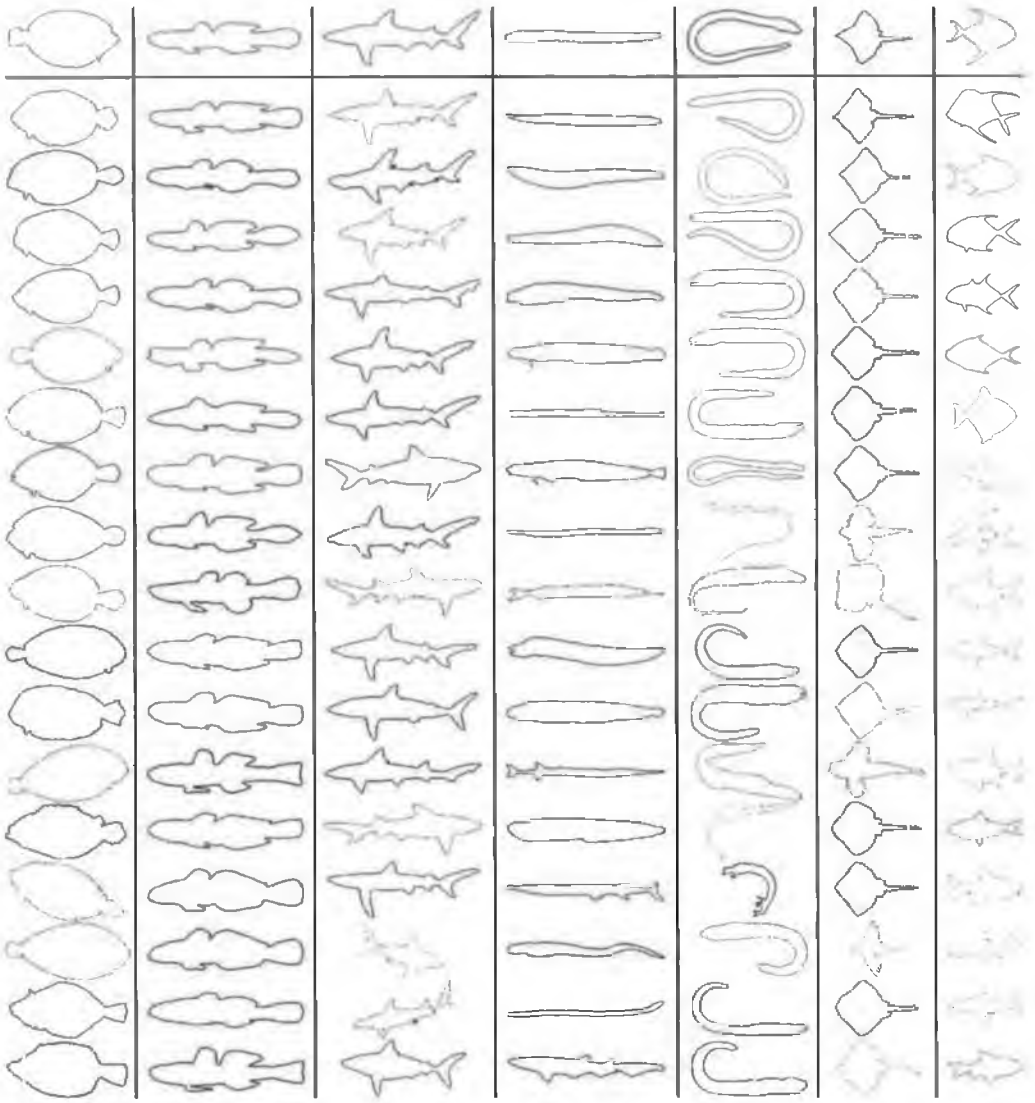


Figure 5.7: Illustrative retrieval results obtained for selected entries from a database of marine creatures. The top row shows query shapes, whilst subsequent rows show the first 17 top ranked matches.

the original CSS approach implemented as in [206][2].

5.4.2 Simulation of MPEG-7 Core Experiment

This section presents quantitative retrieval results obtained by the proposed approach in the author's simulation of the main part (part B) of the Core Experiment "CE-Shape-1" specified during the MPEG-7 standardization process [215].

In this experiment, originally performed as part of the MPEG-7 standardization process, a set of semantically classified images with an associated ground truth is used. The total number of images in the main MPEG-7 collection is 1400

[2]On-line system demonstrating retrieval of marine creatures using the CSS approach is available at `http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/demo.html`

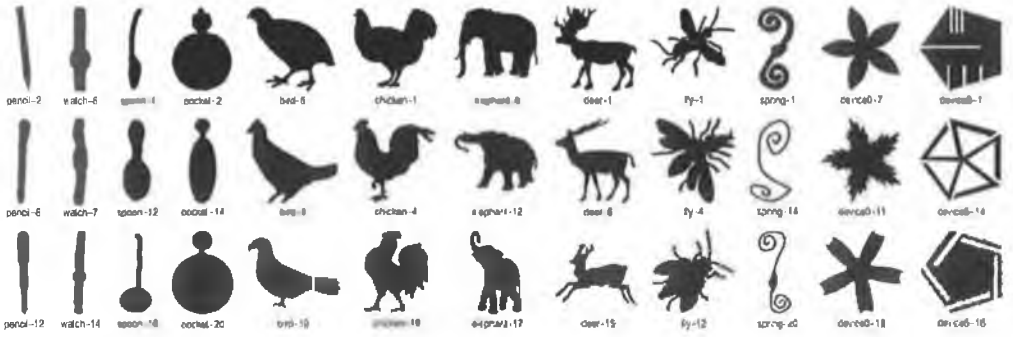


Figure 5.8: Examples of shapes used in part B of "CE-Shape-1" MPEG-7 core experiment.

consisting of 70 classes of various shapes, each class with 20 images. Each shape is used as a query, and the number of similar images (which belong to the same class) is counted in the top 40 matches ("Bulls-Eye" test). Examples of shapes used in this experiment are shown in Figure 5.8.

The collection is the largest and most complete collection of general shapes commonly used in the literature. It contains classes of shapes with different levels of intra class variability, non-rigid deformations, occlusions and cracks. The variety of shape classes and the size of the collection ensures rigorous evaluation of the generality of the approach. As outlined in [208], a 100% retrieval rate is not possible due to high intra class variability. Some classes contain objects whose shape is significantly different so that is not possible to group them into the same class using only shape features. [208] discusses an illustrative example whereby two spoons are more similar to shapes in different classes than to themselves. In fact some researchers predicted that the average retrieval rate of around 80% in MPEG-7 Core Experiment was already close to saturation [216].

One of the best overall performances (76.51%) obtained in the above test was reported in [217] for a method based on shape context descriptors attached to each contour. In the experiment performed as part of the MPEG-7 standardization process, the best performance of 76.45% was reported for the method based on the best possible correspondence of visual parts [192, 218]. A method based on CSS obtained a performance of 75.44%. Results obtained using a more recent optimized version of the CSS approach [216, 211] show a performance of 80.54%. In comparison, the total performance of the approach proposed in this chapter is 84.93%, which shows that the method outperforms the above techniques. It should be noted that this result was obtained for the non-optimized version of the matching algorithm and without adjusting any parameters [211].

Figure 5.9 shows detailed retrieval results for all classes of shapes from the MPEG-7 database for both approaches: MCC and CSS. The classes are sorted according to the retrieval rate obtained by MCC, with corresponding CSS results

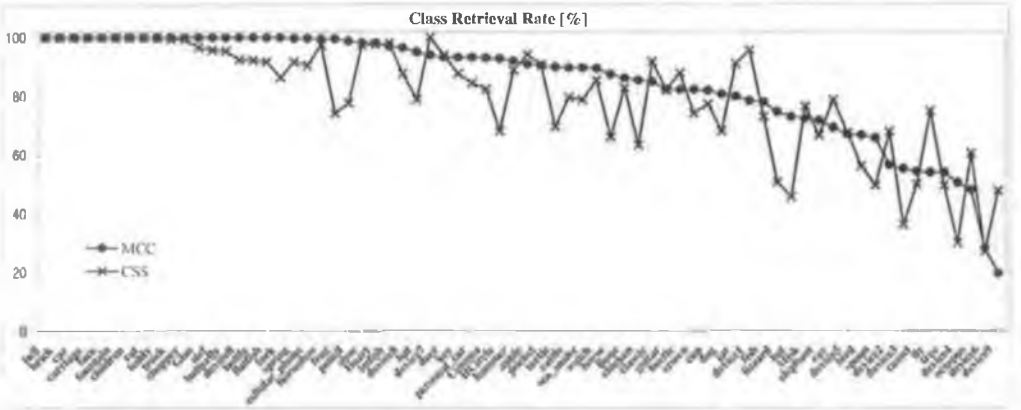


Figure 5.9: Retrieval rates for all classes from the MPEG-7 core experiment (part B of "CE-Shape-1") for both MCC and CSS (sorted by retrieval rate of MCC).

overlaid. The proposed approach outperformed the CSS-based method for the majority of classes. In particular, the proposed approach performs much better then CSS for all totally or almost totally convex classes present in the MPEG-7 database, such as 'pencil', 'HCircle', 'device3', 'device4'. This reflects the fact that CSS image does not have any zero crossings of curvature for totally convex shapes and the method uses simple global descriptors instead. Very good performance was obtained by the MCC-based approach for classes such as 'fork', 'spring', 'frog', 'pencil' and 'device8'. Retrieval results obtained for classes 'bat', 'cattle', 'horse' and 'lizard' are also encouraging, taking into account the high diversity within these classes. The above results illustrate that the proposed representation can capture important (from a human's point of view) shape features.

The poor results obtained for two classes 'device6' and 'device9', can be explained in both cases by the fact that they contain shapes representing geometrical figures (pentagons and circles respectively) with very deep cracks and holes - see Figure 5.8. These shapes are similar according to region-based similarity criteria, but are very different according to contour-based similarity criteria. Clearly, for shape-based retrieval of such classes, region-based descriptors rather than contour-based descriptors should be used.

To complete the comparison between the proposed and the CSS[3] methods Figure 5.10 shows their Precision-Recall curves which are a standard evaluation technique in the information retrieval community [220].

To the best of authors knowledge the quantitative results obtained by the proposed approach in both the bulls-eye test and in terms of precision-recall curves for the MPEG-7 collection are the best reported so far in the literature.

In order to directly illustrate the agreement between dissimilarities between shapes

[3]CSS extraction and matching was performed using the MPEG7 eXperimentation Model (XM software v5.6) [219].

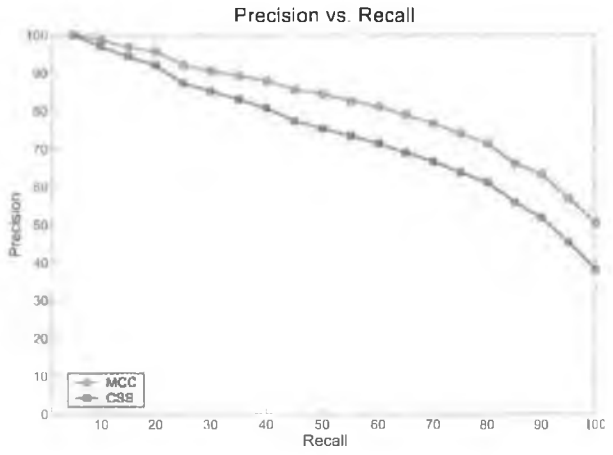


Figure 5.10: Precision versus Recall for the proposed and the CSS approach.

obtained by the proposed approach and human notion of dissimilarity distances between all shapes from the MPEG-7 collection are shown in Figure 5.11. It can be observed that distances between shapes belonging to the same class are smaller than inter-class distances for majority of shapes.

5.5 Computational Complexity

In this section the computational costs of extracting the MCC representation and the matching stage are analyzed independently, as they are usually performed separately.

Extraction of the MCC representation mainly involves iterative filtering of the original contour and calculating the distance between evolved contours. Therefore, the complexity of the extraction stage is $O(\sigma_{max}N^2)$, where σ_{max} denotes the number of scale levels and N is the number of contour points used in the MCC representation. In initial experiments ten scales are used as using higher scale resolution does not lead to improvement in recognition efficiency. For comparison, the CSS image is usually extracted (as specified in [15, 206]) using approximately 100 iterations, each involving filtering of the original contour three times (with gaussian kernel and its first and second derivatives). It took approximately 10 seconds to calculate MCC representations for the entire MPEG-7 database containing 1400 shape images running on a standard PC with Pentium IV/1600Mhz processor and using an implementation of the approach in C++ programming language. Due to very low computational cost of extraction compared to other approaches [206, 163] the proposed method is suitable for fast on-line contour-based recognition in small collections, where usually storage requirements do not play key role, e.g. fast recognition of handwritten words from a small dictionary.

On other hand, the presented matching algorithm has quite a high order of

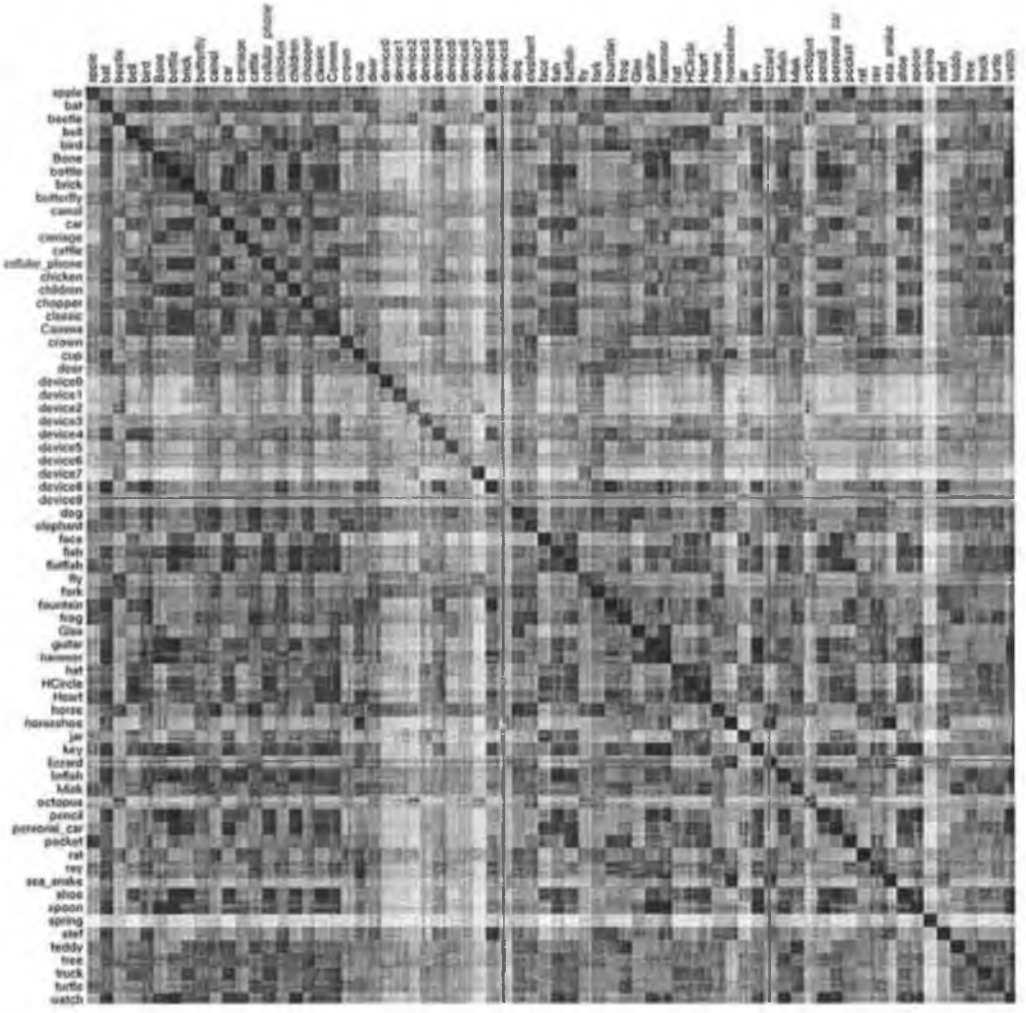


Figure 5.11: Distances between all shapes from the MPEG-7 collection computed using the proposed approach, i.e. Each position/point from the table corresponds to a distance between two shapes. For clarity only class names are shown.

complexity, as the main intention was to find the optimal match between two descriptors and show the maximum retrieval capabilities of the proposed shape representation. The original DTW algorithm requires creation of a $N \times N$ distance table (where N is the number of samples in the matched signals) and one scan through the table to find the least cost path which implies $O(N^2)$ complexity. Since in the case of the algorithm presented here, N circular shifts are checked to obtain invariance to rotation and choice of the starting points the complexity of the entire matching procedure is $O(N^3)$. However, the practical implementation is still quite fast due to the simplicity of elementary operations performed during the matching process. A single comparison between a pair of shapes, including mirrored matching, takes roughly 7ms on a standard PC with Pentium IV/1600Mhz processor. As such, querying the entire MPEG-7 database with a single shape requires less than 10 seconds.

It should be stressed that although the presented algorithm is quite efficient,

much faster matching approaches could be developed in the future. The presented algorithm was developed to find the global optimal match between two MCC representations of closed contours and is described here to illustrate the full capabilities of the proposed descriptor.

The complexity could be significantly reduced (perhaps close to linear) by using only sparse representations to roughly align the shapes before more detailed matching. For example utilization of maxima/minima at high scale levels in a voting scheme, similar to the approach proposed in [170], in order to estimate the most likely global alignments (circular shifts) provided very encouraging results (omitted in this thesis) and will be fully investigated by the author in the future. Also simple shape statistics calculated directly from the contour or from the proposed representation could speed up searching in large databases. One such approach, based on a priori knowledge about the application domain is tested in chapter 6 for holistic word matching. Ultimately, one could perform matching off-line to index the shape database prior to searching similarly to the approach proposed in [221].

5.6 Alternative Representation and Matching Optimization

The remainder of this chapter is devoted to better understanding the advantages and limitations of the proposed method. Possible modifications of the approach are discussed and selected variants are evaluated. In particular, possibilities for establishing optimal combinations of relative weights between distances computed using different scales promising further improvement of the retrieval performance and prospects of reducing storage requirements by decreasing redundancy of the representation are investigated.

5.6.1 An Optimization Framework to Facilitate Evaluation

In the case of shape matching techniques, parameter tuning is frequently performed in an ad-hoc manner [170] or based on trial-and-error. However, it should be remembered that in order to meaningfully evaluate effects of the proposed modifications, optimal testing conditions for different versions of each approach tested have to be ensured.

In the case of the proposed approach, the above requirement means ensuring optimal (according to a chosen criterion) combinations of relative proportions in which information from different contour point features should be combined in the final similarity measure. Until now, distances between contour points computed using different scales were combined with the same importance (see equation 5.5). Clearly, adapting the weights between distances computed using

different scales could be used to tune properties of the matching process. It should be stressed that the weights do not depend on matched contours and are constant for all contour points. Changing the weights should affect the values in each entry of the distance table and directly influence the final dissimilarity measure (calculated as the cumulative cost of those entries along the least cost path). Moreover, altering the values in the distance table might affect the least cost path and therefore the alignment (correspondences between points) of contours itself. In the remainder of this section the influence of the weights on the overall retrieval performance is investigated in a systematic way.

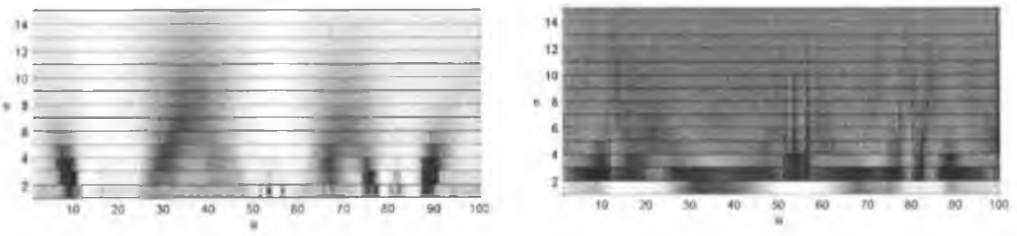
Due to high dimensionality of the solution space ($\sigma_{\max}$ dimensions) further development and testing of the approach requires a systematic framework for optimization of the weights. Such a framework is crucial to investigate the influence of different parameter settings on performance and facilitate the comparison of different versions of the approach by ensuring optimal conditions, and thereby a "level playing field", for all the methods being compared.

In order to optimize retrieval performance, a ground-truth database containing classified shapes is required. Clearly, one could argue that such datasets with ground-truth are often unavailable in real application scenarios. However, even trial-and-error tuning requires a certain number of examples and the optimization method presented here is simply an automatization of this process. The number and type of examples used determine the applicability (generalization) of the optimized method. From another point of view, optimization using a training set specific to a particular application could better tailor the method to this application.

Human judgements of shape similarity differ significantly between observers thereby making the evaluation task very difficult [169, 206]. The desired properties of a shape analysis technique and the associated evaluation criteria depend on a particular application. Adaptation of a shape representation method to a given application can be achieved by utilizing a training step in order to develop the optimal parameters for a given application. The alternative version of the MCC representation and optimization framework proposed in the following sections is one approach to performing this adaptation.

5.6.2 Reduction of Redundancy Between Contour Point Features

Although, a rich representation like MCC provides significant improvement in retrieval performance compared to compact descriptors like CSS, the inherent redundancy between scales implies high storage requirements. Furthermore, high correlation between scale levels make the MCC representation not suited to an optimization process. This section discusses an alternative representation,



(a) (b)

Figure 5.12: MCC-DCT representation: (a) - Original MCC representation of shape *bird19* from Figure 5.2, (b) - MCC-DCT.

termed MCC-DCT, that reduces the redundancy between scales.

MCC-DCT

In order to reduce correlation between multi-scale feature components of each contour point, a 1D DCT is applied to each column of the MCC array. In this new MCC-DCT representation, the feature vector of every contour point consists of the DCT coefficients of its original multi-scale feature vector (Figure 5.12b).

The DCT de-correlates information from different scale levels thus making it more suitable for the optimization process described in the next sections. Additionally, the DCT rearranges the significance of information across different scale levels placing most of the feature vector energy in lower levels. This property, can be utilized to obtain more compact shape representation when combined with a quantization scheme. For clarity, in the remainder of this thesis, DCT coefficients for contour points are indexed by letter i to distinguish them from scale levels σ.

MCC-DCT Matching

The main difference between the original MCC and the modified MCC-DCT representations is the form of the contour point feature vector. Therefore, the MCC-DCT representation can be matched using the same matching algorithm as described previously for the original MCC representation except that the distances between contour points are computed based on their DCT coefficients.

In order to control the relative proportions between distances computed using different coefficients the distances between contour points are computed using a slightly different formula than the one used for the MCC representations (see equation 5.5). Define the MCC-DCT representation as an $I \times N$ matrix $\mathbf{F} = [f_{iu}]$ where f_{iu} denotes the i^{th} DCT coefficient for contour point u. Let contours A and B be represented by $\mathbf{F}^A$ and $\mathbf{F}^B$ respectively. The distance between two contour

points is computed as:

$$d(u^A, u^B) = \sum_{i=1}^{I} p_i \cdot \frac{|f^A_{iu^A} - f^B_{iu^B}|}{x^A_i + x^B_i} \tag{5.11}$$

where x_i denotes dynamic range of i^{th} coefficient computed along the contour and defined as $x_i = \max_u\{f_{iu}\} - \min_u\{f_{iu}\}$.

Weights p_i control the relative proportions between distances computed using different coefficients. The final dissimilarity between two contours is calculated by normalizing the cost of the optimal path (D_{min}) found by the matching algorithm by the number of contour points used in the MCC-DCT representations: $D(A, B) = D_{min}/N$.

5.6.3 Optimization Methodology

The performance of the approach can be adjusted by tuning the weights p_i in equation 5.11. Changing the weights alters the values in the distance table entries and affects the final dissimilarity measure between the contours. The goal of the optimization is to find values of p_i which optimize the retrieval performance for the database.

Assume that a training collection and an optimization criterion (measure of performance) based on the ground-truth are available. The adjustment of the weights p_i can be tackled as the optimization of an objective function of $I - 1$ independent parameters p_i, where I denotes the number of contour point features. It should be noted that the number of free parameters is $I - 1$ and not I due to the fact that the performance depends on the relative proportions (not absolute values) between parameters p_i. In order to make the optimization process unambiguous, the value of one parameter, typically corresponding to the most significant feature, is set to one and the optimization framework is used to determine the relative values of the remaining $I - 1$ parameters.

To avoid re-matching of all shapes in the dataset for each evaluated combination of parameters p_i the optimization process is split into two stages. In the first stage, all shapes in the database are matched using the algorithm described in section 5.3 with an initial combination of parameters p_i and the resulting contour point correspondences are stored. In the second stage, the optimization of the chosen criterion is performed based on the contour point assignments obtained from the first stage. Once the new optimal values of parameters p_i are found, the alignment between contour points for each pair of shapes is updated. Therefore, the whole optimization process is repeated iteratively until convergence. The complete procedure is depicted in Figure 5.13.

The optimization in the second stage is performed by the commonly used *direct*

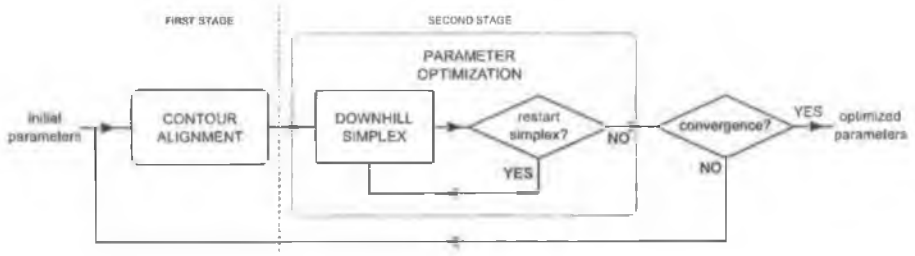


Figure 5.13: Optimization algorithm.

search[4] technique termed *Nelder–Mead Simplex Method* [223, 222], as derivatives of the objective function would be difficult to calculate. In all experiments the initial solution required by the Nelder–Mead simplex method is provided by empirically setting all weights p_i to one. The problem's characteristic length scales for each dimension, also required by the Nelder–Mead simplex method, is assumed to be identical for each dimension and set empirically to 0.5 in all experiments.

Theoretical proof of convergence of the Nelder–Mead simplex method in multi-dimensional space does not exist and the method might be fooled by a single anomalous step. Therefore, it is frequently a good idea to *restart* a multidimensional minimization routine at a point were it claims to have found a minimum. Usually the restart should not be very expensive since the algorithm already converged once.

It should be noted that the above iterative framework separating matching from the optimization can be used only for parameters directly affecting the similarity measure (even if the correspondences between contour points do not change). For example, establishing optimal values of parameters constraining geometry of the least cost path would require rematching of all shapes from the training set.

5.6.4 Optimization Criterion

Many different optimization criteria can be used depending on a particular application. However, smooth and continuous objective functions are advantageous since a discontinuous objective function can affect the convergence of the direct search optimization. One example of a criterion which, for small training sets, may result in a discontinuous objective function is the retrieval rate ratio used in the "Bulls-Eye" test which is based directly on the number of scores (integer values). Therefore, in the experiments presented in the remainder of this chapter, an optimization criterion based on intra- and inter-class distances between shapes from the training set is used instead.

[4]Other direct search algorithms could be used as well, for example pattern search methods or methods with adaptive sets of search directions [222].

Assume a collection $\mathcal{S}$ containing K shapes ($k \in 1, 2, ..K$) from a known set of classes $\mathcal{C}$. Each shape is used as a query and the performance is measured as the ratio between the maximum distance between query shape and a shape from the same class and the minimum distance between the query shape and a shape from a different class:

$$F_q = \frac{\max\limits_{\forall k\ C(s_q)=C(s_k)} D(s_q, s_k)}{\min\limits_{\forall k\ C(s_q)\neq C(s_k)} D(s_q, s_k)} \tag{5.12}$$

where $D(s_q, s_k)$ is the dissimilarity measure between shape s_q and s_k obtained using the evaluated shape matching technique. For a given query shape, s_q, a value of F_q smaller than one means that all shapes from the same class as s_q have lower dissimilarity to s_q than any of shapes not belonging to the class. The overall performance can be calculated as the average of performances obtained for all shapes in the collection:

$$F_{\text{avg}} = \frac{1}{K} \sum_{q=1}^{K} F_q \tag{5.13}$$

Note, that $D(s_q, s_k)$, F_q and F_{avg} are functions of p_i which were omitted in the above equations for clarity. For the remainder of this chapter matching optimization refers to finding combinations of relative values of parameters p_i minimizing F_{avg}.

5.6.5 Optimization Results

The purpose of the following experiments is to demonstrate the capabilities of the proposed alternative version of the MCC representation and, as a byproduct, the optimization framework. The MCC-DCT representation with 10 DCT coefficients per contour point is optimized according to the criterion proposed in section 5.6.4.

Dataset

For all experiments, the dataset from the MPEG-7 Core Experiment "CE-Shape-1" (part *B*), or a subset thereof, is used. As already explained in section 5.4.2, this collection is chosen because of its size, diversity of classes and deformations and also due to its high intra class variability. It is assumed that using such a complete collection in the optimization process will improve overall performance of the general-purpose shape matching technique. Parameters derived by the optimization framework using this collection should provide good performance for many different classes of shapes and deformations. Although, it is still possible to tailor the general-purpose technique to a particular application by using

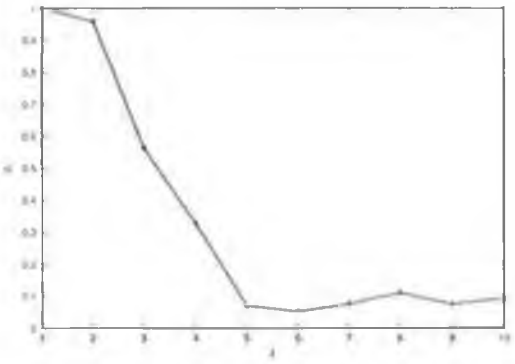


Figure 5.14: Optimal combinations of parameters p_i found by the optimization framework for the MCC-DCT representation.

shapes specific to that application.

The collection has been split into two subsets, one for optimization (training) and another for testing. Every fourth shape from each class was moved into the training set resulting in a collection of 350 shapes from 70 classes (5 shapes per class). The remaining 1050 shapes are used for testing.

Optimization

The optimization framework converged after just two main iterations and five restarts of the Nelder–Mead minimization routine. The optimal combination of parameters p_i found by the optimization framework is shown in Figure 5.14. Weights corresponding to the first five DCT coefficients have significantly bigger values than weights corresponding to the remaining coefficients. This relatively simple form of the p_i combination suggests a good generalization of the training collection. The values of parameters p_i can be interpreted as an estimation of relative importance of those coefficients to effective shape retrieval.

Performance Evaluation

The retrieval performance of the optimized version of the matching algorithm (referred to as MCC-DCT$_{10/O}$) was evaluated using the test dataset. Since each class contained 15 shapes (instead of 20) the retrieval rate ("Bulls-Eye" test) was computed by counting the number of images belonging to the same class in the top 30 matches (instead of 40). The results were compared to the results obtained using the original MCC representation and with the configuration based on a non-optimized matching of MCC-DCT representations ($\forall i\ p_i = 1.0$), referred as MCC-DCT$_{10/E}$.

The evaluation results in terms of average retrieval rate are presented in Table 5.1. The original MCC approach obtained an average performance of 84.49%, which is very close to the result obtained by this approach for the full

configuration	avg. retrieval rate
MCC	84.49%
MCC-DCT$_{10/E}$	84.43%
MCC-DCT$_{10/O}$	86.61%

Table 5.1: Average retrieval rates obtained using different configurations of the matching algorithm.

MPEG-7 collection, indicating that in both cases the testing conditions were similar. The MCC-DCT$_{10/E}$ (non-optimized matching) configuration showed slightly worse performance than the original MCC configuration despite the fact that both descriptors contain exactly the same information. Optimization increased the performance of the MCC-DCT-based technique by approximately 2%. Although, such an increase seems modest it should be noted that some researchers estimated [216] that the average retrieval rate of around 80% in MPEG-7 Core Experiment was already close to saturation.

In order to present a broader perspective on the performance of the techniques, detailed class-by-class performances are shown in Figure 5.15. The collection contains several relatively "easy" classes where all configurations obtained performance close to 100%. The non-optimized MCC-DCT$_{10/E}$ shows improvement compared to the original MCC for the majority of the more challenging classes. This is especially noticeable for classes with distinctive low or high frequency components in the MCC-DCT representation e.g. 'pocked', 'fly', 'tree', 'hCircle'. However, for several classes, MCC-DCT$_{10/E}$ obtained much worse results, e.g. 'device0', 'butterfly', 'chicken', 'device3'. This can be explained by the fact that those classes are characterized by high intra variability occurring at different levels (frequencies). For example, class 'device0' contains approximatively the same number of smooth instances (mainly low frequencies) as more detailed and complex (low and high frequencies) – both kinds of instances from class 'device0' can be seen in Figure 5.8. Because in MCC-DCT$_{10/E}$, all DCT coefficients were incorporated with the same importance, the coefficients corresponding to high frequencies had relatively high impact on the similarity measure resulting in a decrease of the average retrieval performance. The optimization framework was able to find the optimal balance between DCT coefficients resulting in a performance increase for the majority of the challenging classes.

The experiment also confirmed that MCC and MCC-DCT exhibit the same robustness against non-rigid deformations. Both configurations performed similarly for all classes with high non-rigid intra variability, e.g. 'sea_snake', 'spring', 'stef', 'camel'.

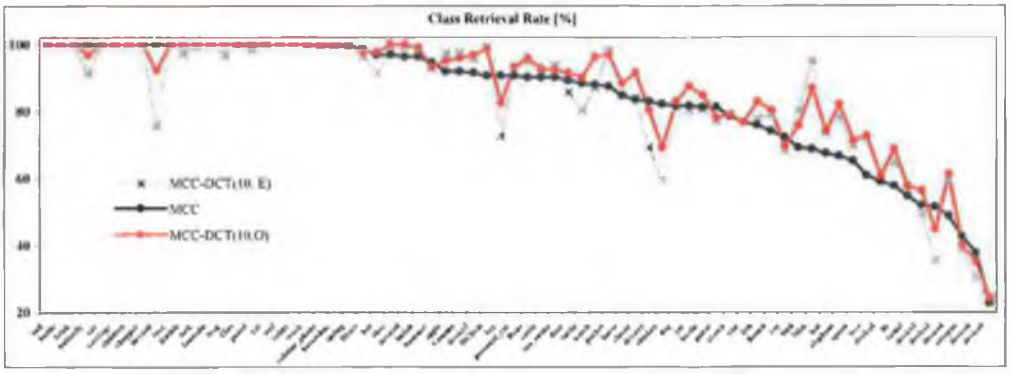


Figure 5.15: Class retrieval ratios for different configurations of the matching algorithm. The classes are sorted according to the performance of the reference MCC configuration, with the corresponding performances of MCC-DCT$_{10/E}$ and MCC-DCT$_{10/O}$ overlaid.

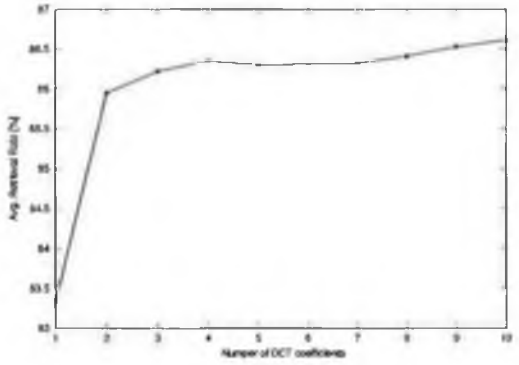


Figure 5.16: Average retrieval ratios vs. number of DCT coefficients.

Compactness versus Performance

The optimal combination of weights p_i (see Figure 5.14) suggests that the DCT coefficients corresponding to low frequencies are somehow more important than the coefficients corresponding to higher frequencies. This is investigated by discarding different numbers of higher frequency DCT coefficients from the matching process. In the experiment, the MCC-DCT$_{10/O}$ configuration with the matching algorithm optimized as in the previous section was used. Retrieval performance is plotted against the number of DCT coefficients used in Figure 5.16. The results indicate that performance obtained by using only the first two DCT coefficients is comparable to performance obtained by using all ten. Therefore, the MCC-DCT descriptor with only the first two DCT coefficients per contour point represents a very good trade-off between performance and compactness. It should be noted that relatively good performance of 83.33% was obtained by using the DC coefficient alone. Although these encouraging results may not hold true for all applications (clearly more complex shapes would require more coefficients), they do illustrate an important property of the new representation.

Finally, it should be stressed that a meaningful comparison between the performances of configurations with different numbers of DCT coefficients was possible only because of the prior optimization. For example, if the non-optimized version of the matching algorithm was used, a conclusion could be reached that 10 DCT coefficients results in worse performance than using only two coefficients.

5.7 Relation to Existing Methods

To the best of author's knowledge, the use of a measure of convexity/concavity for every contour point at multiple scales as a means of matching 2D curves is novel. The most closely related idea in previous work is that of Quddus et al. in their work on wavelet-based multi-level techniques for contour description and matching [209]. This technique uses simple features extracted at high curvature points. These points are detected at each level of the wavelet decomposition as *Wavelet Transform Modulus Maxima* (WTMM). During the matching process, the similarity score is computed at each level and the final similarity score is computed as the mean of the scores at each level. Despite the fact that the method was tested with the MPEG-7 collection [3] the results of its performance specifically in the "Bulls-Eye" test are not available in the literature.

The idea of using dynamic programming (DP) for matching whole contours or parts thereof is not new [204, 172]. Multi-scale methods have been combined with dynamic programming in the past [189]. Recently, very powerful matching algorithms have been proposed by Gdalyahu and Weinshall [170] and by Petrakis et al. [167]. All these algorithms are quite complex. Usually some edit transformation which maps one curve to the other is defined and merging, skipping or stretching parts of the curves is penalized during the matching process. The main advantage of all these approaches is their robustness to occlusions. Some allow uniform matching of the open and closed curves [167]. In contrast, the proposed matching approach is rather simple and originates directly from *Dynamic Time Warping* (DTW) [214] as used in speech recognition and does not require updating the distance table as the matching progresses. It is assumed that the contours are closed. Occlusions are not detected explicitly, but penalized implicitly by the cost of matching them to the non-occluded contour. There is no extra penalty for merging or skipping parts of the contours during the matching process. Rather, a global optimal match between two contours is found using a rich multi-scale feature for each contour point and the differences between shapes are quantified using only distances between feature vectors of corresponding points. The author believes that the above simplifications, allowed by the rich multi-scale representation computed prior to the matching, greatly improved the generality of the approach and contributed to the high

retrieval performance.

5.8 Discussion and Future Work

In this section some aspects and limitations of the proposed approach are discussed and targeted as future research.

5.8.1 Alternative Approaches to Contour Evolution and Convexity/Concavity Measurement

Although the utilized contour evolution method and measure of convexity/ concavity appear intuitive, their employment is driven by ease of implementation. Many alternative approaches could be employed here. For example the so-called *arc-length evolution* could be used [17] which has been shown to be equivalent to heat diffusion [13]. In this approach the curve is parameterized by the normalized arc length parameter and convolved with a Gaussian filter of small value of the standard deviation. The resulting curve is re-parameterized by the normalized arc length parameter and convolved with the same filter. This process is repeated until the curve is convex. The convexity/concavity could be measured as the movement of contour points along normals. Although such approach promises further reduction of computational cost, re-sampling at each iteration could introduce approximation errors.

5.8.2 Different Quantization Schemes for σ

In all of the above experiments, MCC representations were extracted using 10 equally spaced scales. The maximum value of σ was chosen based on the observation that the majority of contours become indistinguishable after filtering with a kernel for which $\sigma > 10$. Adaptation of the number of scales and adaptation of the sampling rate of σ at each level, e.g. more events occur at the lower scales possibly requiring a higher resolution of scale space could be more efficient from a storage point of view. The simplest solution would be to employ commonly used dyadic scale where $\sigma \in 1, 2, 4, 8, 16$. Another alternative would be to adapt the sampling of σ using the training collection by, for example, equalizing the number of events (e.g. number of disappearing extrema) at each sampled level. The above modifications would have to be accompanied by the matching optimization as described in previous sections.

5.8.3 Redundancy Between Contour Points

The MCC-DCT representation still exhibits redundancy along the contour (u axis), which could be reduced by non-uniform sampling, i.e. adapted to the

frequencies present at different scales, or by segment approximation using, for example, B-splines. These solutions were not investigated to date as this chapter is primarily concerned with the investigation of possible benefits of utilizing rich multi-scale representations.

Yet another approach to redundancy reduction could be based on an orthogonal transform, for example *Discrete Fourier Transform* (DFT), applied along the u axis. However such an approach would not allow elastic matching [147].

5.8.4 Limitations of the Optimization Framework

The proposed optimization framework does not make any assumption about the nature of the parameters to be tuned. However, the performance of the Nelder–Mead Simplex method requires specification of the so-called problems' characteristic length scales for each optimized dimension [223]. Also, the method is prone to being trapped in a local minima. An alternative solution to that proposed here would be to directly estimate the weights, e.g. setting their values proportional to the retrieval rates obtained using each DCT coefficient.

5.9 Conclusion

A new method for efficient matching of non-rigid shapes represented with a single closed contour quantifying dissimilarity between them in a way aligned with human intuition has been presented. The approach offers accurate contour matching. Experimental results reveal that the approach supports search for shapes that are semantically similar for humans, even when significant intra-class variability exists. In particular, to the best of the author's knowledge the quantitative results obtained by the proposed approach in both the bulls-eye test and in terms of precision-recall curves for the MPEG-7 collection are the best reported so far in the literature.

Several aspects of the approach remains to be investigated. However, the following chapters focus rather on evaluation/demonstration of the feasibility of the approach in various applications. This itself should lead to better understanding of the advantages and disadvantages of the current solution, identify possible drawbacks, and help to plan future work on the development of the method.

CHAPTER 6

CASE STUDY 1: WORD MATCHING IN HANDWRITTEN HISTORICAL DOCUMENTS

DIGITAL archive construction from historical handwritten manuscripts is an important image analysis application that is of interest for cultural and scientific reasons. Effective indexing is crucial for providing convenient access to scanned versions of large collections of historically valuable handwritten manuscripts. Since traditional handwriting recognizers based on Optical Character Recognition (OCR) do not perform well on historical documents, recently a holistic word recognition approach has gained in popularity as an attractive and more straightforward solution [25]. Such techniques attempt to recognize words based on scalar and profile-based features extracted from whole word images.

This chapter discusses an application of the elastic contour matching technique described in chapter 5 to holistic word recognition in historical handwritten manuscripts. The proposed approach consists of robust extraction of closed word contours and the application of the MCC-DCT contour matching technique. It is demonstrated that contour-based descriptors can effectively capture intrinsic word features. Experimental results provide evidence that word recognition based on contour matching can considerably exceed the performance of other systems reported in the literature.

6.1 Introduction

6.1.1 Motivation

There is an increasing need to digitally preserve and provide access to historical document collections [224]. However, the application of existing technology to this challenging problem exposes numerous problems. Off-the-shelf OCR or

commercial document processing systems are often not able to cope with problems peculiar to historical manuscripts such as poor quality of the manuscript itself, noise, stained paper, contrast variations, faded ink and differences in pressure of the writing instrument. Therefore, recently a significant research effort has been devoted specifically to the analysis of historical documents [225, 226, 227, 25, 224, 228].

In order to be commercially successful, systems for construction of digital archives have to permit integration and subsequent configuration of multiple image and text analysis tools allowing easy re-targeting for different archives. An example of such a configurable system has been described recently in [229] and tested on indexing type-written cards. The main justification of the research carried out here was provided by the fact that handwriting recognizers are key components of such systems since many historical documents are handwritten or contain handwritten annotations.

The approach described here is motivated by the work of Lavrenko, Rath and Manmatha [226, 227, 25] who promote the idea of holistic word recognition for handwritten historical documents. Their main focus is on achieving reasonable recognition accuracy, which enables retrieval of handwritten pages from a user-supplied ASCII query. They argue that, although currently it is not feasible to produce near-perfect recognition results, satisfactory retrieval can still be performed using the noisy outputs of recognizers [230]. In their word spotting approach for indexing historical handwritten manuscripts, pages are segmented into words which are then matched as images and grouped into clusters which contain all instances of the same word. A partial index is constructed for the collection by tagging a number of the resulting clusters.

In this chapter, the primary goal will be to investigate the possibility of matching words using their contours instead of whole images or word profiles. Intuitively, contour-based descriptors should better capture a word's important details and eliminate the need for skew and slant angle normalization. The proposed approach utilizes a rich multi-scale contour representation which eliminates several problems related to holistic representation of words. Since multi-scale representation implicitly captures details at multiple levels there is no need to choose between low-, intermediate- and high-level features (described in subsequent sections) or their integration. Unlike the case of intermediate- and high-level features there is no need for making any decision before the actual matching can proceed. Furthermore, elastic contour matching based on a Dynamic Programming (DP) technique, proposed in chapter 5, should provide a flexible way to compensate for inter-character and intra-character spacing variations.

All experiments are performed on a set of 20 pages from the George Washington collection at the Library of Congress [25]. Since the objective of this chapter is

to illustrate a real application of the proposed shape matching approach, rather then completely study the full problem of indexing handwritten manuscripts the following two simplifications are introduced. It is assumed that bounding boxes for each word are already known[1] and segmented as in [24, 25] and inner holes in letters and dots are ignored.

6.1.2 Previous Work

There has been a large number of approaches to handwritten text recognition proposed in the past. Although typically they cannot be directly used for word spotting in historical documents, some solutions and observations are still relevant to the work presented here.

There has been extensive research devoted to isolated handwritten character recognition, which in theory could be applied to word recognition [231, 232, 121]. A comprehensive review of handwriting recognition approaches considering mainly on-line methods can be found in [231]. Many approaches to alphanumerics [231, 232] and Chinese character recognition are based on some form of elastic matching. Importantly, some studies [231] of the character recognition task reported that elastic matching can halve the error rate of linear matching.

The approaches intended solely for word recognition can be classified as either *analytical* or *holistic*. The approaches in the first category proceed by first identifying the characters and then building a word interpretation. Typical approaches adopt an over-segmentation methodology followed by character model-based recognition using dynamic programming [233]. On the other hand, the holistic approaches attempt to recognize the word as a whole without preliminary letter identification. An excellent discussion of the holistic paradigm can be found in [234]. The review presented here is primarily based on that discussion.

[234] identifies three levels of word features which can be used by holistic recognizers, namely: low-, intermediate-, and high-level. Examples of typical low-level features include stroke direction distribution [235], and various profile-based features [226]. An early approach in this category, presented in [235], used elastic matching with eight direction codes to recognize ten cursively written key words. Typical representative of intermediate-level features are edges, end-points, concavities, diagonal and horizontal strokes [236]. High-level features include ascenders, descenders, loops, i-dots, t-bars and length [237]. For example, the approach in [238] relied on detection of features such as gaps between words, word length, ascenders and descenders for verification of handwritten phrases (lexicon reduction). A dynamic programming algorithm was used to match the predicted features with the extracted image features.

[1] Details of automatic word segmentation can be found in [24].

It should be noted that the introduction of intermediate- or high-level features is mainly dictated by the need for constructing synthetic word models from character models (ASCII), since in many applications the words or phrases to be verified or recognized are not known a priori, i.e. they are not available in the training collection. On the other hand, word spotting in historical documents, as presented in [226, 224], can use a stricter holistic paradigm and directly match the unknown words with annotated words from the training collection. Consequently, word spotting does not necessarily require extraction of intermediate- or high-level features.

To summarize, many approaches to handwritten text recognition currently reported in the literature are not suitable for word spotting in historical documents due to their (i) limitation to single alphanumerics [231, 232], (ii) reliance on character recognition [233], (iii) limitation to applications with small lexicons [235], or (iv) weak discriminatory power suitable for lexicon reduction rather than for word recognition [238].

Given that traditional handwriting recognizers, analytical or holistic, do not perform well on historical documents, recently an increasing research effort has been devoted specifically to the analysis of historical documents [225, 25]. To the best of the authors' knowledge all approaches currently available in the literature are based on the holistic paradigm. So far, the most advanced approaches in this area were proposed by Lavrenko, Rath and Manmatha [226, 227, 25, 224].

Since the quality of historical documents is often significantly degraded it is crucial to select the right features for word matching. The feature set proposed by Rath and Manmatha [226] includes smoothed versions of the word images and various profile-based features. For example, the shape of a word is captured by upper and lower word profiles extracted by traversing the upper (lower) boundary of a word's bounding box and recording for each image column the distance to the nearest "ink" pixel in that column. In order to capture the "inner" structure of a word these features were extended by the number of background to "ink" transitions. It appears that the biggest limitation of the above set of features is their strong dependency on good normalization, mainly skew/slant angle normalization and baseline[2] detection.

In their early approach, profile-based features were compared using *Dynamic Time Warping* (DTW) [227]. They demonstrated on two different data sets from the George Washington collection that their approach outperforms competing matching techniques, including the shape context approach [121] which is currently the best classifier for handwritten digits. They showed in [25] that application of statistical language models can further improve word recognition accuracy as a post-processing step. A simple Hidden Markov Model with one

[2]Baseline - imaginary line upon which a line of text rests.

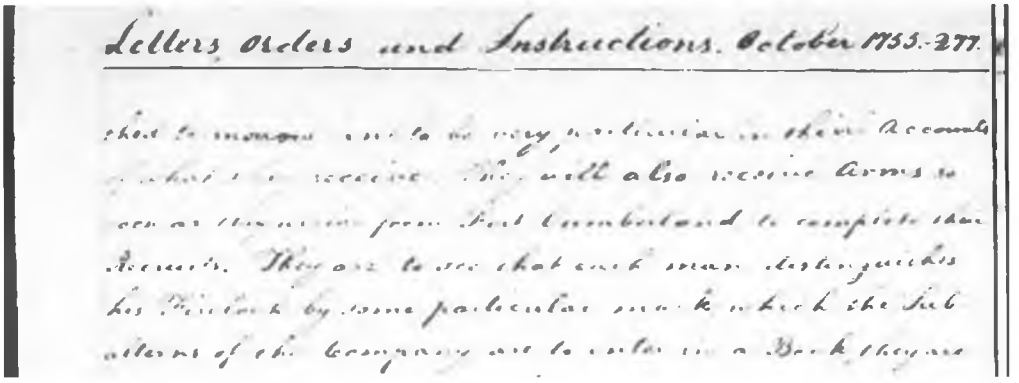


Figure 6.1: Manuscript sample (2770277).

state for each word resulted in over 10% improvement in recognition accuracy. Recently they proposed a complete search engine for historical manuscript images [224]. The main innovation of this system is that it is based on ideas developed in information retrieval, cross-lingual retrieval and automatic image annotation and retrieval rather than on direct matching of word features. Although the system shows promising results on an impressive collection of 987 page images of George Washington's manuscripts, the objective evaluation was carried out only for 26 queries due to the lack of relevance judgments for such a large collection.

6.1.3 Chapter Structure

The remainder of this chapter is organized as follows: the next section describes extraction of a single closed contour for each word. Then, application of the matching algorithm from chapter 5 to word contours is described in section 6.3. Next, initial experimental results followed by a proposal of further adaptations of the matching algorithm are presented in section 6.4. Outlook on future research is discussed in section 6.5 and conclusions are formulated in section 6.6.

6.2 Contour Extraction

Extraction of a single closed contour for each word is a key component of the proposed approach. The extraction procedure has to deal with the poor quality typical of manuscripts (noise, stained paper, etc), contrast variations (faded ink, differences in pressure on the writing instrument) and the most challenging problem – disconnected letters. Figure 6.1 shows a sample of a typical manuscript under consideration. Contour extraction is performed in five steps: binarization, localization of the main body of lower case letters, connected components labelling, connecting disconnected letters and contour tracing.

All parameters controlling the contour extraction are set empirically using

scanned pages of modern handwritten text. It should be noted that theoretically these parameters could be optimized based on the annotated training collection.

6.2.1 Binarization

In the first step, the pixels from the input grey level image are classified as either "ink" or "paper" by comparing their values with a threshold. Since the contrast of the input word image can vary considerably, mainly due to differences in pressure on the writing instrument, there is no single optimal threshold that provides acceptable binarization for different parts of the word. Therefore, the optimal threshold has to be estimated individually for each pixel based on its local neighborhood. A straightforward solution to such problems is offered by the commonly used approach proposed by Niblack [239], where the threshold T for a given pixel is computed using the mean μ and the standard deviation σ of the gray values in a small window centred at the pixel. The threshold is calculated according to the formula proposed by Sauvola et al. [240]:

$$T = \mu \cdot \left(1 - k \cdot (1 - \frac{\sigma}{R})\right) \tag{6.1}$$

where k is a constant set empirically to 0.02 and R denotes the dynamics of the standard deviation and is set to 128. The window size was fixed manually to cover approximatively the width of one stroke.

However, the above approach does not produce smooth word outlines and as such was modified. Prior to binarization, morphological filtering [241] is used to create eroded and opened versions of the input image - the structuring element employed is shown in Figure 6.2a. The dynamic threshold T is calculated based on the eroded image. The binary classification of a single pixel as "ink" or "paper" is made by comparing its value from the opened image with the dynamic threshold calculated for this pixel based on the eroded image.

An example of binarization is shown in Figure 6.2c. Typically, weak strokes (in this example top of double 't' and connections between 'r' and 's') are detected and even dilated without widening the strong strokes. However, it should be stressed that the above approach is just an adaptation of one of many binarization methods available in the literature – see for example [242]. Detailed evaluation of the influence of the binarization stage on the overall performance of the system is outside the scope of this thesis.

6.2.2 Position Estimation

Often binarization alone is not sufficient to obtain a single connected "ink" region for each word. In such cases, separated parts of the word have to be

(a) - kernel (b) - grey-scale image (c) - thresholded image

Figure 6.2: Binarization using dynamic threshold.

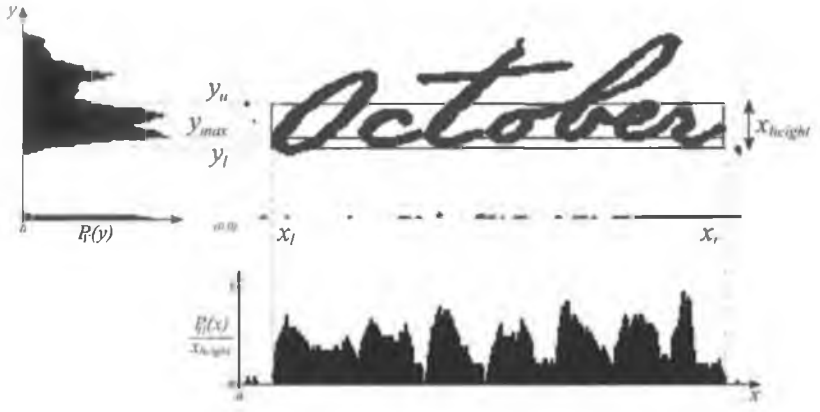


Figure 6.3: Position estimation.

connected. The proposed connecting procedure utilizes the position of the word baseline and the x-height[3]. Detection is performed by analysis of the number of "ink" pixels in each line of the word binary image as illustrated in Figure 6.3.

Let us assume that the origin of the coordinate system lies at the left bottom corner of the word bounding box and that the baseline and the line defining the top of the main body of lower case letters are vertical. Their vertical coordinates are denoted as y_l and y_u respectively. Let x_{height} denote the x-height. Note that $y_u = y_l + x_{height}$. Also, let $P_V(y)$ be a function representing number of "ink" pixels in line y and y_{max} be a vertical coordinate of the line with the maximum number of "ink" pixels. The y_u is located by finding the closest line to y_{max} fulfilling the conditions: $y > y_{max}$ and $P_V(y) < \alpha_u P_V(y_{max})$. Analogically y_l is located by finding the closest line to y_{max} fulfilling the conditions: $y < y_{max}$ and $P_V(y) < \alpha_l P_V(y_{max})$. Parameters α_u and α_l were set empirically to 0.5 and 0.3 respectively.

Similarly, horizontal boundaries of lower case letters are detected by analyzing the number of "ink" pixels in each column of the binary word image. Let $P_H(x)$ be a function representing number of "ink" pixels in column x counted between y_l and y_u. The first column for which $P_H(x)/x_{height} > \alpha_x$ found during scanning the word image from left(right) to right(left) is chosen as the left(right) boundary of lower case letters. The parameter α_x was set empirically to 0.1. Additional rules can be applied to eliminate the residuals of the vertical margin line as

[3]x-height - distance between the baseline and tops of the main body of lower case letters – see Figure 6.3.

illustrated in the example shown in Figure 6.4c.

It should be noted that the position of the baseline and the length of x-height are not directly used by the matching algorithm, but serve only as guidelines for the robust extraction of word contours. Moreover, if all "ink" pixels are already connected in the binarization step, the proper contour can be extracted even if the estimated baseline and x-height are inaccurate. Although precise identification of baseline and x-height is not crucial to the presented approach the above procedure would fail to accurately identify the baseline and x-height in the presence of significant skew. In such cases an alternative methods for baseline detection should be investigates like for example the method presented in [237].

6.2.3 Connected Component Labelling

The connected component labelling algorithm scans the word's binary image and groups "ink" pixels into components based on pixel 8-neighborhood connectivity. Once all groups are determined, each pixel is labelled according to the component it is assigned to. Only components containing a sufficient number of pixels within the area of the main body of lower case letters are retained for further processing, where the required number of pixels is computed as $[0.1x^2_{height}]$. As a result, small "ink" regions and letter residuals from the line above or below the current word are eliminated – examples are shown in Figure 6.4 (4th column).

The above procedure removes punctuation and diacritic marks. This may be problematic for languages which utilize many diacritic marks. However, the results presented here indicate that ignoring diacritic marks, especially in the presence of significant noise and stains, is not critical for Western script.

6.2.4 Connecting Disconnected Letters

Once the relevant components are identified they are sorted based on the horizontal positions of their centre of gravity. Then, successive components are connected by adding the best connecting link (synthetic "ink" line) into the binary image. The author's observations indicate that disconnections within parts of a letter are rare due to the dynamic thresholding. Therefore it is reasonable to assume that disconnections occurs only between letters. The best candidate for the link between two disconnected components is chosen by the following procedure.

At first, candidates for the beginning and the end of the link are chosen from the region on the left and right respectively. The candidates for the beginning (end) of the link are chosen as the first contour pixels from the right (left) in each line

of the left (right) region. Then, candidate links are created by pairing selected points from the left and right region. This is followed by link validation. A link is considered as valid only if:

- both ends of the link are "strictly" inside of the main body of lower case letters – see examples (a), (b) and (e) in Figure 6.4; An end of a link is considered to be "strictly" inside of the main body of lower case letters if its vertical coordinate y fulfils the condition: $y_l < y < (y_u - \alpha_c \cdot x_{height})$, where the value of parameter α_c was set empirically to 0.2.
- both ends of the link are considerably above the main body of lower case letters. – see the top of letter 'I' in example 6.4(h). An end of a link is considered to be considerably above the main body of lower case letters if its vertical coordinate y fulfils the condition: $y > 2y_u - y_l$.

The shortest valid link is chosen as the best "ink" line connecting two consecutive ink regions.

The above heuristic, although simple, performed surprisingly well in all tested manuscripts, typically connecting word components in a very intuitive manner. It should be noted that the above procedure does not necessarily need to connect components in exactly the same manner as a human would. It was assumed that the shape of connections (straight line) and the small discontinuities introduced around them were not crucial for matching as the similarity measure should be influenced by more pronounced parts corresponding to ascenders and descenders. Moreover, even multi-digit numbers (see the last example from Figure 6.4) or words with misplaced connections should be recognized correctly if there are examples in the training collection connected consistently in a similar way.

6.2.5 Contour Tracing

Once all components are connected with the links, the final binary mask for the word can be created – see examples from the 5th column in Figure 6.4. This is followed by a contour tracing procedure which extracts a single ordered contour for each word – see examples from the last column in Figure 6.4.

6.3 Contour Matching

Similarities between word contours are measured using the contour matching technique proposed originally for comparing general shapes in chapter 5. Intuitively, contour-based descriptors should better capture a word's important details than the feature set proposed by Rath and Manmatha [226] and eliminate

Figure 6.4: Intermediate results of the contour extraction process (columns in order from left to right): input gray-scale image, binarization, localization of the main body of lower case letters and removal of margin line residua, connected components alg.(gray levels) and best connecting links (red lines), binary mask, word contour.

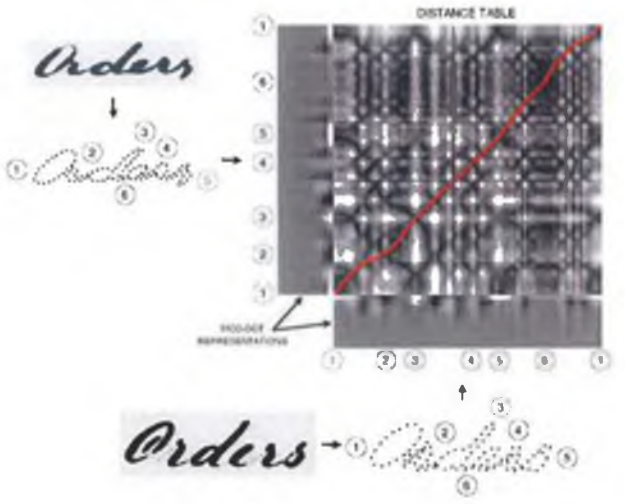


Figure 6.5: Matching words. Finding the optimal match between two contour representations corresponds to finding the lowest cost diagonal path through the table containing pairwise distances between contour point features from both words. The optimal path has to begin and end at the cell corresponding to the alignment of the two starting contour points under consideration. Various strategies can be employed for pairing starting points. Deviations of the path from a straight diagonal path compensates for elastic deformations between contours i.e. stretching and shrinking of parts.

the need for skew and slant angle normalization. Furthermore, elastic contour matching based on a Dynamic Programming (DP) technique should provide a flexible way to compensate for inter-character and intra-character spacing variations.

For the purpose of this chapter, the MCC-DCT version of the contour representation and optimized matching is used. The matching procedure was optimized using the shape collection from the MPEG-7 Core Experiment "CE-Shape-1" (part B) [215]. It should be noted that this dataset contains general shapes and does not contain any examples of word contours. An example of matching word contours using the above method is shown in Figure 6.5.

6.4 Experiments

6.4.1 Overall Results

The approach was tested using a set of 20 pages from the George Washington collection at the Library of Congress. These contain 4,856 word occurrences of 1,187 unique words. The collection was accurately segmented into words using the algorithm described in [24]. Ground-truth annotations for the word images were created manually as in [25].

The contours for all words were extracted according to the procedure described

in section 6.2. Then, for each word contour the MCC-DCT descriptor with 100 equally spaced contour points, each with a feature vector consisting of 10 DCT coefficients ($\sigma_{max} = 10$), was extracted and stored.

Each of the 4,856 word occurrences was used as a query and its contour representation was matched against word representations from all manuscript pages except the page with the query word. A 1-nearest neighbor method was used to classify the query word and the classification result was compared against manual annotation. It should be noted that words from the same manuscript page as the query are excluded from matching for comparability reasons with the results reported in [25].

The results are measured in terms of average Word Error Rate (WER). To separate out-of-vocabulary (OOV) errors from mismatches, two types of WER are reported, one that includes OOV words and one that omits them from the evaluation. In other words, WER that includes OOV errors is the average rate of all incorrectly classified query words - irrespective of the fact that the error was attributed to the performance of the recognition algorithm or simply due to the fact that there was no such word in the searched collection (outside the manuscript page containing the query word). On the other hand, WER that excludes OOV errors is the average rate of all incorrectly classified words for all queries which had at least one instance of the word in the searched collection (outside the manuscript page containing the query word). Therefore, WER that excludes OOV errors will characterize solely the performance of the used recognition method while WER that includes OOV errors characterizes the overall performance of the recognition system which depends also on the size of the annotated collection.

The lowest WER reported until now in the literature for the George Washington collection was 0.449 when OOV words were included and 0.349 when OOV words were excluded [25]. It should be noted that this result was obtained after utilizing additional resources to train a statistical language model. Using a 27-dimensional feature vector representation of each word, without any language model post-processing yielded a WER of 0.603 when OOV words are included and 0.531 when OOV words are excluded. In comparison the method described in this chapter obtained an average WER of 0.306 with OOV words and only 0.174 without OOV words. This represents approximately 50% reduction of the non-OOV error rate obtained by the proposed system despite the fact that it is based on contour mapping only.

6.4.2 Towards Fast Recognition

The contour matching technique described in chapter 5 was proposed originally for comparing general shapes [211, 243]. Although the method can be used

without any changes, this section discusses possibilities for further adaptation to the specific task of word matching in order to improve performance and reduce computational load. In addition, pruning techniques for discarding unlikely matches are proposed. All experiments are performed using a standard PC with 1.6GHz Pentium 4 processor. For clarity, only WER without OOV words is considered in the experiments. Note that the goal of the experiments is to demonstrate the influence of various parameter settings on speed and recognition performance and not a proposal for the final tuning of the algorithm – the latter case would require use of two collections, one for tuning and one for testing.

Modifications to the Matching Algorithm

When matching two shapes, their relative rotation and therefore the circular shift between their contour representations, is generally unknown. Hence, to ensure an optimal match, all circular shifts have to be examined, implying a total complexity of $O(N^3)$, where N is the number of contour points used. In the case of word matching, a single starting point can be located for each contour in such a way that its position along the contour is consistent among different instances of the same word. Matching can then be performed by examining only one circular shift corresponding to the alignment of starting points and therefore reducing the matching complexity to $O(N^2)$.

Here, starting points are selected from the part of the contour corresponding to the end of the word, rather than the beginning, due to better shape consistency. The above choice was motivated by the authors observations that in many cases writers tend to write ends of words in a more consistent way than their beginnings. Moreover, in many cases the end of the word finishes with a well defined "tail" whereas it is often relatively harder to identify a well defined characteristic point at the beginning of the word.

Therefore, starting points are located by scanning the main body of lower case letters and searching for the first "ink" pixel. The scanning starts at the right-bottom and is performed from right to left and from bottom to top as shown in Figure 6.6. Intuitively, the success of the above procedure may depend on writing style and therefore an integrated indexing tool should allow the choice of the most appropriate method for identifying such points.

Utilizing the location of starting points reduced the average matching time for two words from $6ms$ to less than $289\mu s$ whilst maintaining low average WER of 0.182 (without OOV words).

As explained already in chapter 5, finding the optimal match corresponds to finding the lowest cost diagonal path through the table containing pairwise distances between contour points from both contours – see Figure 6.5. The

Figure 6.6: Selection strategy for starting point.

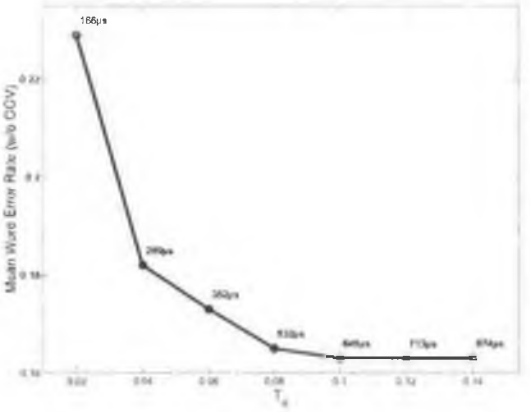


Figure 6.7: Word Error Rate (excluding OOV) as a function of allowed path deviation T_d (starting points utilized). Labels at each data point denote times required for matching two word contours for a given value of T_d.

path has to begin and end at the cell corresponding to the alignment of the two starting points. Deviations of the path from a straight diagonal path compensates for inter-character and intra-character spacing variations.

As has been already discussed in chapter 5, it is common practice in the case of DTW techniques [214], to restrict the path to lie in the area close to the ideal straight diagonal line in order to speed up the matching process. Restricting the path to deviate from the straight diagonal path by no more than the predefined deviation threshold $\pm T_d \cdot N$ reduces the complexity to $O(2T_dN^2)$. The influence of such a restriction on the word recognition rate is demonstrated in Figure 6.7 by plotting WER against different values of the parameter T_d together with corresponding times required for matching two word contours. It can be observed that constraining the path to the diagonal line would result in WER above 0.230 which confirms the requirement of elastic matching (despite the relatively consistent handwriting style in the tested collection). This result suggests that utilization of any method for measuring similarity between contours not taking into account non-rigid deformations would lead to a significant decrease in performance. The experiment also verified the linear dependency between T_d and computational load of the matching. For the remainder of this chapter $T_d = 0.08$ will be adopted as a good trade-off between recognition efficiency and matching speed. However, this setting increases the average matching time to $532\mu s$ whilst the average WER without OOV words is

further reduced to 0.165.

In the current implementation, an equal number of 100 contour points was used for each word contour. Although it may seem reasonable to relate the number of contour points to the length of the word contour in order to reduce the computational load, the author's observations indicate that such modifications could adversely affect recognition performance. Typically, long words have a more distinctive shape (i.e. a distinctive sequence of convexities and concavities) and often can be successfully recognized using a relatively small number of contour points (compared to the contour length). On the other hand, short words are more difficult to recognize, especially words with a small number of ascenders and descenders, therefore requiring a disproportionately large number of contour points to capture more subtle contour features.

Pruning Unlikely Matches

Pruning techniques can be used to quickly discard unlikely matches by requiring word images to have similar statistics [227]. In [244], pruning of word pairs was performed based on the area and aspect ratio of their bounding boxes. This idea was further extended in [245] by the additional requirement of two words to have the same number of descenders (strokes below the baseline). In the proposed approach, the pruning statistics are extracted directly from word contours. Pruning is performed based on contour complexity and the number of descenders and ascenders. For clarity, each pruning rule is first discussed independently and only the best settings for each of the rules is used in the final combined pruning criterion.

The shape complexity x_i of a word contour C_i can be defined as the ratio between its perimeter length l_i and the square root of its area a_i: $x_i = l_i/\sqrt{a_i}$. Both can be easily estimated directly from the word contour [246]. The following rule was used to discard unlikely matches [247]:

Rule 1: Two word contours C_i and C_j are similar only
if $|x_i - x_j|/min\{x_i, x_j\} \leq \tau_x$.

Clearly, increasing τ_x in *Rule 1* allows for more matches to be processed by the matching algorithm (reducing recognition speed) and potentially leads to lower values of average WER. Table 6.1 demonstrates empirically the effect of pruning using different values of threshold τ_x on WER and number of pruned pairs. The results show that setting $\tau_x = 0.3$ would prune almost half of the total number of word pairs, which would otherwise have to be processed by the matching algorithm, while maintaining a low average WER of 0.165.

Word's descenders and ascenders were identified by traversing their contours and labelling contour points as belonging to a descender or an ascender based

Table 6.1: Effect of pruning of word pairs based on contour complexity.

τ_x	$\frac{\sharp pruned\ pairs}{\sharp total\ pairs}$ [%]	WER (w/o OOV words)
0.1	80%	0.215
0.2	63%	0.170
0.3	**48%**	**0.165**
0.4	37%	0.165
...	...	...
∞	0%	0.165

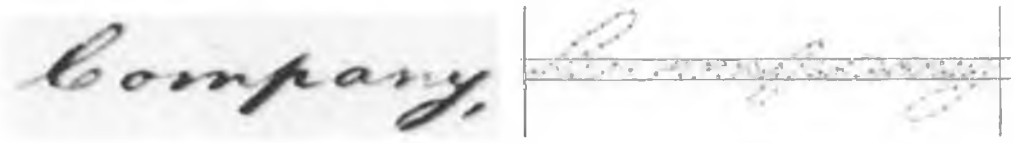


Figure 6.8: Identification of descenders and ascenders.

on the comparison between their vertical positions with the vertical limits of the body of lower case letters – see an example in Figure 6.8. The final number of ascenders and descenders were found by counting the number of continuous sequences of points marked as ascenders or descenders. It should be noted that it may not be important whether or not the "real" ascenders and descenders are detected. Reliable pruning should be possible as long as the handwriting style is consistent, i.e. the identification (or missing) of ascenders and descenders is performed in a consistent way.

Let $DESC_i$ and ASC_i be the estimated number of descenders and ascenders of word contour C_i. The following two rules are used to discard unlikely matches:

Rule 2: Two word contours C_i and C_j are similar only
if $|DESC_i - DESC_j| \leq \tau_{desc}$.

Rule 3: Two word contours C_i and C_j are similar only
if $|ASC_i - ASC_j| \leq \tau_{asc}$.

The thresholds τ_{desc} and τ_{asc} control the tolerance for the absolute differences between numbers of descenders and ascenders when pruning word pairs and should have only integer values. Tables 6.2 and 6.3 demonstrate the effects of pruning on the average WER and the number of pruned pairs using rules 2 and 3 respectively. Table 6.2 shows that the number of descenders provides very reliable evidence for identifying unlikely matches. Specifically, $\tau_{desc} = 0$ would lead to early discarding of 50% of word pairs without noticeable difference in the average WER. In contrast, pruning based on the number of ascenders (Table 6.3) is less reliable most likely due to the generally more complex shape of the upper parts of the words. Therefore, to ensure that only a small proportion of valid matches will be discarded by *Rule 3* a certain tolerance should be allowed, e.g. $\tau_{asc} = 1$ would result in the average WER of 0.164.

Table 6.2: Effect of pruning of word pairs based on number of descenders.

τ_{desc}	$\frac{\sharp pruned\ pairs}{\sharp total\ pairs}$ [%]	WER (w/o OOV words)
0	**50%**	**0.180**
1	9%	0.164
2	2%	0.165
...	...	...
∞	0%	0.165

Table 6.3: Effect of pruning of word pairs based on number of ascenders.

τ_{asc}	$\frac{\sharp pruned\ pairs}{\sharp total\ pairs}$ [%]	WER (w/o OOV words)
0	73%	0.204
1	**30%**	**0.164**
2	9%	0.165
...	...	...
∞	0%	0.165

In the last experiment all three pruning rules were combined:

Rule 4: Two word contours C_i and C_j are similar only if Rules 1,2 and 3 are satisfied.

Incorporating the above rule and setting $\tau_x = 0.3$, $\tau_{desc} = 0$ and $\tau_{asc} = 1$ led to discarding 85% of the total matches while maintaining a low average WER of 0.183 (excluding OOV words).

In summary, the relatively minor adaptation of the matching algorithm and the incorporation of the pruning technique could reduce the average time required to recognize a single word to $0.4s$ (for the particular size of the annotated collection) whilst maintaining a low average WER. This corresponds roughly to a 70 fold increase in speed compared to the original approach.

6.4.3 Comparison with CSS

In the past, different shape matching techniques (typically not requiring parameterization of word contours) have been tested for their potential capabilities to match words. For example, [227] demonstrated that the *Shape Context* approach [121], which is currently the best classifier for handwritten digits, performs poorly in terms of recognition precision and speed for word matching.

The procedure for extraction of single closed contours for each word presented in section 6.2 opens up the possibility of utilizing matching techniques requiring ordering of contour pixels. In this section, the performance of the CSS approach [13] which was adopted as contour-based shape descriptor by the ISO/IEC MPEG-7 standard and discussed in section 4.5.5 is evaluated in this context.

Table 6.4: WER (excluding OOV words) for the discussed methods.

method	WER (w/o OOV words)
profiles-based featurest and HMM [25]	0.349
MCC-DCT	**0.174**
MCC-DCT with pruning	0.183
CSS [13]	0.350

In the experiment, the CSS extraction and matching was performed using the MPEG7 eXperimentation Model (XM software v5.6) [219].

To facilitate comparison of all three approaches, their results are collected in Table 6.4. Adopting the CSS technique for recognition of words from the George Washington collection resulted in relatively high average WER of 0.350 (excluding OOV words). Such poor performance can be explained by two major drawbacks of the CSS technique: the occurrence of ambiguity with regard to concave segments and its inability to represent convex segments. This result confirms the need for utilization of a rich shape descriptor, as discussed in chapter 5.

6.5 Future work

It is acknowledged that the contour extraction procedure has been tested with only one dataset. Therefore more tests should be performed with various handwriting styles and with particular emphasis on generality of rules used for connecting word components and also on the effects of manuscript quality degradation. Also, it would be interesting to evaluate different binarization methods proposed recently (at the time of writing this thesis) in [242] with the variant of Niblack's method described in section 6.2.1.

An obvious improvement to further reduce WER would be the optimization of the similarity measure using the framework proposed in chapter 5 in conjunction with a suitable training collection of handwritten words. Furthermore, currently all contours from the database are re-scaled to a predefined size prior to the extraction of the MCC representations. In the case of word matching, the x-height should be sufficient for size alignment. Replacing the size invariance property with x-height invariance could further improve discrimination between words. The incorporation of additional features in order to capture the "inner" structure of a word and dots, e.g. size and centroid of the inner holes [232], should also be investigated. The number of contour matchings necessary to classify a single word could be further reduced by developing an appropriate indexing strategy for a training dictionary.

One interesting challenge would be augmentation of training dictionaries by synthesizing new word contours based on initial training sets to reduce WER

caused by out-of-vocabulary words. Finally, it should be stressed that, word matching is just a first step in word recognition. Incorporating a statistical language model as a post-process could further improve the recognition accuracy of the system as was shown in [25].

In the future, this algorithm could serve as a starting point for building a complete indexing system for large collections of handwritten historical documents, such as those used for experimentation purposes here, but also illuminated Gaelic manuscripts. Furthermore, in the latter case, a specific challenge of differentiation of multiple scribes within a single document based on characterizing the shapes of words could be investigated. Finally, a feasibility study should be carried out for application of the matching algorithm to on-line handwriting recognition as is required in handheld devices.

6.6 Conclusion

In this chapter, application of the contour matching technique described in chapter 5 to word recognition in historical manuscripts has been proposed and tested. Extensive experimentation conducted using the George Washington collection shows that systems based on word contour matching can significantly outperform existing techniques. Specifically, it has been shown that on a set of twenty pages of good quality, the average recognition accuracy was 83%. Adaptation of the contour matching to the specific task of word recognition in order to improve performance and reduce computational load together with a simple pruning technique were also discussed. Moreover, preliminary investigation of potential simplifications allow speculation that additional development would further improve the performance.

CHAPTER 7

CASE STUDY 2: AN INTEGRATED APPROACH FOR OBJECT SHAPE REGISTRATION AND MODELING

In this chapter, an integrated approach/tool to fast and efficient construction of statistical shape models will be proposed that is a potentially useful tool in *Content-based Image Retrieval* (CBIR). The tool allows intuitive extraction of accurate contour examples from a set of images using the semi-automatic segmentation approach described in chapter 3. A set of labelled points (landmarks) is identified automatically on the set of examples thereby allowing statistical modeling of the objects' shape. The chapter discusses the feasibility of utilizing the method for pairwise correspondence proposed in chapter 5 for automatic landmark identification in sets of contours. The landmarks are used to train statistical shape models known as *Point Distribution Models* (PDM) [70]. Qualitative results are presented for 3 classes of shape which exhibit various types of nonrigid deformation.

7.1 Introduction

7.1.1 Motivation

Object-based retrieval requires important underpinning technology for both object extraction (i.e. segmentation) and indexing in terms of features such as shape, colour and texture. To date, segmentation and indexing have usually been performed independently. For instance, the matching algorithm discussed in chapter 5 allows effective retrieval by shape assuming that objects have been already segmented and their binary masks are available. From another point of

view, the fully automatic segmentation approach described in chapter 3 utilizes only image data (bottom-up approach) without using any prior information about objects present in the scene. In this chapter, an integrated generic approach to shape registration and subsequent generation of shape models for particular classes of objects (e.g. human head and shoulders) that are useful for top-down segmentation and recognition purposes is presented. This chapter focuses on the issues of model generation and not on their subsequent use in a retrieval context.

More than a decade ago Cootes and Taylor introduced one of the more influential ideas within the the image analysis community, the so-called *Active Shape Model*(ASMs) or "Smart Snakes" [70]. Active Shape Models are deformable models with global shape constraints learned through observations. Objects are represented by a set of labelled points. Each point is placed on a particular part of the object. By examining the statistics of the position of the labelled points a *Point Distribution Model* (PDM) is derived. The model gives the average position of the points, and a description of the main modes of variation found in the training set. The statistical analysis of shapes is widely applicable to many areas of image analysis including: analysis of medical images [184, 185], industrial inspection tools, modeling of faces, hands and walking people [117, 186]. The use of PDMs to automatically identify examples of the model object in unseen images and the relations of PDMs with other forms of flexible templates have been presented in [117]. For detailed studies on PDMs the reader can refer to the vast amount of literature available [248, 249, 184, 78, 185, 250].

Statistical shape modeling methods are based on examining the statistics of the coordinates of the labelled points over the training set. The only requirement is a labelled set of examples upon which to train the model. Correspondence is often established by manual annotation, which is subjective, labor intensive and for improved accuracy, intra and inter-annotator variability studies are required.

The goal of this chapter is to develop algorithms necessary for fast and efficient construction of PDMs eliminating the burden of manual landmarking. The proposed approach is based on two key technologies developed in previous chapters: the semi-automatic segmentation approach for fast extraction of examples of closed contours from a set of images discussed in chapter 3 and a new method for automatic identification of landmarks on the set of examples based on the matching technique described in chapter 5.

The typical scenario would be a situation when the user collects several images containing instances of the object under consideration and constructs its statistical shape model, using as small a number of interactions as possible. The process would involve extraction of several object contours by the supervised segmentation. The tool would then automatically place a set of landmarks on each contour and build the required statistical shape model.

7.1.2 Previous Work

In the past, semi-automatic landmarking systems have been developed where the image containing a new example is searched using the model built from the current set of examples and the result edited before adding the new example to the training set. Although such approaches considerably reduce the required effort, many researchers recognize/suggested the need for fully automatic landmark placement. The aim is to place points to build a model which best captures the shape variation and also which has minimal representation error [78]. The methods for finding correspondences across sets of shapes can be classified into two classes [78]: *pair-wise* methods [251, 252, 253] which perform a sequence of individual pairwise correspondences optimizing a pairwise function and *group-wise* methods [254, 250, 255, 256] which explicitly aim to optimize a function defined on the whole set of shapes.

An elegant solution was proposed in [257], where salient image features such as corners and edges were used to calculate a Gaussian-weighted proximity matrix measuring the distance between each feature. A Singular Value Decomposition (SVD) is performed on the matrix to establish a correspondence measure for the features and effectively finds the minimum least-squared distance between each feature. However, this approach is unable to handle rotations of more than 15 degrees and is generally unstable [78]. This method was further extended in [197] and [200]. Although, the above extensions provide good results on certain shapes, stability problems were reported in cases when the two shapes are similar as well as an inability to deal with loops in the boundary [251].

Other methods [252, 253] assume that contours (usually closed) have already been segmented and use curvature information to select landmark points. However, they perform well only if the corresponding points lie on boundaries that have the same curvature. Hill et. al. proposed an alternative method for determining non-rigid correspondence between pairs of closed boundaries [251] based on generating sparse polygonal approximations for each shape. The landmarks were further improved by an iterative refinement step. An interesting system for automated 2D shape model design by registering similar shapes, clustering examples and discarding outliers was described in [258].

Since the above pair-wise methods may not find the best global solution, recently group-wise methods were proposed. In [254], landmarks are placed on sets of closed curves using direct optimization. The approach is based on a measure of compactness and specificity of a model which is a function of the landmark position on the training set. Although in some cases the method produces results better than manual annotation, this is a large, nonlinear optimization problem and the algorithm does not always converge. Recently, Davies et.al. [259] proposed an objective function for group-wise correspondences of shapes based on the *Minimum Description Length* (MDL) principle. The landmarks are chosen

so the data is represented as efficiently as possible. This approach was further extended in [255] by adding curvature measures. A steepest descent version of MDL shape matching was proposed in [256]. Currently, methods based on the MDL principle are seen by many as the state of the art solution: they are fully automatic and in many cases provide more meaningful models than manual annotations. However, the quality of the model directly relies on the choice of the objective function [259, 256] and optimization method which has to take into account many local minima. Often, convergence is slow and scales poorly with the number of examples. The duration of one model-evaluation-tuning loop can take hours or even days. Also, implementation and tuning of the MDL framework requires a lot of experimentation and parameter tweaking [260].

For a more detailed review of methods finding correspondences across contour sets one may refer to [250, 78, 261].

7.1.3 Chapter Structure

The remainder of this chapter is organized as follows: The next section gives a short introduction to the field of statistical shape analysis by describing alignment of sets of points using *Procrustes Analysis* and modeling of shape variation using PDMs. The proposed landmarking framework is described in detail in section 7.4 and section 7.5 presents selected qualitative results. Possible alternative approaches and directions for future research in this area are discussed in section 7.6. Finally, conclusions are formulated in section 7.7.

7.2 Point Distribution Model

This section aims at giving a brief introduction to the field of statistical shape analysis by describing two basic techniques: alignment of sets of points using *Procrustes Analysis* and modeling of shape variation using PDMs.

7.2.1 Pairwise Alignment

According to the definition from [248], shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. Therefore, equivalent points from different shapes can be compared only after establishing a reference pose (position, scale and rotation) to which all shapes are aligned.

A commonly used procedure for aligning corresponding point sets is *Procrustes Analysis*. The Procrustes analysis requires one-to-one correspondence of points often called *landmarks*[1], i.e. points which identify a salient feature on an object

[1]Synonyms for landmarks include homologous points, nodes, vertices, anchor points, markers,

and which are present on every example of the class. Each planar shape is characterized by a set of n landmarks (common for the whole set of examples) and its vector representation is denoted as $\mathbf{x} = [x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n]^T$. Procrustes analysis is based on minimization of the Procrustes distance between two shapes, $\mathbf{x}_1$ and $\mathbf{x}_2$, which is simply the sum of the squared point distances:

$$P_d^2 = \sum_{j=1}^{n} [(x_{j1} - x_{j2})^2 + (y_{j1} - y_{j2})^2] \tag{7.1}$$

The alignment of two shapes $\mathbf{x}_1$ and $\mathbf{x}_2$ involves three steps:

1. Align position of the two centroids,
2. Re-scale each shape to have equal size,
3. Align orientation by rotation.

The centroid is defined as the centre of a mass of the physical system consisting of unit masses at each landmark:

$$(\bar{x}, \bar{y}) = \left(\frac{1}{n} \sum_{j=1}^{n} x_j, \frac{1}{n} \sum_{j=1}^{n} y_j \right) \tag{7.2}$$

Size normalization is performed using a scale metric called *centroid size*:

$$S(\mathbf{x}) = \sum_{j=1}^{n} \sqrt{(x_j - \bar{x})^2 + (y_j - \bar{y})^2)} \tag{7.3}$$

Orientation alignment is achieved by maximizing the correlation between the two sets of landmarks. The shape $\mathbf{x}_1$ is optimally superimposed upon $\mathbf{x}_2$ by performing a *Singular Value Decomposition* (SVD) of matrix $\mathbf{x}_1^T \mathbf{x}_2$ and using $\mathbf{V}\mathbf{U}^T$ as the rotation matrix.

7.2.2 Generalized Procrustes Analysis

The alignment of a set of shapes is typically performed in an iterative[2] manner [184]:

model points and key points.

[2] An analytic solution can be found in [248].

1. Chose an initial estimate of the mean shape (e.g.the first shape in the set),
2. Align all shapes to the mean shape,
3. Re-calculate the estimate of the mean from the aligned shapes,
4. If the mean has changed return to step 2.

The most frequently used estimate of the mean shape is the *Procrustes mean shape*:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{7.4}$$

where N denotes the number of shapes.

It is important to fix the size and orientation at each iteration by normalization in order to avoid any shrinking or drifting of the mean shape.

However, in fact, not all sets require orientation adjustment during alignment of examples to the mean (in step 2). In many applications orientation of examples is already fixed within the scene, e.g. the "Head & Shoulders" example from section 7.5.2. The author's observations indicate that, in such cases, correcting orientation of examples in order to minimize the Procrustes Distance (equation 7.1) may result in additional artificial (non-intuitive) modes of variation. Therefore, for flexibility the tool developed in this chapter allows skipping the orientation alignment. In practice, it might be useful to allow utilization of alternative alignment methods.

7.2.3 Modeling Shape Variation

Once a set of aligned shapes is available the mean shape and variability can be found. In this work, statistical shape models known as *Point Distribution Models* (PDM) [70] are used.

Consider the case of having N planar shapes consisting of n points, covering a certain class of shapes and all aligned into a common frame of reference. Training PDMs relies upon exploiting the inter-landmark correlation in order to reduce dimensionality. It involves calculating the mean of the aligned examples $\bar{\mathbf{x}}$ (according to equation 7.4), and the deviation from the mean of each aligned example $\delta\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, and calculating the eigensystem of the $2n \times 2n$ covariance matrix of the deviations $\sum_{\mathbf{x}} = 1/N \sum_{i=1}^{N} (\delta\mathbf{x}_i)(\delta\mathbf{x}_i)^T$. The modes of variation are described by $\mathbf{p}_k$ ($k = 1, \ldots, 2n$), the unit eigenvectors of $\sum_{\mathbf{x}}$ such that:

$$\sum_{\mathbf{x}}\mathbf{p}_k = \lambda_k\mathbf{p}_k \tag{7.5}$$

where λ_k is the k^{th} eigenvalue of $\sum_{\mathbf{x}}$ and $\lambda_k \geq \lambda_{k+1}$. This results in an ordered

basis where each component is ranked according to variance. Modifying one component at a time gives the *principal modes* of variation. Any shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first t modes:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} \tag{7.6}$$

where $\mathbf{P} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_t})$ is the $2n \times t$ matrix of the first t eigenvectors and $\mathbf{b} = (b_1, b_2, \ldots, b_t)^T$ is a t element vector of weights (shape parameters), one for each eigenvector. New examples of shapes can be generated by varying the parameters $\mathbf{b}$ within suitable limits, so that the new shapes will be similar to those in the training set. t is typically chosen as the smallest number of modes such that the sum of their variances explain a sufficiently large proportion of the total variance of all the variables.

The modes are often comparable to those a human would select to design a parameterized model. However, as explained in [78], they may not always separate shape variation in an obvious manner since they are derived directly from the statistics of a training set.

7.3 Semi-Automatic Segmentation

Since a typical model requires tens of examples, special consideration is devoted to ensure easy and intuitive user interactions in a segmentation process so that generation of these examples is as simple as possible. Contour examples are extracted using the semi-automatic segmentation approach discussed in chapter 3. This approach has proven to be extremely practical and efficient, allowing rapid segmentation while providing the flexibility necessary to obtain high accuracy in extracted examples.

7.4 Landmark Identification

The strategy used in the framework for automatic identification of landmarks on a set of contours is similar to the scheme proposed by Fleute and Lavalée for landmarking sets of closed 3D surfaces [262]. Each training example is initially matched to a single reference template and a mean is built from these matched examples. Then, iteratively, each example is matched to the current mean and the procedure is repeated until convergence.

7.4.1 Pair-wise Matching

The above framework relies upon the ability to match pairs of shapes (in order to match each shape with the mean) and to measure the quality of the match (in order to decide the initial reference shape). In other words, the mean cannot be built and subsequently updated if there is no dense correspondence between contour points of the whole set. In order for this process to be successful, an accurate, robust method for establishing pairwise correspondence is required. It is also required that the pairwise matching is invariant with respect to pose (translation, scale and rotation). Furthermore, the method has to perform well in cases where the two boundaries represent different examples from the same class of objects and a nonrigid transformation is required to map one boundary onto the other.

This chapter studies the usefulness of the pairwise shape matching technique presented in chapter 5 for the above landmarking framework. The method is particulary attractive for the above application as has been demonstrated, in chapters 5 and 6, it performs well in cases of elastic deformations and where the similarity between curves is weak [211]. For the purpose of the landmarking framework, the MCC-DCT version [211, 243] of the contour representation is used. The matching procedure was optimized using the shape collection from the MPEG-7 Core Experiment "CE-Shape-1" (part B) [215] as discussed in section 5.6. This choice of this configuration was based on the assumption that its better recognition performance on collections containing general shapes has been resulted by more intuitive matching.

It should be noted that the matching algorithm requires the two contours to be represented by an equal number of equidistant points. This will lead to an additional re-sampling step in the landmarking algorithm. The extension of the matching algorithm to a non-uniformly sampled contour is straightforward and will be part of authors' future work. Also, the landmarking scheme requires dense correspondence (ideally between every pixel). To allow such fine matching without increasing the computational cost, faster versions of the matching algorithm could be utilized/tested, e.g. based on the strategy for fast selection of starting points [170].

7.4.2 Correspondence for a Set of Curves

Before the landmarking process begins, all examples are down-sampled and represented by a fixed number of equidistant contour points (300 in all experiments in this chapter). Then, the position and size of the contours are normalized according to equations 7.2 and 7.3. Next, an initial reference contour is selected from the set in such a way that the total value of dissimilarities between the reference and all other examples is minimized. This ensures that the initial ref-

erence is chosen close to the centre of the modelled shape space. Subsequently, each example is matched to the reference using the pairwise matching method discussed in section 7.4.1. This provides a common basis for identification of points from all examples. In other words, a contour point in any of the examples is identified by the point from the reference contour to which it is best matched. This allows optimal rotation of examples to superimpose with the reference (as explained in section 7.2.1) and finally construction of the mean contour according to equation 7.4. If the distance between the current reference and the newly estimated mean (computed using equation 7.1 and normalized by the centroid size of the reference), exceeds a predefined threshold the mean is taken as the new reference and the procedure is repeated, otherwise the procedure stops. As a result, a dense correspondence between each example and the mean is obtained. Using these dense correspondences the landmarks placed on the mean shape can be projected back to each example. The main steps of the proposed framework are outlined below:

1. Sample uniformly all examples,
2. Normalize all examples:
   a) Translate centroid to position $(0, 0)$;
   b) Re-scale to unit size,
3. Select the initial reference contour,
4. Find correspondences between examples and the reference,
5. Rotate each example to superimpose with the reference,
6. Create a mean contour according to Eq. 7.4,
7. Take the mean as the new reference,
8. If the difference between the old and the new reference exceeds a predefined threshold then go to step 4.,
9. Generate landmarks in the reference contour,
10. Project landmarks back to each example.

To avoid any drifting of the mean shape its orientation is adjusted at each iteration by optimal superposition (minimization the Procrustes distance) with the initial reference. Additionally, the mean contour is re-sampled before a new iteration begins to ensure equidistance of its contour points. This additional step could be eliminated by adapting the pairwise matching algorithm to match non-uniformly sampled contours.

The objective of step 4. is to identify, on each example, positions corresponding best to the contour points from the reference mean. However, the matching algorithm must compensate for nonrigid transformations as well as small differences in the position of contour points caused by sampling. The matching is symmetrical which means that both contours (an example and the reference

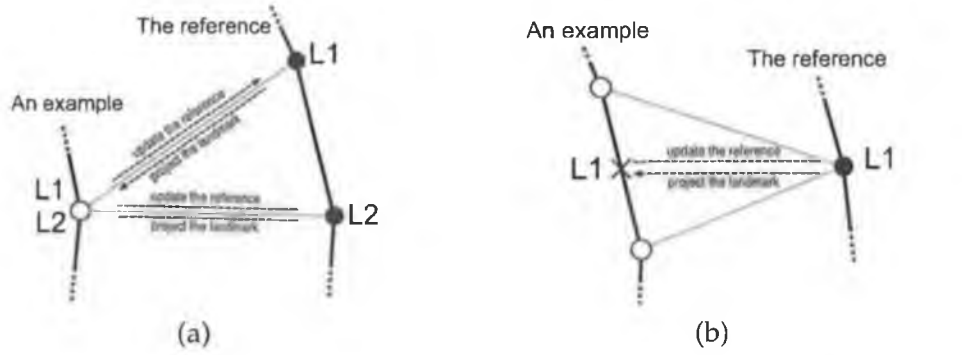


Figure 7.1: Updating of the mean and landmark propagation in cases of multiple assignments of points.

mean) are treated equally. Therefore, a single contour point from an example can be matched with two reference points from the mean, and vice versa. The first case, depicted in Figure 7.1a, does not require any special treatment. The position of the single point from the considered example can be used to update the positions of both matched points from the reference, e.g. to construct the new mean. Also, during stage 8, if both reference points from the mean are selected to become landmarks, the position of the single point can be used for both of them without causing any conflict. In contrast, the second case requires special consideration since the single landmark (L1) cannot be used for two different points from the considered example. Therefore, only the centre of the segment connecting both points is used to update the position of the matched reference point and can potentially become a landmark – see Figure 7.1b.

The number of contour points should be considerably large compared to the required precision in localization of the final landmarks. For maximum precision, the number of contour points should be comparable with the average number of pixels in each example.

The set of landmarks can be generated automatically on the mean using any sensible method, for example the approach based on detection of *Critical Points* presented in [263]. In the current implementation all points from the mean were selected as landmarks. However, choosing the extrema of convexity/concavity measure from the MCC representation as major landmarks (due to high matching reliability at those points) and placing the minor landmarks equally between them could further improve specificity of the final PDMs.

The selection of the initial reference contour requires computation of pairwise dissimilarities between all N examples implying $O(N^2)$ complexity. Since a coarse measure of the dissimilarity is sufficient for the selection, the computational load of this step can be reduced by using a reduced number of contour points.

7.4.3 Modeling Symmetrical Shapes

Often one would like to model shapes exhibiting mirror symmetry, e.g. a face or a human head & shoulders. In such cases, extending the training set by mirrored versions of the boundaries usually leads to more intuitive modes of variations – especially compared with cases where originally there is a discrepancy between the numbers of examples from each view (e.g. more faces looking to the left than to the right) and the number of examples is small. Clearly such an approach does not require any modification of the proposed landmarking method, however it implies additional cost of matching of these additional examples with the mean. An alternative strategy requires adaptation of steps 4-6 of the proposed framework. In such a case, each iteration begins by matching only the original set to the mean. Based on this match, a new mean and its mirrored version are created. Then, both versions of the mean (original and mirrored) are matched and their correspondence is used to establish correspondences between the mirrored set of contours and the mean obtained from the original set. Finally, the new symmetrical version of the mean, including original and mirrored sets, is created. This approach avoids the need for individual matching of mirrored examples, explicitly ensures that each mirrored boundary is matched with the reference in exactly the same way as its original version[3], and further improves convergence.

7.5 Results

This section presents qualitative results of training PDMs for 3 classes of shapes exhibiting various types of nonrigid deformation. In all three cases, the supervised segmentation of the set took less than 5 minutes. The landmarking process converged after 2-5 iterations in all cases, requiring typically less than 30 seconds on a standard PC.

7.5.1 Cross-sections of Pork Carcasses

The data in this experiment consists of 14 gray scale images (768x576 pixels each) [77]. An example image is shown in Figure 7.2a. The results are comparable with the manual annotations presented in [77]. It can be observed from Figure 7.2b that the model is compact (the three first eigenvectors explain more than 77% of all variations) and qualitative results, shown in Figure 7.2c, indicate good specificity [4].

[3] With the precision limited by the contour sampling.

[4] A specific model should only generate instances of the object class that are similar to those in the training set.

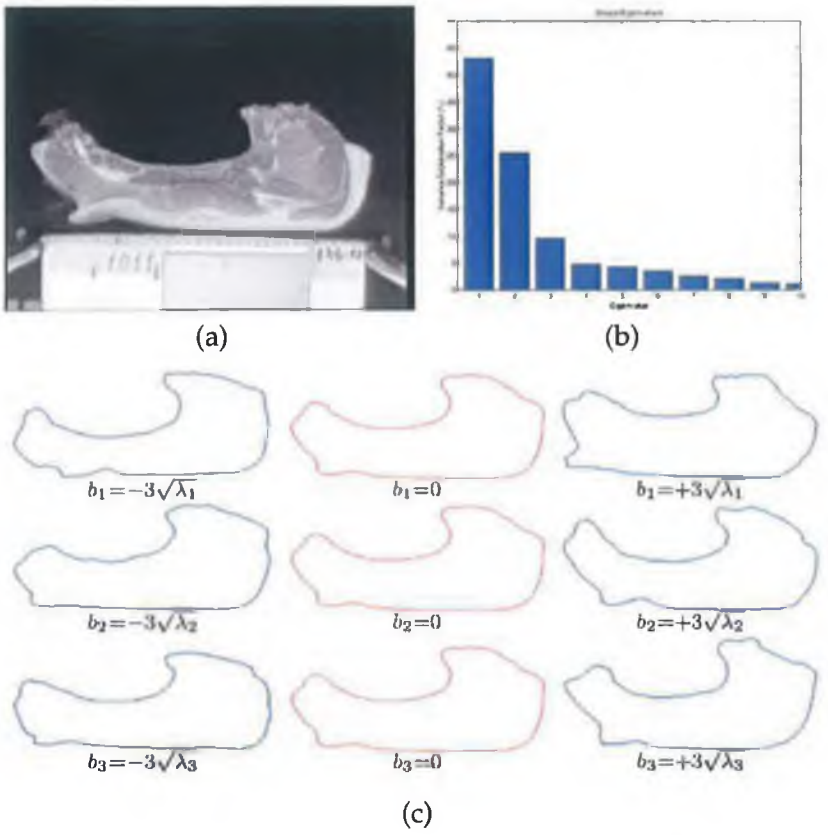


Figure 7.2: Pork carcasses: (a) - Example image of pork carcass cross-section, (b) - Eigenvalues, (c) - Deformation using 1st, 2nd and 3rd principal modes.

7.5.2 Head & Shoulders

The data in this experiment consists of 19 colour images of CIF size. Example images (with extracted object boundaries) are shown in Figure 7.3a. The training set was extended using mirrored versions of the examples as described in section 7.4.3. The final model is compact (three first eigenvectors explain more than 76% of all variations) and qualitative results shown in Figure 7.3c indicate good specificity.

7.5.3 Synthetic Example

In this experiment, synthetic data designed to "stress test" the model, in order to illustrate some potential limitations of the proposed approach is used. The training set, shown in Figure 7.4, is a synthetic "bump" example introduced and discussed originally in [250]. The examples were generated from a synthetic object that exhibits a single mode of shape variation where the "bump" moves along the top of the box. Therefore, it should be possible to describe the variation with a single parameter.

Qualitative results for the proposed approach are shown in Figure 7.4. The algorithm failed to find "ideal" (the most intuitive) correspondences for the

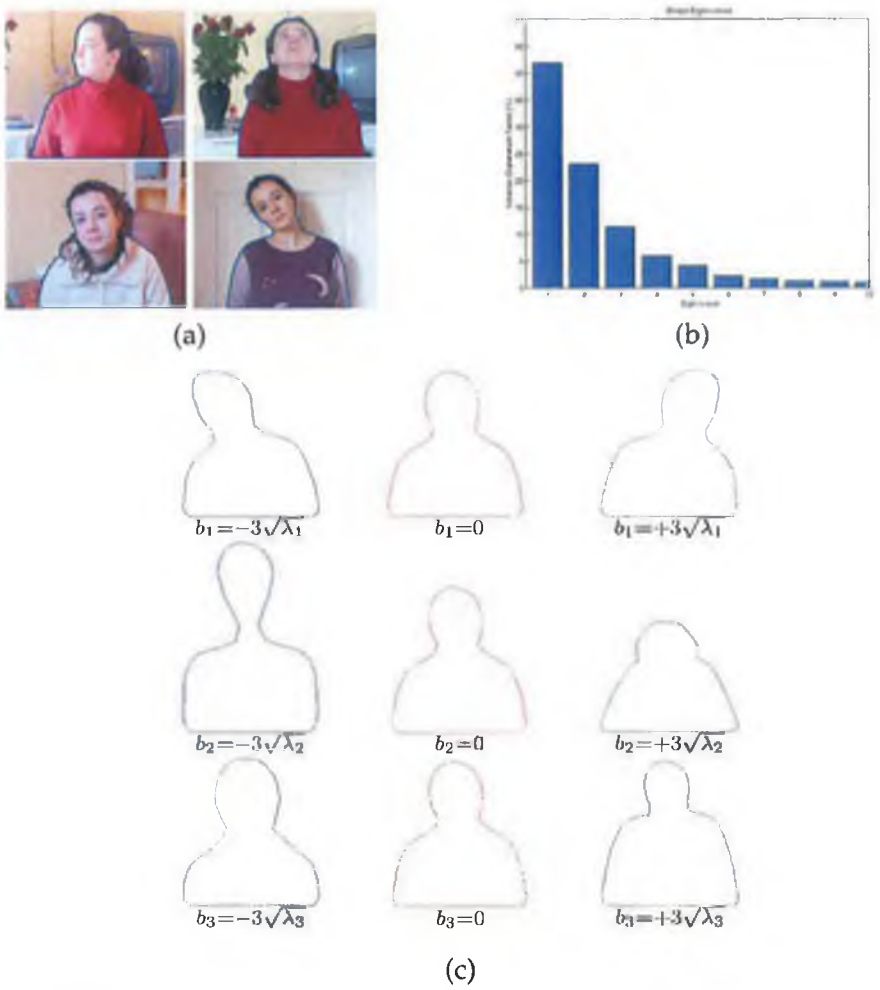


Figure 7.3: Head & Shoulders: (a) - Examples of contours extracted using user-driven segmentation tool for the head & shoulders model, (b) - Eigenvalues, (c) - Deformation using 1st, 2nd and 3rd principal modes.

set as some of the "bump" end-points lie on boundaries with concavity while others correspond to convex parts. The matching algorithm could only warp the parts with different curvature attempting to find optimal global correspondence. The poor localization of the end-points of the "bump" for some examples compromised the specificity of the model as shown in Figure 7.4b. It should be noted however, that the mean has desired shape and the first eigenvector still describes almost 98% of the total model variations. This example suggests that making the approach fully functional (applicable to all classes of shape deformations) would require allowing manual corrections of some landmarks and/or further optimization, e.g using a MDL-based approach.

7.6 Discussion

The main limitation of the approach is related to the fact that the matching relies on the assumption that the corresponding points lie on boundaries that have

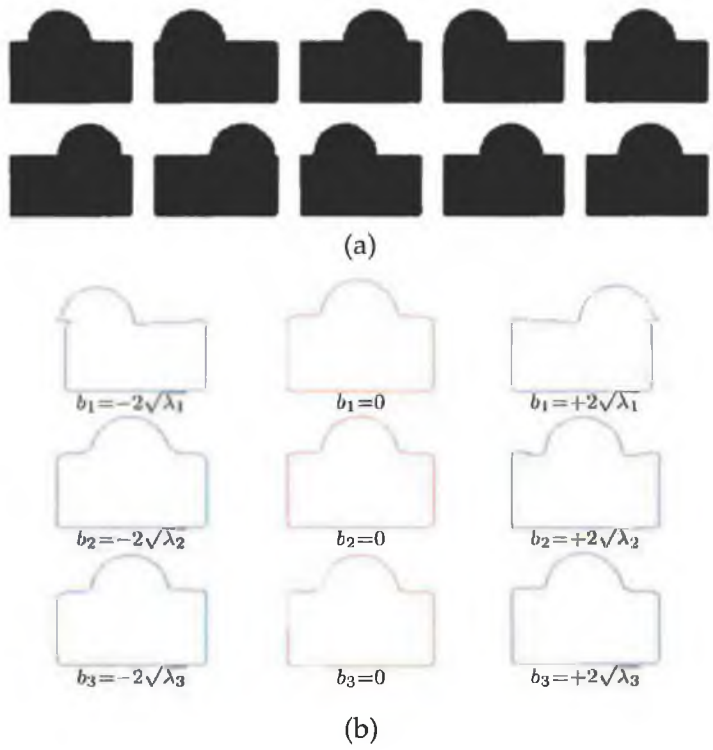


Figure 7.4: "Bump": (a) - The "Bump" training set, (b) - Deformation using 1st, 2nd and 3rd principal modes.

similar values of curvature (convexity/concavity). As shown in section 7.5.3, this may compromise the compactness and specificity of the final model. One solution would be to allow the user to correct the position of misplaced landmarks. Alternatively, a MDL-based method could be used to further refine the correspondences as these approaches do not rely on pairing any boundary features. In other words, the proposed method could provide initial correspondences which could be then refined using a MDL-based technique. Such a hybrid approach could reduce the sensitivity of the MDL method to the search parameters and reduce the time required for convergence. Also, neither MDL methods or the proposed method are suitable for objects represented by multiple open/closed boundaries, e.g. face expressions. Addressing this problem will be part of author's future research. One approach would be to integrate a method based on *Shape Context* [121] or its recently proposed extension with a figural continuity constraint [264].

The extension of the proposed approach to open curves would require extension of the user interface to allow extraction of open curves (e.g. by marking end-points of the closed curves) and a relatively straightforward modification of the pairwise matching algorithm. In fact, the matching problem could be simplified by assuming that the end-points of the curves correspond.

Summarizing, the tool would have to integrate multiple modules allowing different types of interaction and generated models for full flexibility in different

applications.

7.7 Conclusion

The proposed approach allows rapid creation of statistical shape models for classes of objects represented by single closed boundaries. The tool allows intuitive semi-automatic segmentation of objects' examples from a set of images, automatically identifies correspondences for the extracted set of contours and creates Point Distribution Models. The proposed landmarking method is fast and does not require any tweaking of the parameters. In all presented examples, it took less than 5 minutes for an unexperienced user to segment and create a final model.

CHAPTER 8

GENERAL CONCLUSIONS AND FUTURE WORK

THIS final chapter summarizes the thesis by reviewing the achieved research objectives and recalling the main conclusions drawn throughout this thesis. Also, it indicates directions for further research by discussing possibilities for further improvement of the proposed techniques as well as their use in related challenges in the field.

8.1 General Discussion

In recent years content based retrieval of multimedia data has become a very active research area resulting in initiatives such as a dedicated ISO standard for description of multimedia content named MPEG-7, large scale evaluation activities, e.g. TrecVid, and a large number of large-scale academic and industrial CBIR systems. However, it is safe to say that the current CBIR systems are in their infancy and their performance when dealing with general content is rather poor. Moreover, the majority of currently available systems are able to deal only with collections which are reasonably homogenous. Clearly providing content based functionalities for heterogenous collections such as the content of World Wide Web is much more demanding. Therefore, one can expect the problem of content based retrieval of multimedia data to remain a challenging research area for many years to come.

Although image or video content is clearly much more versatile compared to text, their interpretation has rarely a unique meaning. Such ambiguity implies considerable difficulties in finding optimal ways to abstract images or video using visual descriptors.

In recent years, it has become clear that in order to provide content based functionalities both low-level visual features and high-level semantic features

should be be integrated. However, both the optimal choice of low-level features and also capturing of the semantic concepts based on the automatically extracted low-level features is context, user, and query dependent. Furthermore, it is also not clear how to intertwine contextual information into numerical similarity measures used by computers.

Currently the problem of extraction of semantically meaningful entities from visual data can be successfully solved only by techniques tuned to a very narrow application scope and which are not easily extensible to other application contexts. In other words, current image processing techniques are unable to automatically extract semantically meaningful entities from general visual content.

Although the capabilities of existing CBIR systems and their underpinning technologies seem quite limited taking into account the typical variability of the multimedia content, they certainly help to better understand the problem of retrieval by content which hopefully will lead to new advances in the area.

It is essential that future CBIR systems further utilize contextual information and knowledge of the world. Also it should be accepted that interpretation of the content depends strongly on the user and his/her query and that the development objective for future CBIR systems should not be a further reduction of the user's role but rather embracing it by providing better querying models and relevance feedback mechanisms.

Many of the major obstacles to the development of efficient CBIR systems may be solved by future intelligent acquisition devices. In many applications new devices could provide valuable contextual information, e.g. time and location of capturing the data. Also, the problem of image segmentation and extraction of semantic entities would simplify if the new acquisition devices could provide additional modalities, e.g. depth or temperature. However it should be noted that even in such cases many of the fundamental problems of image segmentation would remain essentially the same, e.g. fusion of the information from multiple sources or the choice between bottom-up and top-down processing. Besides, even in the case where all semantically meaningful entities are successfully extracted, the problem of measuring similarities between such entities and integrating contextual information into numerical measures remains to be solved.

8.2 Thesis Summary & Research Contributions

The content analysis work reported in this thesis has focused on developments in the areas of image segmentation and shape matching in the context of content based functionalities required by future multimedia applications. The

limitations of current techniques from both of these areas have been evaluated and extended. Strengths and weaknesses of the proposed solutions have been demonstrated in the context of selected applications. The author hopes that this contribution leads to a better understanding and perhaps further advances in the field of content analysis.

The first chapters of the thesis consider segmentation of images of real world scenes while the later chapters focus on estimation of similarities between objects' silhouettes. The final chapters present case studies demonstrating the usefulness of the proposed solutions in selected applications.

The research areas of image segmentation and shape analysis are discussed in the context of content-based image retrieval and motivations for the investigation carried out by the author are outlined. The current state of the art in image segmentation and shape matching is illustrated by reviewing the most characteristic categories of techniques in both areas of research and briefly discussing the most interesting and promising approaches from each category.

The feasibility study of utilizing spatial configuration of regions and structural analysis of their contours (so-called *syntactic visual features* [18]) for improving the quality of fully automatic image segmentation performed via region merging is successfully carried out in chapter 3. Several extensions to the well-known *Recursive Shortest Spanning Tree* (RSST) algorithm [19] are proposed among which the most important is a novel framework for integration of evidence from multiple sources of information, each with its own accuracy and reliability, e.g. colour homogeneity and a region's geometric properties, with the RSST region merging process. The new integration framework is based on *Theory of Belief* (BeT) [20, 21, 22, 23]. The "strength" of the evidence provided by the geometric properties of regions and their spatial configurations is assessed and compared with the evidence provided solely by the colour homogeneity criterion by extensive experimentation. Other contributions in this area include a new colour model and colour homogeneity criteria, practical solutions to structure analysis based on shape and spatial configuration of image regions, and a new simple stopping criterion aimed at producing partitions containing the most salient objects present in the scene.

An application of the automatic segmentation method to the problem of semi-automatic segmentation is also discussed. An easy to use and intuitive semi-automatic segmentation tool utilizing the automatic approach is described in detail. This example demonstrates that syntactic features can be useful also in the case of supervised scenarios where they can facilitate more intuitive user interactions e.g. by allowing the segmentation process make more "intelligent" decisions whenever the information provided by the user is not sufficient. Utilization of the approach in an integrated tool allowing intuitive extraction and registration of contour examples from a set of images for construction of

statistical shape models is discussed in chapter 7.

Selected issues of shape analysis, in particular contour representation, matching, and estimation of similarity between silhouettes are investigated in chapter 5. A novel rich multi-scale representation for non-rigid shapes with a single closed contours and a study of its properties in the context of silhouette matching and similarity estimation is presented. An initial approach for optimization of the matching in order to achieve higher performance in discriminating between shape classes is also proposed. The efficiency of the proposed approach is demonstrated in a variety of applications.

In particular, the retrieval results are compared to those of the *Curvature Scale Space* (CSS)[13, 14, 15](adopted by the MPEG-7 standard [16, 17]) and it is shown that the proposed scheme performs better than CSS in two applications: retrieval in collections of closed silhouettes and holistic word recognition in handwritten historical manuscripts. As a matter of fact, it is shown that the proposed contour based approach outperforms all techniques for word matching proposed in the literature when tested on a set of 20 pages from the George Washington collection at the Library of Congress [24, 25]. Additionally, it is shown that the matching approach can be extended with some success to the problem of matching multiple contours and therefore facilitating establishment of dense correspondence between multiple silhouettes which can then be employed for contour registration.

Through presentation of various applications of the proposed solutions it is demonstrated that all contributions have the potential of providing the key technologies enabling content-based functionalities in the future multimedia applications. In each case, the performance and generality of the proposed techniques is analyzed based on extensive and rigorous experimentation using as large as possible test collections.

8.3 Future Work

This section revisits the discussion about the role of image segmentation and shape matching in the context of content based retrieval and proposes a number of research problems which, according to the author, are worth investigating to fully/further exploit the potential of the approaches proposed in this thesis.

It should be noted that incremental improvements to all proposed techniques have already been discussed in previous chapters. This section focuses on applications of the proposed approaches which have not been sufficiently investigated in this thesis.

8.3.1 Image Segmentation

Although the problem of partitioning an image into a set of homogenous regions or semantic entities is typically considered to be a fundamental enabling technology for understanding scene structure and identification of relevant objects, recently its role in the context of content based retrieval has been questioned.

For example, despite recent advances in the area of region-based image segmentation (including the one proposed in this thesis) to the best of the author's knowledge no one has reported quantitative results demonstrating that utilizing local features representing homogenous regions leads to better retrieval capabilities in AV material with general content than using features representing image blocks or even entire images.

To better illustrate the problem let us consider CBIR systems for visual data with general content. In order to allow reasonable response time to the query, the computation of visual descriptors and the creation of the index should be done off-line. However, often the selection of the optimal features is both context and query dependent. For example, extraction of local features, where ideally each feature describes a semantic object or at least an important part of an object, requires image segmentation. However, automatic bottom-up segmentation is often under-constrained as well as context and query dependent due to the inherent ambiguity of the visual content. An alternative top-down solution to image segmentation requires utilization of domain-specific models of the semantic entities to be extracted. However the relevancy of the objects is again query dependent and it is currently impossible to build such models for all semantic objects that could possibly appear in the scene. Also, even in cases where the semantic objects can be automatically extracted, the problem of estimation of similarity between the query and the target image remains to be solved. In fact, the problem of similarity estimation based on local features appears to be much more difficult when utilizing local features than simple global descriptors. For example, some researchers reported that in the case of CBIR systems allowing query by example with replacement of the global features with local features describing individual homogenous regions leads to retrieval improvement only in some very specific cases (e.g. scenes with a single dominant semantic object with homogenous colour) while decreases the performance in others (e.g. scenes defined by composition of multiple visually distinctive objects). Also, the vast majority of approaches to scene classification into semantic classes, e.g. indoor/outdoor or citycsape/landscape, use features representing predefined blocks rather than homogenous regions.

Therefore the author firmly believes that to fully asses the role of image segmentation in the content based retrieval problem further research effort should be devoted to investigate new ways of utilizing the results of image

segmentation in CBIR systems.

One possibility would be a departure from representing the results of segmentation as a single segmentation mask for each image with a view to utilizing more flexible representations able to capture the structure of the scene and allowing utilization of the contextual information, e.g. the *Binary Partition Tree* (BPT) advocated in [107]. Employment of such structures, or in more general employment of visual features extracted based on multiple partitioning hypothesis, to detection of high-level concepts or scene classification should also be considered.

Also, new flexible querying models allowing both: image-based and object-based queries by example and also new ways of establishing similarity between the query and the target suitable for such models should be investigated.

Finally, detection and extraction of new classes of semantic entities potentially important in many content based applications should be investigated. For example, currently it is widely recognized that detection and recognition of faces present in the visual content can provide very valuable information for content-based retrieval. Although many powerful approaches to face detection have been proposed [265, 266] the problem of face recognition in general content remains to be solved [252, 79, 115]. However, additional information potentially valuable for content based functionalities in many applications could be provided by detection and extraction of person's head & shoulders. The presence of the shoulders should be useful for validation of the face detection. Also, the extracted head & shoulder should provide more reliable evidence for person's identification than face in many applications, e.g. identification of the player's team in close-ups present in most sports videos.

8.3.2 Shape Matching

Clearly shape analysis and shape matching in particular are very important in the context of CBIR however it is also a challenging task and many problems from this area remain to be solved.

Shape similarity depends highly on the context, e.g. different applications may require employment of different notions of shape similarity. Until now, most of the shape matching techniques were proposed in the context of a very specific application and tested on relatively homogenous collections. Incorporation of contextual information in the shape similarity measures is currently in its infancy [3] and certainly deserves greater attention. Also combining the output of a number of methods to establishing similarity between shapes could open up interesting possibilities, e.g. increased retrieval performance or improved efficiency by using coarse to fine strategies.

Although establishing shape similarity is a challenging task, it can be performed with reasonable success, as has been shown in this thesis. Currently, the main

problem in the way to fully exploit shape information in CBIR systems appears to be the limited ability to extract the semantically meaningful entities. One means of extracting such objects is by utilization of shape matching techniques which can match a template shape directly with an object present in an unseen image or by employing image segmentation driven by a shape models, e.g. the PDM models discussed in chapter 7. Such techniques have been successfully applied to a very specific problems and their application to more general content certainly requires more research. The author acknowledges that such techniques, unlike the methods investigated in this thesis, should employ shape templates consisting of multiple disconnected closed and open contours or use models not only of the outline of the object but also of its internal parts in order to be applicable to a wider class of problems.

Finally, the preliminary investigation of utilization of the contour matching approach proposed in this thesis to word spotting in handwritten documents carried out in chapter 6 provided very encouraging results and should be further investigated. Particulary interesting is the possibility of matching the trajectory of the pen instead of the the external outlines of the words. This should be easily achieved in the case of on-line handwriting recognition where the trajectory of the pen is readily available but should be also possible in off-line recognition scenarios where the pen movements, or at least their approximation consistent across many instances of the same word, could be recreated from the word's images.

8.4 Final Word

The author believes that all these challenges outlined above will provide ample opportunity for interesting research for many years to come. Extending the state of the art in these areas will significantly enhance existing CBIR applications whilst enabling other, as yet unknown, future applications.

APPENDIX A

THE HUMAN VISUAL SYSTEM AND OBJECT RECOGNITION

THIS appendix briefly discusses selected theories from the field of *Cognitive Psychology* regarding human perception and object recognition which influenced Computer Vision and Image Understanding and in consequence also systems for indexing and retrieval of visual content.

Although none of the abstract models presented in this section is explicitly implemented in this thesis several theories, concepts and suggestions proposed in the field of Cognitive Psychology provide context and justification for the choices made by the author and can serve as explanation of the results obtained throughout the thesis.

A.1 Gestalt Principles of Visual Organization Of Perception

Gestalt is a psychology term which means "unified whole" and refers to theories of visual perception developed in the 1920s by German psychologists (Max Wertheimer (1880 – 1943), Wolfgang Köher (1887 – 1967) and Kurt Koffka (1886 – 1941)) attempting to describe how humans tend to organize visual elements into groups or unified wholes. The Gestalt psychologists outlined several principles of perceptual organization on the basis of their use in identifying ambiguous patterns. They advocated that humans, when confronted by any abstract or symbolic visual image, tend to separate a dominant shape ("figure") from "background" (or "ground") using several principles of identifying ambiguous patterns which include: (i) Proximity (things closer together are associated), (ii) Similarity (things sharing visual characteristics such as shape, size, colour, texture, value or orientation are associated), (iii) Continuity (preference for continuous figures), (iv) Closure (interpretations which produce "closed" rather

than "open" figures are favored), (v) Smallness (smaller of two overlapping areas tend to be seen as figures while the larger is regarded as background), (vi) Surroundedness (areas seen as surrounded by others tend to be perceived as figures), (vii) Symmetry (symmetrical areas tend to be seen as figures). The above principles serve the overarching principle of pragnänz, which states that the simplest and most stable interpretations are favored.

The original Gestalt principles were formulated based on two dimensional abstract or symbolic pictures and their application to complex real world scenes is rather difficult. Moreover, the Gestalt laws simply provide descriptions of phenomena, not an explanation of them. Nevertheless, the pioneering suggestion that humans may be predisposed towards interpreting ambiguous images in one way rather than another have important implications for designing modern visual systems. Mimicking such natural predispositions by an automatic bottom-up approach to scene segmentation could help in dealing with the inherent ambiguities of such approach. This suggestion provides the basis for investigation of the usefulness of the so-called syntactic features in the bottom-up segmentation process.

A.2 Structural-Description vs. Image-Based Models

A.2.1 Reconstruction of the Outside World from the Retinal Image

Computational models of recognition proposed in the past were based largely on the work of Marr (1982) [267] who popularized a solution that the goal of vision is to reconstruct the 3D scene by producing a conclusive representation of the outside world from the retinal image. He suggested that human visual processing passed through a series of stages, each corresponding to a different representation: (i) the retinal image, (ii) a "primal sketch" representing explicitly lines, edges and junctions, (iii) a "$2\frac{1}{2}$D sketch" including information about surface slant and depth, (iv) a "3D model representation" of objects built up hierarchically from simple primitives such as cylinders and (v) a "space frame centred on the observer" where the 3D models reside.

Based on Marr's type of bottom-up computational approach Biederman (1987) proposed a theory known as *Recognition by Components* (RBC) [268, 183] according to which objects are represented in terms of their simple parts, the 3D volumetric primitives called *geons* (geometric ions), e.g. bricks, bent cylinders, and wedges. Geons are defined by viewpoint-invariant properties of 2D images [183] such as smooth continuation, co-termination, parallelism and symmetry. Since these properties, referred often as "non-accidental", represent properties of the real world they are sufficient to define 3D component representations (no need for $2\frac{1}{2}$D sketch). Objects are defined as relationships between geons. In other

words, according to Biederman, object recognition only requires detecting edges, arranging them into geons based on nonaccidental properties and object classification based on geon's types and relationships. Other sources of information (e.g. colour, texture etc.) are relatively unimportant [268]. Biederman argues that geon representation is necessary since representing objects directly in terms of local image features would require a different representation for each slightly altered orientation of the object or for each arrangement of occluding contour. Moreover, since accurate geon activation requires only categorical classification of edge characteristics the RBC theory explains well human ability to rapidly recognize objects.

Although the above paradigm is attractive because it decomposes the problem of image analysis into independent and manageable components the reconstruction is difficult to achieve computationally. For example, the problems of edge detection and contour extraction are under-constrained due to inherent ambiguity present in real scenes. Therefore, hierarchical reconstruction may lead to irreversible propagation of errors from one processing stage to another. Moreover, many researchers (e.g. Tarr & Bülthoff, 1998) suggest that the RBC theory only explains well coarse discriminations between object types (e.g. a cup vs. a car) but not within an object type (e.g. one car vs. another).

Despite the difficulties with computational implementation of bottom-up type of processing Biedermans' work provides extremely strong empirical evidence suggesting that, in the case of humans, priming is largely visual, rather than conceptual or lexical [268, 183] therefore de-emphasizing the role of context. Biederman argues that extraordinary speed of object recognition in the absence of any context disagrees with a central role for top-down information. This is a potentially important implication in designing CBIR systems where off-line bottom-up processing is often unavoidable to ensure fast responses to user queries. In other words, Biedermans' work provides hope that if components performing off-line processing in CBIR systems could imitate, to a certain degree, human ability to resolve inherent ambiguities present in complex real world scenes, the result of such processing could efficiently serve indexing and retrieval of visual content. This provides justification of the research carried out in chapter 3 investigating alternative sources of information for improving the perceptual quality of the output of bottom-up segmentation.

A.2.2 Exemplar-Based Multiple-Views Mechanism

An alternative approach to the RBC paradigm was suggested by Tarr & Bülthoff who advocate an exemplar-based multiple-views mechanism as an important component of both exemplar-specific and categorical recognition [269]. In the multiple-views approach object representations are collections of views that depict the appearance of objects from specific viewpoints. They argue that

image-based models containing multiple views explain well recent behavioral results, e.g. recognition performance is viewpoint dependent because both recognition time and accuracy vary with the degree of mismatch between the percept and target view. A variety of different mechanisms for generalizing from unfamiliar to familiar views have been proposed, including mental rotation (Tarr & Pinker, 1989), view interpolation (Poggio & Edelman, 1990), and linear combinations of views (Ullman & Basri, 1991). Tarr & Bülthoff also suggest that there must be some mechanism for measuring the perceptual similarity (within some domain) between an input image and known objects and between an input image and known views of that same object.

Tarr and his collaborators also suggest that a substantial part of the information present in an object view is carried by outline shape [164, 165] and often advocate the need for the computation of silhouette similarity. [164] demonstrates the utility of silhouette representations for a variety of visual tasks, ranging from basic-level categorization to finding the best view of an object. They argue that using only silhouettes or boundary contour information a large of exemplars can be separated into stable perceptual categories corresponding closely to human perceivers. The above suggestion, coming from cognitive science, provides very strong motivation for utilization of silhouettes for classification and recognition in computer vision and provides justification of the research carried out by the author in several chapters of this thesis.

A.2.3 Which Model Explains All Facts of Human Visual Perception?

Currently many researchers, including advocates of both approaches presented above, Biederman [183] and Tarr [269], seem to recognize that no existing single theory can explain all facets of human visual perception. In fact, there seems be general consensus that several approaches (including RBC and exemplar-based multiple-views mechanism) are not mutually exclusive but simply address different aspects of object recognition under different conditions determined for example by a particular task, context, familiarity, and visual similarity.

Indeed, the presence of several recognition mechanisms may somehow explain difficulties in developing a single "universal" computational model able to perform general object detection and recognition. So far, computer vision is able to successfully address only very specific problems.

APPENDIX B

IMAGE COLLECTION USED IN CHAPTER 3

THIS appendix presents the collection of images and ground-truth segmentation developed by the author in collaboration with a colleague working in another university. The collection contains 100 images from the Corel gallery [132] and 20 images from various sources such as keyframes from well known MPEG-4 test sequences, digital camera, and mobile phone camera segmented by four annotators.

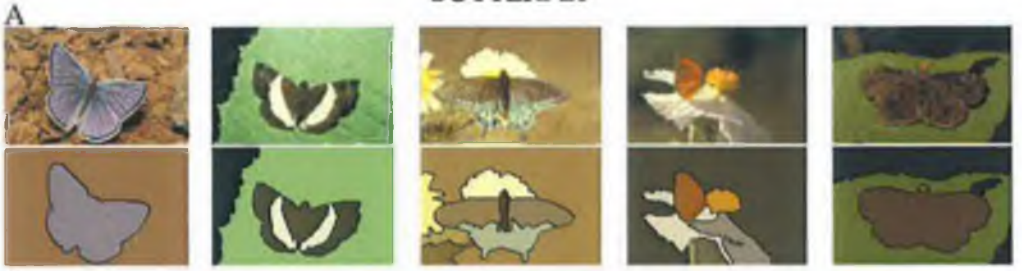
BEACH

A

B

BUTTERFLY

A



B

CAR

A

B

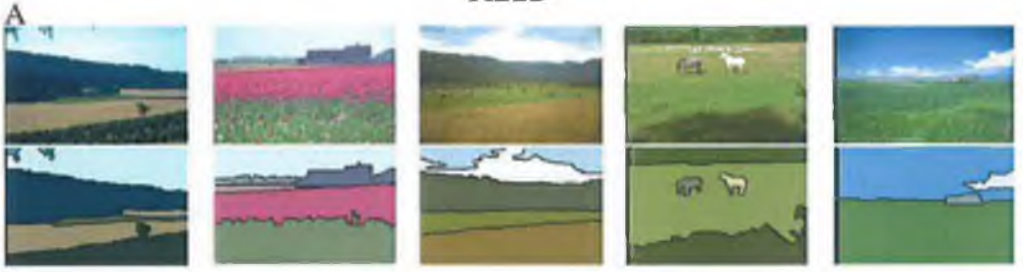
EAGLE

A

B

FIELD

A



B

HORSE

A

B

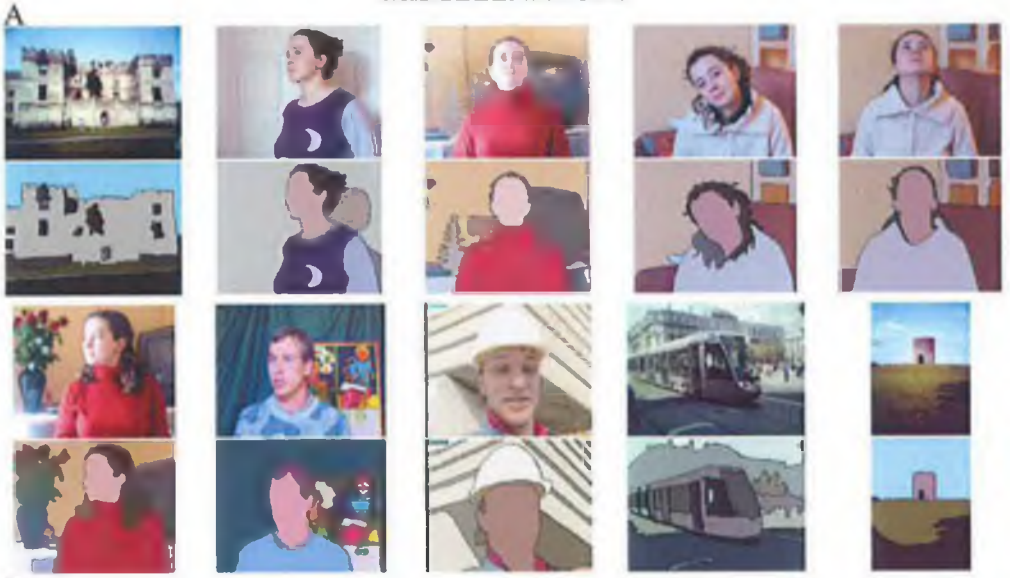
JET

A

B

MISCELLANEOUS

A



B

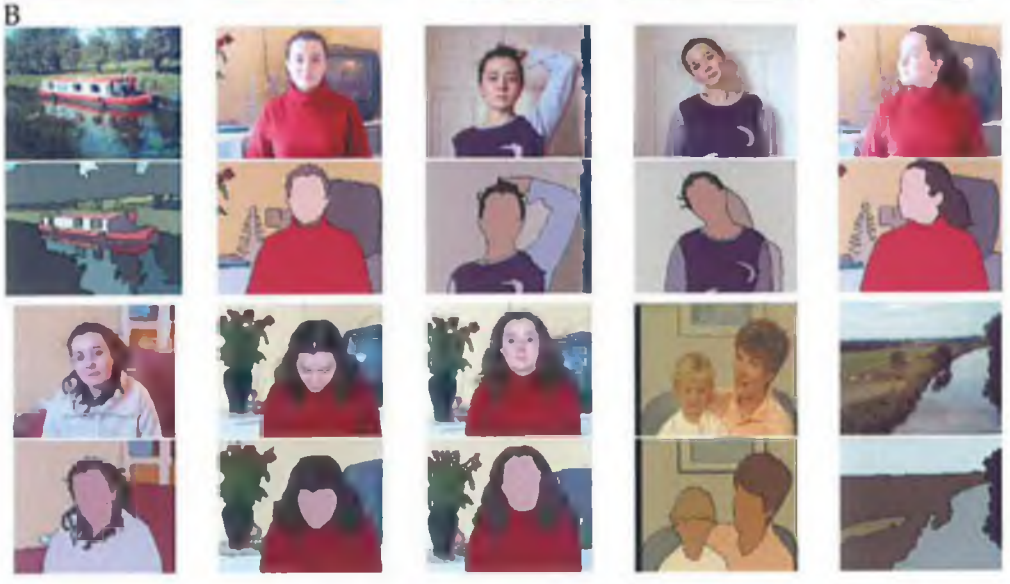


Figure B.1: Image collection used in chapter 3.

BIBLIOGRAPHY

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proc. the 7th Int'l WWW Conf. (WWW'98). Brisbane, Australia*, 1998.

[2] J. P. Eakins and M. E. Graham, "Content-based image retrieval: A report to the JISC technology application programme," Institute for Image Data Research, University of Northumbria at Newcastle, UK, `www.unn.ac.uk/iidr/report.html`, Technical Report, 1999.

[3] F. A. Cheikh, "MUVIS: A system for content-based image retrieval," Ph.D. dissertation, Tampere University of Technology, 2004.

[4] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Language*. New York: John Wiley & Sons Ltd., ISBN: 0-471-48678-7, 2002.

[5] J. M. Martínez, "MPEG-7 overview," ISO/IEC JTC1/SC29/WG11, Tech. Rep. N6828, Mar. 2003.

[6] N. O'Connor, E. Cooke, H. Le Borgne, M. Blighe, and T. Adamek, "The AceToolbox: Low-level audiovisual feature extraction for retrieval and classification," in *Proc. 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, U.K.*, Dec. 2005.

[7] A. Smeaton, H. Lee, and K. McDonald, "Experiences of creating four video library collections with the Físchlár system," *Journal of Digital Libraries: Special Issue on Digital Libraries as Experienced by the Editors of the Journal*, vol. 4, no. 1, pp. 42–44, 2004.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. IEEE Int'l Conf. on Computer Vision*, 2002, pp. 97–112.

[9] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[10] S. Sav, H. Lee, A. F. Smeaton, and N. E. O'Connor, "Using segmented objects in ostensive video shot retrieval," in *Proc. 3rd Int'l Workshop on Adaptive Multimedia Retrieval*, July 2005.

[11] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkhani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer Magazine*, vol. 28, no. 9, pp. 23–32, Sept. 1995.

[12] S. Lambert, E. de Leau, and L. Vuurpijl, "Using pen-based outlines for object-based annotation and image-based queries," in *Proc. 3rd Int'l Conf. on Visual Information and Information Systems (VISUAL'99), Amsterdam, The Netherlands*, ser. LNCS 1614, D. Huijsmans and A. Smeulders, Eds. Springer, June 1999, pp. 585–592.

[13] F. Mokhtarian and A. K. Mackworth, "A theory of multiscale, curvature-based shape representation for planar curves," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 8, pp. 789–805, Aug. 1992.

[14] F. Mokhtarian, "Silhouette-based isolated object recognition through curvature scale space," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 17, no. 5, pp. 539–544, May 1995.

[15] F. Mokhtarian, S. Abbasi, and J. Kittler, "Robust and afficient shape indexing through curvature scale space," in *British Machine Vision Conf. (BMVC'96)*, 1996, pp. 53–62.

[16] M. Bober, "MPEG-7 visual shape descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 716–719, 2001.

[17] F. Mokhtarian and M. Bober, *Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standarization*, ser. ISBN 1-4020-1233-0. Kluwer Academic Publishers, 2003, vol. 25.

[18] C. F. Bennstrom and J. R. Casas, "Binary-partition-tree creation using a quasi-inclusion criterion," in *Proc. 8th Int'l Conf. on Information Visualization (IV'04), IEEE Computer Society Press, London, UK*, 2004.

[19] O. Morris, M. Lee, and A. Constantinides, "Graph theory for image analysis: an approach based on the shortest spanning tree," in *IEE Proceedings*, vol. 133, Apr. 1986, pp. 146–152.

[20] A. Dempster, "Upper and lower probabilities induced by multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.

[21] G. Shafer, "A mathematical theory of evidence," *Princetown Univ. Press, Princetown New Jersey*, 1976.

[22] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[23] P. Smets, "The normative representation of quantified beliefs by belief functions," *Artificial Intelligence*, vol. 92, pp. 229–242, 1997.

[24] R. Manmatha and N. Srimal, "Scale space technique for word segmentation in handwritten manuscripts," in *Proc. 2nd Int'l Conf. on Scale-Space Theories in Computer Vision, Corfu, Greece*, Sept. 1999, pp. 22–33.

[25] V. Lavrenko, T. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proc. Document Image Analysis for Libraries (DIAL'04)*, 2004, pp. 278–287.

[26] MPEG Requirements Group, "MPEG-7 context and objectives," Doc. ISO/MPEC, MPEG Atlantic City Meeting, Tech. Rep. 2460, Oct. 1998.

[27] D. Santa-Cruz and T. Ebrahimi, "An analytical study of JPEG 2000 functionalities," in *Proc. IEEE Int'l Conf. on Image Processing*, vol. 2, 2000, pp. 49–52.

[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[29] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Color- and texture-based image segmentation using EM and its application to image querying and classification," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 8, pp. 1026–1037, Aug. 2002.

[30] V. Mezaris, I.Kompatsiaris, and M.G.Strintzis, "Still image segmentation tools for object-based multimedia applications," *Int'l Journal of Pattern Recognition and Artificial Intell.*, vol. 18, no. 4, pp. 701–725, June 2004.

[31] J. Fauqueur and N. Boujemaa, "Region-based image retrieval: Fast coarse segmentation and fine color description," *Journal of Visual Languages and Computing, special issue on Visual Information Systems*, vol. 15, pp. 69–95, 2004.

[32] N. E. O'Connor, "Video object segmentation for future multimedia applications," Ph.D. dissertation, Dublin City University, School of Electronic Engineering, 1998.

[33] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 5, 2002.

[34] S. Sun, D. R. Haynor, and Y. Kim, "Semiautomatic video object segmentation using vsnakes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 75–82, Jan. 2003.

[35] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[36] W. Skarbek and A. Koschan, "Colour image segmentation - a survey," Technical University of Berlin, Department of Computer Science, Germany, Technical Report 94-32, Oct. 1994.

[37] Special Issue, "Special issue on image and video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 19–31, Sept. 1998.

[38] P. Salembier and F. Marqués, "Region-based representations of image and video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1149–1167, Dec. 1999.

[39] H. Cheng, X. H. Jiang, Y. Sun, and J. L. Wang, "Color image segmentation: Advances & prospects," *Pattern Recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[40] S. Horowitz and T. Pavlidis, "Picture segmentation by a tree traversal algorithm," *J. Assoc. Compt. Math.*, vol. 23, no. 2, pp. 368–388, 1976.

[41] O. Salerno, M. Pardas, V. Vilaplana, and F. Marqués, "Object recognition based on binary partition trees," in *Proc. Int'l Conf. on Image Processing (ICIP '04)*, vol. 2, Oct. 2004, pp. 929–932.

[42] G. Aggarwal, S. Ghosal, and P. Dubey, "Efficient query modification for image retrieval," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'00)*, vol. II, June 2000, pp. 255–261.

[43] I. Kompatsiaris, E. Triantafillou, and M. G. Strintzis, "Region-based colour image indexing and retrieval," in *Proc. Int'l Conf. on Image Processing (ICIP'01), Thessaloniki, Greece*, vol. 1, 2001, pp. 658–661.

[44] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Region-based image retrieval using an object ontology and relevance feedback," *EURASIP Journal on Applied Signal Processing*, no. 6, pp. 886–901, June 2004.

[45] F. Souvannavong, B. Merialdo, and B. Huet, "Region-based video content indexing and retrieval," in *Proc. 4th Int'l Workshop on Content-Based Multimedia Indexing (CBMI'05)*, 2005.

[46] W.-Y. Ma and B. S. Manjunath, "NeTra: A toolbox for navigating large image databases," *Multimedia Systems*, vol. 7, no. 3, pp. 184–198, 1999.

[47] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An effective region-based image retrieval framework," in *Proc. the ACM Int'l Conf. on Multimedia*, 2002.

[48] R. C. Veltkamp and M. Tanase, "Content-based image retrieval systems: A survey," Department of Computing Science, Utrecht University, Technical Report UU-CS-2000-34, Tech. Rep., Oct. 2000.

[49] IBM, "QBIC project," `http://wwwqbic.almaden.ibm.com`.

[50] R. Barber, W. Niblack, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: Querying images by content using colour, texture, and shape," in *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, Feb. 1993, pp. 173–187.

[51] A. Del Bimbo, M. Mugnaini, P. Pala, and F. Turco, "Picasso: Visual querying by color perceptive regions," in *Proc. 2nd Int'l Conf. on Visual Information Systems, San Diego*, Dec. 1997, pp. 125–131.

[52] P. Pala and S. Santini, "Image retrieval by shape and texture," *Pattern Recognition*, vol. 32, no. 3, pp. 517–527, 1999.

[53] W.-Y. Ma, "NeTra: A toolbox for navigating large image databases," Ph.D. dissertation, Dept. of Electrical and Computer Engineering, University of California at Santa Barbara, June 1997.

[54] W. Ma and B. Manjunath, "Edge flow: a framework of boundary detection and image segmentation," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition, Puerto Rico*, June 1997.

[55] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: Colour image segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'97)*, June 1997, pp. 750–755.

[56] E. Choi and P. Hall, "Data sharpening as a prelude to density estimation," *Biometrika*, vol. 86, 1999.

[57] M. Markkula and E. Sormunen, "End-user searching challenges indexing practices in the digital newspaper photo archive," *Information Retrieval*, vol. 1, pp. 259–285, 2000.

[58] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, "The effects of segmentation and feature choice in a translation model of object recognition," in *Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR'03)*, 2003.

[59] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for filtering, segmentation, and information retrieval," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP'98), Chicago (IL), USA*, Oct. 1998.

[60] J. Malik, S. Belongie, J. Shi, and T. Leung, "Textons, contours and regions: Cue integration in image segmentation," in *Proc. IEEE Int'l Conf. on Computer Vision, Corfu, Greece*, Sept. 1999.

[61] J. Garding and T. Lindeberg, "Direct computation of shape cues using scale-adapted spatial derivative operators," *Int'l J. Computer Vision*, vol. 17, no. 2, pp. 163–191, Feb. 1996.

[62] Y. Zeng and A. G. Constantinides, "Perceptual saliency weighted segmentation algorithm," in *Proc. 7th Int'l Conf. on An Image Processing And Its Applications*, July 1999.

[63] Y. Zeng, "Perceptual segmentation algorithm and its application to image coding," in *Proc. Int'l Conf. on Image Processing*, 1999, pp. 820–824.

[64] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int'l Journal of Computer Vision*, vol. 1, pp. 312–331, 1988.

[65] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing and bayes/MDL for multiband image segmentation," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 18, no. 9, pp. 884–900, 1996.

[66] M. Rousson, T. Brox, and R. Deriche, "Active unsupervised texture segmentation on a diffusion-based space," in *Proc. Int'l Conf. on Computer Vision & Pattern Recognition)*, 2003.

[67] "Special section on perceptual organization in computer vision," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 25, no. 4, Apr. 2003.

[68] T. Adamek, N. E. O'Connor, and N. Murphy, "Region-based segmentation of images using syntactic visual features," in *Proc. 6th Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05), Montreux, Switzerland*, Apr. 2005.

[69] E. Engbers and A. Smeulders, "Design considerations for generic grouping in vision," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 25, no. 4, pp. 445–457, Apr. 2003.

[70] T. F. Cootes and C. J. Taylor, "Active shape models - smart snakes," in *Proc. British Machine Vision Conf. (BMVC'92), Springer Verlag*, 1992, p. 266.

[71] S. Sclaroff and L. Liu, "Deformable shape detection and description via model-based region grouping," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 5, 2001.

[72] M. E. Leventon, W. L. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'00)*, vol. I, 2000, pp. 316–323.

[73] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, "Curve evolution technique for image segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, vol. I, 2001, pp. 463–468.

[74] Y. Chen, S. Thiruvenkadam, H. Tagare, F. Huang, and D. Wilson, "On the incorporation of shape priors into geometric active contours," in *Proc. (VLSM'01)*, 2001, pp. 145–152.

[75] D. Cremers, T. Kohlberger, and C. Schnorr, "Nonlinear shape statistics in mumford-shah based segmentation," in *Proc. European Conf. on Computer Vision (ECCV'02)*, vol. II, 2002, pp. 93–108.

[76] M. Rousson and N. Paragios, "Shape priors for level set representations," in *Proc. European Conf. on Computer Vision (ECCV'02, LNCS 2351)*, 2002, pp. 78–92.

[77] M. B. Stegmann, "Active appearance models: Theory, extensions and cases," Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, `http://www.imm.dtu.dk/~aam/`, 2000.

[78] T. Cootes and C.J.Taylor, "Statistical models of appearance for computer vision," University of Manchester, `http://www.isbe.man.ac.uk/~bim`, Tech. Rep., Mar. 2004.

[79] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[80] B. Marcotegui, P. Correia, F. Marqués, R. Mech, R. Rosa, M. Wollborn, and F. Zanoguera, "A video object generation tool allowing friendly user interaction," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP'99), Kobe, Japan*, 1999.

[81] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 16, no. 6, pp. 641–647, June 1994.

[82] E. Chalom and V. Bove, "Segmentation of an image seguence using multi-dimensional image attributes," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP'96), Lausanne, Switzerland*, vol. 2, Sept. 1996, pp. 525–528.

[83] J. Cichosz and F. Meyer, "Morphological multiscale image segmentation," in *Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'97), Louvain-la-Neuve, Belgium*, June 1997, pp. 161–166.

[84] C. Gu and M.-C. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 572–584, Sept. 1998.

[85] S. Cooray, N. E. O'Connor, S. Marlow, N. Murphy, and T. Curran, "Semi-automatic video object segmentation using recursive shortest spanning tree and binary partition tree," in *Proc. 3rd Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'01), Tampere, Finland*, 2001.

[86] N. Otsu, "A threshold selection method from grey level histograms," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 8, pp. 62–66, 1978.

[87] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," *IEEE Trans. on Image Processing*, vol. 7, no. 6, pp. 918–920, 1998.

[88] J. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkely Symp. on Math. Stat. and Prob.*, vol. 1, 1967, pp. 281–296.

[89] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[90] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[91] M. Herbin, N. Bonnet, and P. Vautrot, "A clustering method based on the estimation of the probability density function and on the skeleton by influence zones," *Pattern Recognition Letters*, vol. 17, pp. 1141–1150, 1996.

[92] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 13, no. 6, pp. 583–598, June 1991.

[93] D. Wang, "A multi-scale gradient algorithm for image segmentation using watersheds," *Pattern Recognition*, vol. 30, no. 12, pp. 2043–2052, 1997.

[94] T. Vlachos and A. G. Constantinides, "A graph-theoretic approach to color image segmentation and contour classification," in *Proc. 4th Int'l Conf. Image Processing and Its Applications, Maastrict, The Netherlands*, Apr. 1992.

[95] S. H. Kwok, A. G. Constantinides, and W.-C. Siu, "An efficient recursive shortest spanning tree algorithm using linking properties," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 852–863, June 2004.

[96] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions and the bayesian restoration of images," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 6, no. 6, pp. 721–741, 1984.

[97] Z. Tu and S. Zhu, "Image segmentation by data-driven markov chain monte carlo," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 5, pp. 657–673, 2002.

[98] E. D. Gelasca, T. Ebrahimi, M. C. Q. Farias, and S. K. Mitra, "Impact of topology changes in video segmentation evaluation," in *Proc. 5th Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS2004), Lisboa, Portugal*, Apr. 2004.

[99] W. Perkins, "Area segmentation of images using edge points," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 2, no. 1, pp. 8–15, 1980.

[100] J. Prager, "Extracting and labelling boundary segments in natural scenes," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 2, no. 1, pp. 16–27, 1980.

[101] N. Papamarkos, C. Strouthopoulos, and I. Andreadis, "Multithresholding of colour and gray-level images through a neural network technique," *Image and Vision Computing*, vol. 18, pp. 213–222, 2000.

[102] J. Guo, J. Kim, and C. Kuo, "Fast and accurate moving object extraction technique for MPEG-4 object based video coding," in *Proc. SPIE Visual Comm. and Image Processing*, 1999, pp. 1210–1221.

[103] L.Liu and S. Sclaroff, "Deformable shape detection and description via model-based region grouping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'99)*.

[104] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 26, no. 11, pp. 1452–1458, Nov. 2004.

[105] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. London, UK: Chapman and Hall, 1995.

[106] S. H. Kwok and A. G. Constantinides, "A fast recursive shortest spanning tree for image segmentation and edge detection," *IEEE Trans. Image Processing*, vol. 6, no. 2, pp. 328–332, Feb. 1997.

[107] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp. 561–576, Apr. 2000.

[108] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Proc. Int'l Conf. on Computer Vision, Vancouver, Canada*, vol. I, July 2001.

[109] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: A survey," *Medical Image Analysis*, vol. 1, no. 2, pp. 91–108, 1996.

[110] L. Cohen, "On active contour models and balloons," *CVGIP:IU*, vol. 53, no. 2, 1991.

[111] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *Int'l Journal of Computer Vision (IJCV)*, vol. 22, no. 1, pp. 61–79, 1997.

[112] J. Sethian, *Level Set Methods.* Cambridge University Press, 1996.

[113] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational-problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[114] N. Paragios and R. Deriche, "Geodesic active contours and level-set methods for supervised texture segmentation," *Int'l Journal of Computer Vision*, 2002.

[115] A. L. Yuille, D. S. Cohen, and P. Hallinan, "Feature extraction from faces using deformable templates," *Int'l Journal of Computer Vision*, vol. 8, no. 2, pp. 99–112, 1992.

[116] A. P. Pentland and S. Sclaroff, "Closed-form solutions for physically based modelling and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 7, pp. 715–729, 1991.

[117] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 95.

[118] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburt, R. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. on Computers*, vol. 42, pp. 300–311, 1993.

[119] K. V. Mardia, J. T. Kent, and A. N. Walder, "Statistical shape models in image analysis," in *Proc. Computer Science and Statistics: 23rd INTERFACE Symposium, E. Keramidas, editor, Interface Foundation, Fairfax Station)*, 1991, pp. 550–557.

[120] U. Grenander and M. Miller, "Representations of knowledge in complex systems," *Journal of the Royal Statistical Society*, vol. 56, pp. 249–603, 1993.

[121] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[122] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, Aug. 1996.

[123] P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proc. Int'l Conf. on Image Processing (ICIP'00), Vancouver, Canada*, Sept. 2000, pp. 308–311.

[124] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Recommendation BT.500-7, 1995.

[125] ITU-T, "Subjective video quality assessment methods for multimedia applications," Recommendation P.910, Aug. 1996.

[126] P. Correia and F. Pereira, "Standalone objective evaluation of segmentation quality," in *Proc. 3rd Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'01), Tampere, Finland*, May 2001.

[127] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Still image objective segmentation evaluation using ground truth," in *Proc. 5th COST 276 Workshop (2003), Berlin*, 2003, pp. 9–14.

[128] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video object generation algorithms," in *Doc. ISO/IEC/JTC1/SC29/WG11 M3448, 43rd MPEG Meeting, Tokyo, Japan*, Mar. 1998.

[129] P. Villegas, X. Marichal, and A. Salcedo, "Objective evaluation of segmentation masks in video sequences," in *Proc. Workshop on Image Analysis For Multimedia Interactive Services (WIAMIS'99), Berlin*, May 1999, pp. 85–88.

[130] P. Correia and F. Pereira, "Estimation of video object's relevance," in *Proc. European Conf. on Signal Processing (EUSIPCO'2000), Tampere, Finland*, Sept. 2000.

[131] MIT Vision Texture (VisTex) database, `http://www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html`.

[132] Corel stock photo library, Corel Corp., Ontario, Canada.

[133] M. Wirth, M. Fraschini, M. Masek, and M. Bruynooghe, "Performance evaluation in image processing," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 45 742, 3 pages, 2006.

[134] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervised performance evaluation of image segmentation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 96 306, 12 pages, 2006.

[135] P. Correia and F. Pereira, "Video object relevance metrics for overall segmentation quality evaluation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 82 195, 11 pages, 2006.

[136] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance measures for image segmentation evaluation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 35 909, 10 pages, 2006.

[137] R. Piroddi and T. Vlachos, "A method for single-stimulus quality assessment of segmented video," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 39 482, 22 pages, 2006.

[138] R. Usamentiaga, D. F. Garcia, C. Lopez, and D. Gonzalez, "A method for assessment of segmentation success considering uncertainty in the edge positions," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 21 746, 12 pages, 2006.

[139] T. Y. Lui and E. Izquierdo, "Scalable object-based image retrieval," in *Proc. IEEE Int'l Conf. on Image Processing (ICIP'03), Barcelona, Spain,* Sept. 2003.

[140] A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image sequence analysis for emerging interactive multimedia services - the European COST 211 Framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 7, pp. 802–813, Nov. 1998.

[141] E. Tuncel and L. Onural, "Utilization of the recursive shortest spanning tree algorithm for video-object segmentation by 2-d affine motion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 5, pp. 776–781, Aug. 2000.

[142] J. H. Ward, "Hierarchical grouping to optimize an objective function," *American Stat. Assoc.*, vol. 58, pp. 236–245, 1963.

[143] L. Garrido, P. Salembier, and D. Garcia, "Extansive operators in partition lattices for image sequence analysis," *EURASIP Signal Processing*, vol. 66, no. 2, pp. 157–180, Apr. 1998.

[144] T. Brox, D. Farin, and P. de With, "Multi-stage region merging for image segmentation," in *Proc. 22nd Symposium on Information and Communication Theory in the Benelux, Enschede, The Netherlands,* May 2001.

[145] J. Hafner, H. Sawhnay, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 17, no. 7, pp. 729–736, July 1995.

[146] Y. Chen and H. Sundaram, "Estimating the complexity of 2D shapes," in *Proc. Int'l Workshop on Multimedia Signal Processing (MMSP'05),* May 2005.

[147] S. Loncaric, "A survey of shape analysis techniques," *Pattern Recognition*, vol. 31, no. 5, pp. 983–1001, 1998.

[148] H. Moravec, "Towards authomatic visual obstacle avoidance," in *Proc. 5th Int'l Joint Conf. on Artificial Inteligence, Caregie-Mellon University, Pittsburg, PA,* Aug. 1977.

[149] J. Bezdek, "Fuzziness vs. probability - the n-th round," *IEEE Trans. on Fuzzy Systems*, vol. 2, no. 1, pp. 1–42, 1994.

[150] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.

[151] D. Dubois, J. Lang, and H. Prade, "Automated reasoning using possibilistic logic: Semantics, belief revision, and varaible certainty weights," *IEEE Trans. on Knowledge and Data Engineering*, vol. 6, pp. 64–71, 1994.

[152] Z. Hammal, A. Caplier, and M. Rombaut, "A fusion process based on belief theory for classification of facial basic emotions," in *Proc. 8th Int'l Conf. on Information Fusion, Philadelphia, USA,* July 2005.

[153] A. Al-Ani and M. Deriche, "A new technique for combining multiple classifiers using the dempster-shafer theory of evidence," *Journal of Artificial Intelligence Research*, vol. 17, pp. 333–361, 2002.

[154] V. Girondel, A. Caplier, and L. Bonnaud, "A belief theory-based static posture recognition system for real-time video surveillance applications," in *Proc. IEEE Int'l Conf. on Advanced Video and Signal based Surveillance (AVSS'05), Como, Italy*, Sept. 2005.

[155] P. Smets, E. Mamdami, D. Dubois, and H. Prade, *Non-Standard Logics for Automated Reasoning*, ser. ISBN 0126495203. Academic Press, Harcourt Brace Jovanovich Publisher, 1988.

[156] L. Zadeh, "A mathematical theory of evidence (book review) - a critique of the normalization factor in dempster's rule of combination," *AI Magazine*, vol. 5, no. 3, pp. 81–83, 1984.

[157] M. E. Cattaneo, "Combining belief functions issued from dependent sources," in *Proc. 3rd Int'l Symposium on Imprecise Probabilities and Their Applications (ISIPTA'03), Lugano, Switzerland*, 2003.

[158] P. Rosin, "Unimodal thresholding," *Pattern Recognition*, vol. 34, no. 11, pp. 2083–2096, Nov. 2001.

[159] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Computer, Graphics and Image Processing*, vol. 1, pp. 244–256, 1972.

[160] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. Intl Conf. Computer Vision*, 2001.

[161] "Berkeley segmentation and boundary detection benchmark and dataset," `http://www.cs.berkeley.edu/projects/vision/grouping/segbench`, 2003.

[162] P. Correia and F. Pereira, "Classification of video segmentation application scenarios," *IEEE Trans. Circuits Syst. Video Technol., special issue on Audio and Video Analysis for Multimedia Interactive Services*, vol. 14, no. 5, pp. 735–741, May 2004.

[163] L. F. Costa and R. M. Cesar, *Shape Analysis and Classification, Theory and Practise*, ser. ISBN 0-8493-3493-4. CRC Press LLC, 2000.

[164] F. Cutzu and M. J. Tarr, "The representation of three-dimensional object similarity in human vision," in *SPIE Proc. from Electronic Imaging: Human Vision and Electronic Imaging II*, 1997, pp. 460–471.

[165] W. G. Hayward, "Effects of outline shape in object recognition," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 2, pp. 427–440, 1998.

[166] F. Mokhtarian and A. K. Mackworth, "Scale-based description and recognition of planar curves and two dimensional shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, no. 1, pp. 34–43, Jan. 1986.

[167] E. G. M. Petrakis, A. Diplaros, and E. Millos, "Matching and retrieval of distorted and occluded shapes using dynamic programing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1501–1516, 2002.

[168] R. Basri, L. Costa, D. Geiger, and D. Jacobs, "Determining the similarity of deformable shapes," in *Proc. IEEE Workshop on Physics Based Modeling in Computer Vision*, 1995, pp. 135–143.

[169] H. Kauppinen, T. Seppanen, and M. Pietikainen, "An experimental comparison of autoregressive and fourier-based descriptors in 2D shape classification," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 17, pp. 201–207, 1995.

[170] Y. Gdalyahu and D. Weinshall, "Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 12, pp. 1312–1328, Dec. 1999.

[171] R. Veltkamp, "Shape matching: Similarity measures and algorithms," Utrecht University, Technical Report UU-CS-2001-03, 2001.

[172] N. Ansari and E. J. Delp, "Partial shape recognition: A landmark-based approach," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 5, pp. 470–483, 1990.

[173] M. Safar, C. Shahabi, and X. Sun, "Image retrieval by shape: A comparative study," in *IEEE Int'l Conf. on Multimedia and Expo (ICME'00), NY, USA*, Aug. 2000, pp. 141–144.

[174] Y. Rui, A. C.She, and T. S. Huang, "Modified fourier descriptors for shape representation - a practical approach," in *1st Int'l Workshop on Image Databases and MultiMedia Search, Amesterdam, Holland*, Aug. 1996.

[175] Y. Rui, A. C.She, and T. S. Huang, "A modified fourier descriptor for shape matching in MARS," *Image Databases and Knowledge Engineering, (ed. S.K. Chang), World Scientific Publishing House in Singapore*, vol. 8, pp. 165–180, 1998.

[176] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. of Information Theory*, vol. 8, 1962.

[177] A. Khotanzad and Y. Hong, "Invariant image recognition by zernike moments," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.

[178] D. Shen and H. Horace, "Discriminative wavelet shape descriptors for recognition of 2D patterns," *Pattern Recognition*, vol. 32, no. 2, pp. 151–156, 1999.

[179] B. Zitova and J. Flusser, "Landmark recognition using invariant features," *Pattern Recognition Letters*, vol. 20, pp. 541–547, 1999.

[180] J. Flusser, "Refined moment calculation using image block representation," *IEEE Trans. on Image Processing*, vol. 9, no. 11, pp. 1977–1978, Nov. 2000.

[181] H. Blum, "A transformation for extracting new descriptors of shape," *Models for Perception of Speech and Visual Forms, MIT Press*, pp. 362–380, 1967.

[182] H. Blum, "Biological shape and visual science (part 1)," *Journal of Theoretical Biology*, no. 38, pp. 205–287, 1973.

[183] I. Bierderman, *Visual Object Recognition, In An Invitation to Cognitive Science (S. F. Kosslyn and D. N. Osherson (Eds.))*. Visual Cognition. MIT Press, 1995, ch. 4, pp. 121–165.

[184] F. L. Bookstein, "Landmark methods for forms without landmarks: Localizing group differences in outline shape," *Medical Image Analysis*, vol. 1, no. 3, pp. 225–244, 1997.

[185] A. Neumann and C. Lorenz, "Statistical shape model based segmentation of medical images," *Computerized Medical Imaging and Graphics*, vol. 22, no. 2, pp. 133–143, 1998.

[186] J. Marchant and C. Onyango, "Fitting grey level point distribution models to animals in scenes," *Image and Vision Computing*, vol. 13, no. 2, pp. 3–12, Feb. 1995.

[187] J. Iivarinen, M. Peura, J. Särelä, and A. Visa, "Comparison of combined shape descriptors for irregular objects," in *Proc. 8th British Machine Vision Conf. (BMVC'97)*, Sept. 1997, pp. 430–439.

[188] V. Mezaris, H. Doulaverakis, S. Herrmann, B. Lehane, N. OConnor, I. Kompatsiaris, W. Stechele, and M. Strintzis, "An extensible modular common reference system for content-based information retrieval: the SCHEMA reference system," in *Proc. 4th Int'l Workshop on Content-Based Multimedia Indexing (CBMI'05), Riga, Latvia*, June 2005.

[189] N. Ueda and S. Suzuki, "Learning visual models from shape contours using multiscale convex/concave structure matching," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 4, pp. 337–352, 1993.

[190] R. Veltkamp and M. Hagedoorn, "State of the art in shape matching," Utrecht University, Technical Report UU-CS-1999-27, 1999.

[191] T. Sebastian, P. Klein, and B. Kimia, "Recognition of shapes by editing shock graphs," in *8th Int'l Conf. on Computer Vision, Vancouver*, 2001, pp. 755–762.

[192] L. J. Latecki and R. Lakämper, "Contour-based shape similarity," in *Proc. Int'l Conf. on Visual Information Systems, Amsterdam*, vol. LNCS 1617, June 1999, pp. 617–624.

[193] I. Cohen, N. Ayache, and P. Sulger, "Tracking points on deformable objects using curvature information," in *Proc. 2nd European Conf. on Computer Vision (ECCV'92)*, Sept. 1997, pp. 458–466.

[194] T. B. Sebastian, P. N. Klein, and B. B. Kimia, "On aligning curves," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 1, pp. 116–125, Jan. 2003.

[195] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell, "An efficiently computable metric for comparing polygonal shapes," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 13, no. 3, p. 209216, Mar. 1991.

[196] E. Keogh, "Exact indexing of dynamic time warping." in *Proc. 28th Int'l Conf. on Very Large Data Bases, Hong Kong*, 2002, pp. 406–417.

[197] L. S. Shapiro and J. M. Brady, "A modal approach to feature-based correspondence," in *Proc. 2nd British Machine Vison Conf. (BMVC'91), Springer-Verlag*, Sept. 1991, pp. 78–85.

[198] D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 9, pp. 850–863, 1993.

[199] S. Umeyama, "Parameterized point pattern matching and its application to recognition of object families," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 2, pp. 136–144, Feb. 1993.

[200] S. Sclaroff and A. P. Pentland, "Modal matching for correspondence and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 6, pp. 545–561, 1995.

[201] X. Yi and O. I. Camps, "Line-based recognition using a multidimensional hausdorff distance," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 21, no. 9, pp. 901–916, Sept. 1999.

[202] W. E. L. Grimson and D. P. Huttenlocher, "On the sensitivity of the hough transform for object recognition," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 12, no. 3, pp. 255–274, Mar. 1990.

[203] W. J. Christmas, J. Kittler, and M. Petrou, "Structural matching in computer vision using probabilistic relaxation," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 17, no. 8, pp. 749–764, Aug. 1995.

[204] J. W. Gorman, O. R. Mitchell, and F. P. Kuhl, "Partial shape recognition using dynamic programing," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 10, no. 2, pp. 257–266, 1988.

[205] A. Witkin, "Scale-space filtering," in *Proc. 8th Int'l Joint Conf. on Artificial Intell. (IJCAI'83)*, 1983, pp. 1019–1022.

[206] F. Mokhtarian, S. Abbasi, and J. Kittler, "Efficient and robust retrieval by shape content through curvature scale space," in *Proc. Int'l Workshop on Image Database and Multimedia Search, Amsterdam, the Netherlands,* `http://www.ee.surrey.ac.uk/Research/VSSP/imagedb/demo.html`, 1996, pp. 35–42.

[207] F. Mokhtarian and R. Suomela, "Robust image corner detection through curvature scale space," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 20, no. 12, pp. 1376–1381, Dec. 1998.

[208] L. J. Latecki, R. Lakämper, and U. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR'00)*, 2000, pp. 424–429.

[209] A. Quddus, F. A. Cheikh, and M. Gabbouj, "Wavelet-based multi-level object retrieval in contour images," in *Proc. the Very Low Bit rate Video Coding (VLBV99) workshop, Kyoto, Japan*, Oct. 1999, pp. 43–46.

[210] E. Milios and E. Petrakis, "Shape retrieval based on dynamic programming," *IEEE Trans. on Image Processing*, vol. 9, no. 1, pp. 141–147, 2000.

[211] T. Adamek and N. E. O'Connor, "A multi-scale representation method for non-rigid shapes with a single closed contour," *IEEE Trans. Circuits Syst. Video Technol., special issue on Audio and Video Analysis for Multimedia Interactive Services*, no. 5, May 2004.

[212] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.

[213] L. Davis, "Understanding shapes: Angles and sides," *IEEE Trans. on Computers*, vol. 26, no. 3, pp. 236–242, Mar. 1977.

[214] C. M. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoff in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 12, 1980.

[215] S. Jeannin and M. Bober, "Description of core experiments for MPEG-7 motion/shape," MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG99/N2690, Seoul, Mar. 1999.

[216] M. Bober and W. Price, "Report on results of the CE-5 on contour-based shape," MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG00/M6293, Beijing, July 2000.

[217] S. Belongie, J. Malik, and J. Puzicha, "Matching shapes," in *8th IEEE Int'l Conf. on Computer Vision*, July 2001, pp. 452–463.

[218] L. Latecki and R. Lakämper, "Shape similarity measure based on correspondence of visual parts," *IEEE Trans. Pattern Anal. and Machine Intell.*, pp. 1185–1190, Oct. 2000.

[219] "ISO/IEC MPEG7 eXperimentation model (XM software)," `www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html`.

[220] C. Van Rijsbergen, *Information Retrieval*, 2nd ed. Dept. of Computer Science, Univ. of Glasgow, 1979.

[221] T. B. Sebastian, P. N. Klein, and B. B. Kimia, "Shock-based indexing into large shape databases," *Lecture Notes in Computer Science*, vol. 2352, pp. 731–746, 2002.

[222] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipies in C: The Art Of Scientific Computing*. Cambridge University Press, Cambridge, 1986.

[223] J. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, no. 7, pp. 308–313, 1965.

[224] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proc. SIGIR Conf. (SIGIR'04)*, July 2004.

[225] C. I. Tomai, B. Zhang, and V. Govindaraju, "Transcript mapping for historic handwritten document images," in *Proc. 8th Intl Workshop on Frontiers in Handwriting Recognition*, Aug. 2002, pp. 413–418.

[226] T. Rath and R. Manmatha, "Features for word spotting in historical manuscripts," in *Proc. Int'l Conf. on Document Analysis and Recognition (ICDAR'03)*, vol. 1, 2003, pp. 218–222.

[227] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR'03)*, vol. 2, 2003, pp. 521–527.

[228] S. Marinai and A. D. (eds.), "Document analysis systems vi," in *Proc. 6th Int'l Workshop DAS'04, Florence, Italy*. Verlag, LNCS 3163, Sept. 2004.

[229] J. He and A. Downton, "User-assisted OCR enhancment for digital archive construction," in *Proc. Int'l Conf. On Visual Information Engineering*, Apr. 2005.

[230] S. M. Harding, W. B. Croft, and C. Weir, "Probabilistic retrieval of OCR degraded text using n-grams," in *Proc. 1st European Conf. on Research and Advanced Technology for Digital Libraries, Pisa, Italy*, Sept. 1997, pp. 345–359.

[231] C. C. Teppert, C. Y. Suen, and T. Wakahara, "The state of the art in on-line handwriting recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 8, pp. 787–808, Aug. 1990.

[232] D. Cheng and H. Yan, "Recognition of handwritten digits based on contour information," *Pattern Recognition*, vol. 31, no. 3, pp. 235–255, 1998.

[233] J. Favata, "Character model word recognition," in *5th Int'l Workshop on Frontiers in Handwriting Recognition, Essex, England*, Sept. 1996, pp. 437–440.

[234] S. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 2, pp. 149–164, Feb. 2001.

[235] R. F. Farag, "Word-level recognition of cursive script," *IEEE Trans. Comput.*, vol. C-28, no. 2, pp. 172–175, Feb. 1979.

[236] G. Dzuba, A. Filatov, D. Gershuny, and I. Kil, "Handwritten word recognition-the approach proved by practice," in *Proc. Sixth Int'l Workshop Frontiers in Handwriting Recognition*, 1998, pp. 99–111.

[237] S. Madhvanath, G. Kim, and V. Govindaraju, "Chaincode contour processing for handwritten word recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 9, pp. 928–932, Sept. 1999.

[238] S. Madhvanath, E. Kleinberg, and V. Govindaraju, "Holistic verification of handwritten phrases," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 12, pp. 1344–1356, Dec. 1999.

[239] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, N.J.: Prentice Hall, 1986.

[240] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen, "Adaptive document binarization," in *Proc. Int'l Conf. on Document Analysis and Recognition*, vol. 1, 1997, pp. 147–152.

[241] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 176–201, Apr. 1993.

[242] J. He, Q. Do, A. Downton, and J. Kim, "A comparison of binarization methods for historical archive documents," in *Proc. 7th Int'l Conf. on Document Analysis and Recognition (ICDAR'2005), Seoul, Korea*, Aug. 2005.

[243] T. Adamek, N. E. O'Connor, and N. Murphy, "Multi-scale representation and optimal matching of non-rigid shapes," in *Proc. 4th Int'l Workshop on Content-Based Multimedia Indexing, CBMI'05, Riga, Latvia*, June 2005.

[244] R. Manmatha and W. B. Croft, "Word spotting: Indexing handwritten archives," in *Proc. Intelligent Multi-media Information Retrieval Collection, M. Maybury (ed.), AAAI/MIT Press*, 1997.

[245] S. Kane, A. Lehman, and E. Partridge, "Indexing George Washington's handwritten manuscripts," Technical Report MM-34, Center for Intelligent Information Retrieval, University of Massachusetts Amherst, Tech. Rep., 2001.

[246] T. Adamek and N. O'Connor, "Efficient contour-based shape representation and matching," in *Proc. 5th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval (MIR'03), Berkeley, CA*, Nov. 2003.

[247] L. Huang and M. Wang, "Efficient shape matching through model-based shape recognition," *Pattern Recognition*, vol. 29, no. 2, pp. 207–215, 1996.

[248] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*. John Wiley & Sons, 1998.

[249] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Jour. Royal Statistical Society, Series B*, vol. 53, pp. 285–339, 1991.

[250] R. H. Davies, "Learning shape: Optimal models for analysing shape variability," Ph.D. dissertation, University Of Manchester, 2002.

[251] A. Hill and C. J. Taylor, "A method of non-rigid correspondence for automatic landmark identification," in *Proc. 7th British Machine Vison Conf. (BMVC'96), BMVA Press*, Sept. 1996, pp. 323–332.

[252] A. Benayoun, N. Ayache, and I. Cohen, "Human face recognition: From views to models - from models to views," in *Proc. Int'l Conf. on Pattern Recognition*, Sept. 1994, pp. 225–243.

[253] C. Kambhamettu and D. Goldgof, "Point correspondence recovery in non-rigid motion," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'92)*, Sept. 1992, pp. 222–227.

[254] A. C. W. Kotcheff and C. J. Taylor, "Automatic construction of eigenshape models by direct optimisation," *Medical Image Analysis*, vol. 2, no. 4, pp. 303–314, Sept. 1998.

[255] H. Thodberg and H. Olafsdottir, "Adding curvature to minimum description length shape models," in *Proc. 14th British Machine Vision Conf. (BMVC'03), BMVA Press*, vol. 2, 2003, pp. 251–260.

[256] A. Ericsson and K. Åström, "Minimizing the description length using steepest descent," in *Proc. 14th British Machine Vision Conf. (BMVC'03), BMVA Press*, vol. 2, 2003, pp. 93–102.

[257] G. Scott and H. Longuet-Higgins, "An algorithm for associating the features of two images," *Phil. Trans. Roy. Soc. London*, vol. B 244, pp. 21–26, 1991.

[258] N.Duta, A. Jain, and M. Dubuisson-Jolly, "Automatic construction of 2d shape models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 5, pp. 433–446, 2001.

[259] R. Davies, T. Cootes, C. Twining, and C. Taylor, "An information theoretic approach to statistical shape modelling," in *Proc. 12th British Machine Vison Conf. (BMVC'01), BMVA Press*, Sept. 2001, pp. 3–11.

[260] J. Hladuvka, "Establishing point correspondence on training set boundaries," VRVis Technical Report # 2003-040, Tech. Rep. 2003-040, 2003.

[261] T. Cootes, "Timeline of developments in algorithms for finding correspondences across sets of shapes and images," University of Manchester, `http://www.isbe.man.ac.uk/~bim`, Tech. Rep., June 2004.

[262] M. Fleute and S. Lavallée, "Building a complete surface model from sparse data using statistical shape models: Application to computer assisted knee surgery system," in *Proc. Int'l Conf. on Medical Image Computing and Computer Assisted Intervention*, 1998, pp. 879–887.

[263] P. Zhu and P. Chirlian, "On critical point detection of digital shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 8, pp. 737–748, Aug. 1995.

[264] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla, "Shape context and chamfer matching in cluttered scenes," in *Proc. Conf. Computer Vision and Pattern Recognition, Madison, USA*, June 2003.

[265] E. Hjelmas and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[266] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 1, pp. 696–706, Jan. 2002.

[267] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman, 1982.

[268] I. Bierderman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[269] M. J. Tarr and H. H. Bülthoff, "Image-based object recognition in man, monkey, and machine," *In Special issue on Object Recognition in Man, Monkey, and Machine, M. J. Tarr and H. H. Bülthoff (Eds.), Cognition*, no. 67, pp. 1–20, 1998.