Semi-Automatic Person-Annotation in Context-Aware Personal Photo Collections

Neil O'Hare

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Prof. Alan Smeaton

September 2007

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy 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: Neil Rhone

ID No. 51164116

Date: 24/09/2007

Acknowledgements

Firstly, I would like to thank my supervisor Alan Smeaton for all his guidance and support over the last four years, particularly in the final months when he pushed me over the finish line. I am indebted to Enterprise Ireland for funding my work on the MediAssist project.

The help and support of all the lecturers within the CDVP, in particular those involved in the MediAssist project, is also greatly appreciated: Gareth Jones and Cathal Gurrin for their input into the MediAssist system, Hyowon Lee for his invaluable suggestions regarding the MediAssist interface, and Noel O'Connor.

I would like to thank the other researchers in the CDVP for providing analysis tools used in my work, particularly Saman Cooray for providing face detection and body patch tools, and Bartlomiej Uscilowski for providing face recognition.

I am very grateful to all users of the MediAssist system for providing photos and for making invaluable comments. I particularly wish to thank all the users who provided the photos used in my experiments in this thesis: Paul Browne, Paul Ferguson, Cathal Gurrin, James Lanagan, Sinéad Mc Givney, Sandra Rothwell, Alan Smeaton and Peter Wilkins.

I want to thank all my colleagues who made my life in DCU fun these past four years: Aiden, Bart, Colum, Daragh, Dave, Eamonn, Fabio, Fabrice, Georgina, Jiamin, Kieran, Kirk, Liadh, Mike and of course everyone else already mentioned above.

I thank both my parents, Margaret and Jim, for all their support over the years.

Finally, I wish to thank my girlfriend Isabelle, your love and support helped get me through those lonely final months.

Abstract

Recent years have seen a revolution in photography with a move away from analogue film capture towards digital capture technologies, resulting in the accumulation of large numbers of personal digital photos. This means that people now have very large collections of their own personal photos, which they must manage and organise.

In this thesis we present a prototype context-aware photo management system called MediAssist, which facilitates browsing, searching and semi-automatic annotation of personal photos. We propose an approach to semi-automatic person-annotation in personal photo collections that facilitates the annotation of people in personal photo collections in a batch manner, by suggesting annotations to users as they interact with the system. We propose person classification and retrieval techniques based on analysis of the context of photo capture in addition to analysis of the image content of the photo We use classification techniques to suggest names for faces detected in photos, and retrieval techniques suggest faces for a query name We implement the proposed techniques and integrate them into the interface of the MediAssist prototype photo management system.

We provide a comprehensive empirical study of the proposed person classification and retrieval techniques, using the real photo collections of a number of users. We also conduct a user study which confirms the effectiveness of the semi-automatic person-annotation approach, and the utility of the system for real users when used as part of a photo management system

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent years have seen a revolution in photography with a move away from analog film towards digital technologies resulting in the accumulation of large numbers of digital photos for personal use. The ease of photo capture and the capacity of storage devices mean that users now acquire more and more photos This increased volume of personal photos means that technologies for efficiently managing and organising digital photos assume more and more importance, as user wish to efficiently browse and search through larger and larger photo collections. Commercial systems tend to facilitate thumbnail views, time-based display and manual annotation. While these tools are fine for smaller photo collections they prove inadequate as collections sizes grow much larger.

Content-based image retrieval techniques (Smeulders et al., 2000) have tried to solve the problems of general image management through analysis of the visual content of images, facilitating searching and browsing of image collections based on this visual image content. The *semantic gap*, which is the gap between the information available from the visual content of an image and the semantic interpretation of that image by a given user, means that these content-based image retrieval techniques are not yet able to provide adequate solutions for the management of personal photos. It seems unlikely, in the short term at least, that this semantic gap will be bridged by the use of content-based analysis alone, and photo management systems based

purely on the analysis of visual image content are unlikely to solve user needs

Personal photos differ from general images, however, in that they have an associated context. The photos are normally captured by the user of the photo management system, and that user will have personal recollection about the time, place and other context information relating to the environment of photo capture. Most digital personal photos make a certain amount of contextual metadata available in their EXIF header (JEITA, 2002), which stores the time of photo capture and some basic camera settings such as lens aperture, exposure time and whether or not the flash was fired GPS location information is also supported and, although not captured by most commercial cameras at the moment, there are ways of 'location-stamping' photos using data from a separate GPS device or manually using a map interface (Toyama et al., 2003), while camera phones are inherently location-aware The N95 camera phone from Nokia[1] includes a build-in GPS sensor. Systems for managing personal photo collections can make use of this contextual metadata in their analysis and organisation of personal photos.

1.1 The MediAssist Context-Aware Personal Photo Management System

In this thesis we will present the MediAssist system, a prototype context-aware photo management system for location-aware photos and we use this to develop and test our hypothesis on context- and content-based photo management. The system incorporates automatic context-based and content-based analysis for the automatic annotation of personal photos with a number of concepts. The context-based analysis converts the location coordinates to place-names, and the time is represented both in terms of physical linear time and in terms of the cyclic temporal schemata by which people remember temporal information, such as month of the year, day of the month, day of the week and hour of the day. In addition to these

[1] http://www.nseries.com/products/n95/

basic contextual features, the system also performs further analysis of the contextual information by performing light status classification in terms of day, night, dusk or dawn, indoor/outdoor classification, and weather status classification. A user's photo collection is also automatically segmented into 'events' based on the contextual information.

In addition to the context-based analysis, the system also indexes photos using content-based analysis tools A face detection module detects faces in the photos. These detected faces are then analysed by both context-based and content-based person matching tools. A building detection module detects buildings in photos. The results of all this analysis are also written to surrogate text files, which can be indexed by a conventional text search engine to allow for conventional text search of photo collections

The MediAssist prototype system interface provides tools for browsing, searching and semi-automatic annotation of personal photo collections based on this analysis, providing a sophisticated filter-based interface for formulating complex queries, alongside a simple text search alternative. Semi-automatic annotation tools allow users to correct any errors in the automatically created annotations, with the option of batch annotation to annotate multiple photos simultaneously. The system also provides tools for semi-automatic person-annotation, as discussed in more detail in the next section. The system provides a background for the research hypotheses proposed and evaluated in this thesis, and the MediAssist system and its users provide a testbed for these hypotheses.

1.2 Hypothesis

In our research into semi-automatic person-annotation in this thesis we are interested in exploring two related hypotheses about person annotation, both of which are stated below.

Hypothesis 1 (H1)

> *Semi-automatic person-annotation techniques for context-aware personal photo collections can be developed and can perform effectively.*

We will propose using person classification and retrieval techniques to suggest annotations for unannotated faces in personal photo collections, and we will propose a user interface that allows users to carry out this annotation in a batch manner.

Hypothesis 2 (H2)

> *Person classification and retrieval techniques, which use context-based analysis in addition to content-based analysis, can be developed and can perform effectively*

We will propose approaches to person classification and person retrieval based on the use of context-based and context-based analysis, and based on a combination of context and content. We are particularly interested in exploring the role of context for person classification and retrieval and in discovering whether is it possible to improve on the results of content-based analysis through the additional use of context.

1.3 Person-Annotation in Personal Photo Collections

In this thesis we are primarily interested in annotating people in personal photo collections. It is possible to use face recognition technologies (Zhao et al., 2003) to classify the names of people in personal photos, although such approaches often do struggle due to the challenging nature of the data, particularly variations in pose, expression and illumination. A number of researchers have proposed a 'body patch' feature to exploit the observation that, within a constrained environment such as an event, people tend to wear the same clothes (Zhang et al., 2003, Anguelov

et al , 2007). It is also possible to exploit the context of photo capture for person classification, assuming that people tend to re-occur in similar locations and that photos taken in close temporal proximity to each other tend to contain the same people (Naaman et al., 2005).

For *Hypothesis 1* we will propose an approach to person-annotation that will facilitate semi-automatic annotation of people in a batch manner We do this by suggesting names for unknown faces and, at appropriate moments as the user browses their collection, suggesting a list of candidate faces for a given name, to be annotated in a batch manner. These complementary annotation mechanisms require person classification techniques to suggest names for an unknown person, and person retrieval techniques to suggest faces to match a person name. Our aim is to take advantage of regularities within a constrained context such as an event, and to use existing annotations from that constrained context to assist in the person classification and retrieval tasks.

For *Hypothesis 2* we will propose the use of a combination of context-based and content-based analysis to perform this person classification and retrieval. For context-based analysis, we will propose a language model approach from information retrieval (Ponte and Croft, 1998), which will use smoothing techniques and hierarchical smoothing techniques to estimate the probabilities of the model. We will also propose a nearest neighbour (Duda et al., 2000) context-based approach, which ranks candidates based on temporal and spatial distance from known faces. For content-based person classification we will use a nearest neighbour approach using face recognition and body patch features, in addition to two features we expect to model the properties of photos in highly constrained environments, such as within the same event. These features are face colour and image colour. We will also propose combined context-based and combined content-based approaches, and an approach to person classification and retrieval that combines context and content information.

1.4 Research Objectives

We will conduct a thorough evaluation of these proposed approaches using a database of photos from 9 users. Our research is primarily concerned with exploring the two related hypotheses *H1* and *H2*, stated above. Our research differs from other approaches in that we thoroughly investigate the use of context for person classification, and examine to what extent the direct use of context can benefit the performance of content-based person classification and retrieval techniques. We also investigate the use of face colour and image colour features in an environment where there are known annotations in the local context of the unknown faces, to determine the utility of these features in such a constrained environment. We will evaluate retrieval performance in addition to classification performance, in order to investigate the performance of our approaches in each of these scenarios, and to determine if similar approaches perform similarly for each scenario.

We will restrict our evaluation of person classification and retrieval to the scenario where there are already known annotations in the local context of candidate face, which in practice means the same event or the same day. This will allow us to thoroughly evaluate the utility of our proposed features, which depend on local context information, when such information from within the same event is available, and to evaluate if they can improve the results given by standard techniques such as face recognition.

Given this scenario, the specific questions that we are interested in answering in this thesis, in addition to verifying hypotheses *H1* and *H2*, are:

- Are the proposed hierarchical smoothing language model and nearest neighbour approaches effective for person retrieval in addition to person classification? How do these approaches compare with each other? Which hierarchical structures are most effective?

- Are the proposed *face colour* and *image colour* features useful for person classification and retrieval?

- How does the performance of context-based approaches compare with the performance of content-based approaches?

- Can combined context-based and content-based approaches to person classification and retrieval improve performance compared with content-based approaches alone and context-based approaches alone? Is is necessary to learn combination weights to achieve this improvement?

- Is it more effective to annotate at the level of the photo and use photo-level features for face classification and retrieval, or is it more effective to annotate and the face region-level and use face region-level features for face classification and retrieval?

We will implement our proposed approaches in the user interface of the MediAssist system, and conduct a user study to evaluate the effectiveness of the system for real users performing annotation tasks. From this user study we will learn whether users can efficiently use our person-annotation system, and we can measure user satisfaction with the system and determine whether they are happy with the quality of the annotation suggestions from the system, and whether they would be likely to use these annotation tools for their real world photo management.

1.5 Thesis Organisation

The remainder of this thesis is organised as follows.

Chapter 2. In chapter 2 we introduce related work in the field of personal photo management. We outline existing approaches to event detection in personal photo collections based on analysis of time, location and image content. We then discuss approaches to person classification, before discussing approaches to simplifying the manual annotation process. We also outline the features generally available in commercial and research photo management systems.

Chapter 3. In chapter 3 we describe the MediAssist system, a prototype context-aware photo management system developed by the author of this thesis. The system uses a location-aware photo collection and facilitates browsing, searching and semi-automatic annotation of personal photos. Context-based and content-based analysis extends the available metadata associated with photos, with semi-automatic annotation allowing users to remedy any errors in the automatic annotations, either one at a time or in a batch manner. The system provides powerful search tools that allow users to browse their collection based on any of the context-based and content-based features that it automatically extracts. We will describe all of the analysis techniques used by the system, and we will also describe the user interface of this system in detail.

Chapter 4. In chapter 4 we propose our approach to semi-automatic person-annotation in personal photo collections. We extend the basic MediAssist system to allow for semi-automatic batch person-annotation. We do this by having the system suggest names for any faces encountered as the user browses their collection and, at appropriate moments as the user interacts with the system, suggesting a set of faces to be annotated with a specific candidate person name. We then describe the implementation of the proposed semi-automatic person-annotation approach in the MediAssist user interface.

Chapter 5. In chapter 5 we propose approaches to person classification and retrieval using a partially annotated photo collection as the training set. We propose a context-based language model approach to person classification, inspired by information retrieval (Ponte and Croft, 1998), which is an extension of an existing approach (Naaman et al., 2005). The language model approach uses Jelinek-Mercer smoothing (Lafferty and Zhai, 2001) and hierarchical smoothing techniques (Westerveld et al., 2003b) to estimate the probabilities of the model. We also propose nearest neighbour (Duda et al., 2000) context-based approaches based on temporal and spatial proximity to annotated faces. For content-based person classification and

retrieval we also propose using nearest neighbour approaches, using face recognition and body patch features in addition to face colour and image colour, two features we propose will be useful for person classification and retrieval if we restrict our analysis to photos taken in the same event. We then propose combined approaches based on combining multiple context-based features, multiple content-based features, and an approach that combines both context-based and content-based features.

Chapter 6. In chapter 6 we evaluate the proposed context-based and content-based approaches to person classification and retrieval. We first evaluate our proposed context-based language model and nearest neighbour approaches. We compare different temporal and spatial hierarchical structures for hierarchical smoothing approaches to both person classification and person retrieval. We then evaluate the relative performance of the content-based features. We evaluate both combined context-based approaches and combined content-based approaches, before evaluating approaches that use both content and context.

Chapter 7. In chapter 7 we describe a study that examines the usability of the proposed annotation interface for real users. We compare the semi-automatic batch annotation system with an alternative that does not facilitate batch annotation, measuring the efficiency of the annotation process, and observing user behaviour with the system. We elicit qualitative user reactions to the system in the form of a questionnaire.

Chapter 8. In the final chapter we summarise the results of our work and the lessons learned. We also outline possible extensions to our approach and directions for future work.

Chapter 2

Organising Personal Photo Collections

Personal photo collections have unique characteristics that set them apart from other image collections, making approaches to managing and organising them markedly different from approaches to other content-based image retrieval (CBIR) applications (Smeulders et al., 2000), which tend to extract low-level image features based on colour and texture and allow users to search image collections based on similarity to a query image using these low-level features. User studies have shown a low level of willingness to use such content-based technologies for managing digital photos. The results of a user study by Rodden and Wood (2003) show that many users never even wanted to use such visual queries when using a system which provided these capabilities. When users did use visual queries with images as example queries, the system's notion of similarity rarely coincided with the user's notion of similarity and the results were generally not considered useful. In fact, the study found that the most important features that users require from photo management systems are the organisation of photos in chronological order and displaying a large number of photos at once. Additionally, they found that most users wanted to be able to browse their photo collections by event, users were unlikely to invest time in manual annotation, and were often interested in the people in their photographs.

In the next section we discuss some of the features of personal photo collections that distinguish them from general image collections Then we will examine automatic approaches to event detection, followed by a description of approaches to person classification. After that we will detail some other photo analysis techniques that have been proposed to assist in the task of photo organisation We will then describe approaches to the annotation of personal photos, and approaches that attempt to make the annotation process easy for the user. Finally we will describe some photo management systems and the general features that they provide.

2.1 Features of Personal Photo Collections

The most important difference between personal photo management and general image management is that in personal photo managment the user is, in general, the author and owner of the photos. The user has a memory of the autobiographical photo capture event as they generally captured the photos themselves and have personal recollections of the time, place and other circumstances surrounding photo capture. That is, personal photos have *context*, of which the owner has intimate knowledge, and a certain amount of this context is available for automatic analysis.

2.1.1 Context in Personal Photos

Personal photo collections, as opposed to general image collections, have a wealth of contextual information associated with them. We will not attempt to give a general definition of context here, but rather we will define context in a way that distinguishes the *content* of a photographic image from the *context* of photo capture. We will use the term context to refer to the circumstances surrounding the photo-capture environment, including the time, the location, the people present, the weather conditions, the camera settings, and any other data relevant to the photograph that is not contained in the photographic image. Other working definitions of context, in the field of image processing, consider visual image content surround-

ing an object as context (Wolf and Bileschi, 2006). So, for example, a car detector could take into account the visual context surrounding a candidate car object by detecting a street scene surrounding the object, making the presence of a car more likely. For alternative definitions of context, see Chen and Kotz (2000), Dey (2001) and Dimitrova (2004).

In this thesis, we will refer to all approaches based on image analysis as content-based because they are based on analysis of image content and they do not take context outside of this image content into account. We will refer to approaches that use the context of photo capture as context-based approaches We will also distinguish between a *photograph* and an *image*. We use the term *photo* or *photograph* to refer to the entire photograph, including whatever contextual metadata we have about that photograph, in addition to the image content of the photo. When we use the term *image*, *photographic image* or *photo-image* we are referring only to the visual image content of the photo.

Some of the context surrounding photo-capture is stored in the EXIF header of digital photographs (JEITA, 2002). The EXIF standard has been adopted by most digital camera manufacturers, and recently camera-phone manufacturers have adopted this standard for their cameras Different cameras vary in the subset of the EXIF data that they actually store in photos, but they will always store the time of photo capture, and some basic camera settings such as the lens aperture, exposure time and whether a flashgun was fired or not. Many cameras will store more advanced camera settings such as the distance of the subject from the camera and the brightness of the scene at the time of photo capture.

The EXIF standard also supports GPS (Global Positioning System) information in the form of latitude/longitude co-ordinates. At present, very few commercially available digital cameras actually support direct capture of location information. There are, however, tools available to allow the time-stamps of photos to be matched against the time-stamps from the log of a GPS device. In the absence of GPS information, photos can be 'location-stamped' retrospectively by dragging them

onto a location in a map interface (Toyama et al., 2003). In addition, mobile phones which include cameras are inherently location-aware, and some researchers have exploited this fact to associate location information with photos captured using camera phones (Davis et al., 2005a).

All of this contextual metadata can be used to assist in the automatic management and organisation of photo collections.

2.2 Event Detection

We can think of a collection of personal photos as being made up of consecutive events, with an event corresponding to some significant occurrence in the user's life, for example a birthday party, a trip to the zoo, etc. We can think of an event, within a photo collection, as a cluster in time and space, because an event can always be defined in terms of time and space Events can be defined hierarchically, for example a holiday to Europe could be divided between trips to Munich, Paris and Rome. Each of these sub-events could in turn be segmented into individual day-trips.

The problem of automatically detecting events within personal photo collections has been extensively researched, with most approaches using the time of photo capture. Some approaches measure the time gap between consecutive photos, and use a (sometimes adaptive) threshold to determine whether this gap is long enough to signal an event transition (Platt, 2000; Graham et al., 2002). Other approaches use a clustering algorithm to cluster photos into groups, each group corresponding to an event (Cooper et al., 2003; Gargi, 2003). Loui and Savakis (2000) use k-means clustering, with $k = 2$, to cluster the values of the time gaps between consecutive photos, making the assumption that one of these clusters will contain only the larger time gaps that signal an event transition, while the other will contain the smaller time gaps between photos within the same event. A number of these approaches also make use of content-based image similarity to refine the event boundaries, although

time-only approaches always work better than content-only approaches (Loui and Savakis, 2000; Platt, 2000; Cooper et al., 2003).

Although most approaches rely on time information, with occasional use of image content, some approaches also use location information. Naaman et al. (2004c) extend the approach of Graham et al. (2002) by clustering time-based events using geographical information to create location clusters. They then merge the location and time-based clusters. Pigeau and Gelgon (2003, 2004, 2005) also use both temporal and geographical information, proposing a statistical approach to create distinct hierarchical temporal and geographical partitions of a user's personal photo collection, which can then be combined to create an event segmentation based on both time and location

Recently, Apple's iPhoto[1] personal photo management application has added an event detection feature, allowing management of personal photos based on events.

2.3 Person Classification

Recognising that finding photos of people in photo collections is of prime importance, many researchers have focused on methods of classifying the identity of people found in photo collections. Most systems that perform person classification in images first automatically detect any faces in the image using face detection techniques. Research in face detection is a mature field, and the current state of the art gives reasonably reliable performance for general images (Yang et al., 2002). Face detection is even availble in the Adobe PhotoShop Elements commercial photo management application.[2]

The traditional approach to person classification is to use face recognition techniques (Zhao et al., 2003), which attempt to model the faces of a database of people in order to identify unknown faces. Personal photo collections represent challenging environments for face recognition techniques due to varying lighting conditions,

[1] http //www apple com/ilife/iphoto/
[2] http //www.adobe com/products/psprelements/

facial expressions, pose etc. In spite of this, a number of researchers have proposed the use of standard face recognition techniques for identifying people in personal photo collections (Kuchinsky et al., 2001; Girgensohn et al., 2004a,b).

The Riya[3] online photo management system uses face recognition technology in the management of personal photos. The interface for this system allows users to 'train' the system to recognise known people in their collections. Automatic face detection detects faces in images, which the user can annotate manually: for this annotation the system does not provide the type of semi-automatic annotation proposed in this thesis, instead the user simply chooses from a list of names known to them. Although the system performs automatic person recognition using face recognition and body patch matching technologies, the system does not provide tools for semi-automatically annotating people in a batch manner, and they do not make direct use of context information as we do in this work.

While personal photo collections are, on one level, an unconstrained environment for photo capture with a wide variety of capture conditions, subsets of an individual's photo collection can exhibit a large amount of uniformity in capture conditions, for example photos taken at the same event people will tend to wear the same clothes. Since personal photographs include capture time information, and since it is possible to automatically segment personal photo collections into events, it is possible to exploit this observation. A number of researchers have proposed a feature to exploit this regularity. The feature has variously been called 'clothes', 'costume', 'torso' and 'body patch' (Zhang et al., 2003; Cooray et al., 2006; Sivic et al., 2006; Song et al., 2006, Anguelov et al., 2007). In our work we refer to this feature as 'body patch'.

The body patch feature has also been proposed for person identification in video (Jaffré and Joly, 2004; Everingham et al., 2006), where some similar assumptions can be made about people wearing the same clothing, and where time information of a slightly different nature is also available Person identification in video is a different problem, however, and different approaches need to be taken.

[3]http://www riya com/

When using the 'body patch' feature in personal photo management environments, the approaches invariably only use this feature if comparing faces from a similar context, such as from the same event or on the same day In this sense, we can say that these approaches are context-aware, in that they indirectly use temporal context to decide whether or not a particular content-based feature (body patch) can be used in a certain situation. Although some of the authors state that their approaches are 'contextual', they do not make *direct* use of context in the sense used in this thesis to refer to context outside the image content, in particular spatial and temporal context. Zhang et al. (2003) also calculate a global prior probability for each candidate person, based the relative frequency of occurrence of the person, and use this prior an input into their classifier, although this prior is not context-based as it is a global prior calculated over all annotations in the user's collection. Sivic et al. (2006), in addition to using the body patch feature for person recognition, also use the feature to detect the re-appearance of people in cases where the face has not been detected by face detection techniques

There are a few existing systems that make direct use of context in the person classification process. Naaman et al. (2005) propose an approach that estimates the probability of a person occurring in a given photo based on previous annotations of other photos with a similar context to the photo. So, for example, if a person is known to occur in photos frequently within a certain time interval of the capture time of a given photo, that person will have a high probability of occurring in that photo. This method does not need to analyse the content of an image in order to suggest person annotations for the photo. We describe the approach in more detail in chapter 5, where we also extend their approach.

Zhao et al. (2006) use a combination of context-based and content-based analysis for the automatic annotation of people in photos For context-based analysis the use the approach proposed by Naaman et al. (2005), while for content-based analysis they use body patch and face recognition features. They use these features to cluster faces for 'automatic' person annotation Although they report that their approach

has the advantage that it is fully automatic, we do not believe that the accuracy of their results is high enough to support such automatic annotation without user correction. Our work differs from this work in that we propose a framework for semi-automatic annotation make the best of the automatic analysis. We conduct a thorough exploration of the utility of context-based approaches and we use additional content-based features, along with body patch and face recognition, in our analysis.

The approach of Davis et al. (2005b, 2006) also makes direct use of both content and context for person classification. Their person classification system is based on the Mobile Media Metadata context-aware camera phone system, that includes bluetooth co-presence information and enables photo sharing. For person classification they use standard face recognition tools. In addition to this they use spatial and temporal contextual features, along with bluetooth co-presence information, which determines if a certain individual was present at the time of photo capture, and photo sharing information about the people the user shared the photo with. These extra features are specific to the Mobile Media Metadata system, and it is unclear if such features would ever be available from general photo capture devices. The approach makes limited use of temporal information, using a weekend/weekday feature, and an hour of the day feature Their approach also has the disadvantage that their algorithm could not make direct use of GPS co-ordinates, forcing them to cluster the GPS co-ordinates. They choose the number of clusters empirically, but do not propose an approach to learning the number of clusters automatically. In this thesis we will explore alternative approaches to using temporal context for person classification, including temporal proximity to known person occurrences, and we will make direct use of location information that does not need any pre-processing in the form of clustering.

2.4 Automatic Photo Analysis

In addition to event detection and person classification, researchers have proposed a number of other approaches to analysing personal photos, based on both the context of photo capture and image content. Lim et al. (2003) propose using content-based image analysis techniques to automatically label photos with a number of concepts, for example people, sky, mountain or building The use of content-based image similarity can be found in commercial photo managment tools such as Adobe Photoshop Elements. The context metadata in photos can also be exploited to enhance the metadata associated with a photo. The PhotoCompas system (Naaman et al., 2004a) leverages extra contextual information from the time and location either automatically or using external resources, extracting features such as season, weather and light status. The system also allows browsing based on time, location, and other contextual cues

A number of researchers propose approaches that use a combination of content and context to automatically label photos. Boutell and Luo (2005) propose enhancing content-based techniques using context information from the EXIF metadata in photos to aid with image classification. They use metadata features such as exposure time, subject distance, flash, and aperture in addition to content-based image analysis. They apply their approach to the tasks of indoor/outdoor classification, sunset detection and man-made/natural scene classification, and show that context improves performance for each classification task Indeed, for the indoor/outdoor classification task, they showed that classification based on context-alone worked better than classification based on content-alone The MyPhotos system (Sun et al., 2002) analyses image content to detect underexposed photos, duplicate photos, photos with red-eye and photos that need to be rotated, so that the user can remove or fix problematic photos.

The MetaXa system (Boll et al., 2007) proposes an architecture for the extraction and enhancement of photo metadata. The authors extract content-based features

such as face detection and an image colour feature, and perform sharpness analysis to determine the quality of photos They enhance the context information to derive additional features such as light conditions and indoor/outdoor classification, and the extracted metadata is used to help the system automatically create photo albums.

The MediAssist system (Gurrin et al., 2005; O'Hare et al., 2005b, 2006, 2007) enhances photo metadata to extract additional context-based and content-based features, in addition to converting all of these extracted features to text form. This 'text surrogate' allows indexing by a text search engine, facilitating automatic search of personal photos using text-based search techniques. This MediAssist system will be described in detail in Chaper 3. The author of this thesis has also proposed an approach to retrieving photos of landmarks, such as buildings, in large, shared, photo collections, based on combining content-based image retrieval techniques with context in the form of location information (O'Hare et al., 2005a).

Other work has looked at inferring semantic content for photos using the available context by sharing available labels, which have been created manually, between photos with a similar context (Naaman et al., 2003; Davis et al., 2004), while the Mobile Media Metadata system (Davis et al., 2005a) uses context to recommend recipients for sharing photos taken with a context-aware phone.

2.5 User Annotation

Automatic analysis will never give users the full range of metadata they require for their personal photos, and can never be expected to be 100% accurate. Many systems provide manual annotation tools to assist the user in adding extra metadata to their photos. Annotations can be in the form of free text annotations or can be ontology-based. Commercial systems such as iPhoto[4] allow users to type free-text annotations, and photo sharing websites like Flickr[5] allow for such photo tagging by

[4]http://www.apple.com/ilife/iphoto/
[5]http://www.flickr.com/

an online community. These systems allow for the annotation of people in personal photo collections, but they do not attempt to use semi-automatic techniques to make the process easier for the user.

Shneiderman and Kang (2000) proposed drag and drop annotation to allow users to easily drag pre-defined concept labels onto images. This approach has been incorporated into commercial photo management software, and extended to allow for batch annotation, a process whereby numerous photos can be labelled simultaneously (for example, Adobe Photoshop Elements and Picasa[6]). Alternative approaches propose audio annotations, which are processed using automatic speech recognition techniques to create text annotations (Srihari et al., 1999, Mills et al., 2000, Chen et al , 2003a).

Some systems attempt to leverage the techniques of content-based image retrieval to enhance the annotation process. The MiAlbum system (Wenyin et al., 2000), for example, uses a semi-automatic approach to image annotation based on image retrieval techniques, and Bissol et al (2003) allow a user to define the indexing concepts (e.g. a car, or a tree) which the system will then learn how to detect automatically, using image classification techniques. *Semi-automatic* annotation approaches use a combination of automatic and manual techniques. automatic image analysis techniques will suggest annotations to the user, who can then confirm them, sometimes one at a time or alternatively in a batch manner. The PhotoCopain system is another system that suggests annotations for photos, based on image analysis, for the user to confirm or reject (Tuffield et al , 2006)

A number of systems also facilitate the semi-automatic annotation of people in photo collections. The first such approach was proposed by Kuchinsky et al. (2001), and used face recognition technologies to suggest names for faces in images, which the user could confirm one at a time. Zhang et al. (2003) use face recognition and body patch features to classify the names of people in images, again prompting the user for confirmation. Girgensohn et al. (2004a) present an approach to batch annotation

[6]http://picasa.google.com/

of people based on similarity to known people using face recognition technology. This is the most similar existing technique to the interactive annotation approach we will propose in this thesis. Unlike our work, however, they rely solely on face recognition technology, and they do not conduct a user evaluation of their system. The Saphari system (Suh and Bederson, 2007) also facilitates batch annotation of people in personal photos. Their approach relies solely on the 'body patch' feature, clustering people based on this feature and allowing the user to annotate person clusters in a batch manner. The system also supports semi-automatic annotation of events by allowing users to refine the event boundaries *automatically* detected by their system.

The person classification work described in Section 2.3 can also be used in the context of semi-automatic person annotation, as they can provide annotation suggestions that can be confirmed or rejected by the user. Other work attempts to propagate labels from the photo level to the face level: that is, given a number of photos known to contain the same person, we can match the name to a specific face within those photos (Zhang et al., 2004). This can be seen as complementary to the above approaches as it attempts to improve the granularity of the annotations submitted to the system Wilhelm et al. (2004) propose a system that facilitates annotation of photos using a camera phone immediately after photo capture. Outside the sphere of personal photos, the Polar Rose system[7], implemented as a web browser plugin, detects faces in images on web pages appearing in the user's browser, and allows users to annotate the names of those images, making such annotations available to all users of Polar Rose as they browse annotated web pages, and using the annotations to train face recognition models of the faces.

[7]http://www.polarrose.com/

2.6 Other Photo Management Systems

In this section we will give a brief overview of some photo management systems not mentioned in the previous sections, emphasising the main interface features that they provide. Most photo management systems rely on time to organise photo collections, often using calendar-based views to allow browsing of collections based on photo capture time Photomesa (Bederson, 2001), presents an alernative method of browsing photo collections, laying them out on a zoomable 2D space for efficient photo collection navigation.

The WWMX project (Toyama et al , 2003), though not exclusively concerned with personal photograph collections but rather providing shared access to a large online archive of location stamped images, allows navigation of a large photograph collections using a map-based interface. GTWeb creates web pages with map-overviews of trips along with associated photographs using GPS location information (Spinellis, 2003). The online photo-sharing website Flickr also allows users to browse their own, and other people's, photos on a map interface They also provide facilities for collective annotation, where users can annotate their own, and other people's, photos, although they do not support annotation in a batch manner. The Photo-Compas system (Naaman et al., 2004b) proposes an interface for browsing personal photo collections using a number of automatically extracted context-based features, and groups photos by event.

The MyLifeBits system (Gemmell et al., 2002, 2003a,b, 2006; Aris et al., 2004) is a project focused on the management of *all* of a user's personal information, not just their photos. It allows browsing of photos using a map interface, annotation of photos and navigation of photos based on links to other personal information items. So, for example, photos can be linked to calendar events, and person annotations linked to the user's contacts. The system also creates animated trip logs from photos and GPS logs, with photos animating the user's progress along a map.

Finally, the Personal Digital Historian system is a tabletop system with facilities

for managing personal information, including personal photos (Shen et al., 2001, 2002, 2003). It is designed to allow groups of people to browse their photos in a collaborative manner, and organises photos based on *'who? when? where? and what?'*. It provides maps and timelines for browsing of photos, and integrates photo annotation tools into the main browsing interface.

2.7 Summary

The availability of contextual metadata for personal photos means that different approaches are used in the management of personal photos compared with the organisation of general image collections A number of techniques have been proposed to automatically segment personal photo collections into the 'events' of which they are composed, where an event is an occurrence such as a birthday party These techniques generally rely on analysis of the time of photo capture, although they may additionally make use of image content and location information

A number of approaches have been proposed for the classification of people in personal photo collections. The approaches often use standard face recognition techniques, and sometimes enhance the face recognition feature with a 'body patch' feature that represents the clothes worn by a person This feature is only used to compare people occurring within a short time span of each other, typically on the same day or within the same event. Other approaches use spatial and temporal context of photo capture directly, in order to classify the people present in images, or they combine context-based approaches with content-based approaches.

Some photo management applications use image analysis techniques to automatically add annotations to photos, based on analysis of image content and the photo context metadata found in the EXIF header of digital photographs, and sometimes using a combination of image content and context.

Many systems provide tools for the manual annotation of personal photographs, with facilities for batch annotation to make the annotation process easier. Semi-

automatic annotation approaches combine automatic image analysis with manual annotation by suggesting labels for photos which the user can subsequently confirm or reject. Semi-automatic approaches to person classification suggest annotations for faces in photo collections, sometimes facilitating batch person annotation.

Interfaces for personal photo management systems typically sort photos by time, and provide calendars and other tools to browse based on time. A number of systems, assuming that location information is available, facilitate browsing photos using a map interface, while other interfaces allow navigation based on automatically extracted content-based and content-based features.

In the next chapter we will describe, in detail, the MediAssist system, a context-aware photo management system developed by the author of this thesis.

Chapter 3

A Prototype Context-Aware Photo Management System

In this chapter we describe MediAssist, a prototype context-aware photo management system for location-aware personal photos. The system incorporates automatic context-based and content-based analysis for the automatic annotation of personal photos with a number of concepts, in addition to providing semi-automatic tools which enable users to refine the output of the automatic annotation tools. We believe that this system represents the state of the art in terms of automatic organisation of, and retrieval from, personal photo collections.

This system was developed as part of the MediAssist project, funded by Enterprise Ireland over a three-year period, until September 2006. A number of researchers were involved in the project and in the development of the system. A number of the concept detectors used in the system were developed by other researchers working on the MediAssist project, and these are credited and cited where appropriate. The main implementation, and a most of the design and planning, of the demonstrator system for the MediAssist project was carried out by the author. In Section 3.1 we will outline in detail the exact contribution of the author of this thesis to the MediAssist system

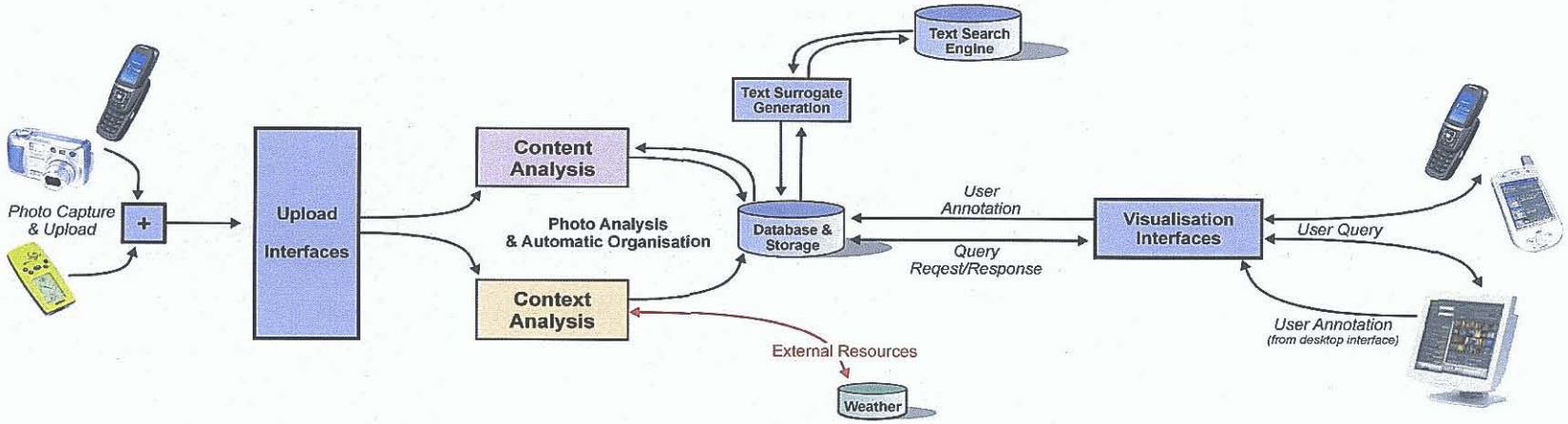


Figure 3.1: The architecture of the MediAssist prototype photo management system.

The architecture of the MediAssist system is illustrated in Figure 3.1. Various aspects of the system have been described previously in O'Hare et al. (2005b, 2006, 2007) and Gurrin et al. (2005). The system described in this chapter serves as a platform for the main research of this thesis in the area of person annotation, and it will be used as a testbed for our main research hypotheses in this area.

3.1 Chapter Outline and Author's Contribution

In this section we give an overview of this chapter and, given that the MediAssist project involved a number of researchers, we will outline the exact contribution of the author of this thesis to the system.

MediAssist Personal Photo Archive

Section 3.2 describes the MediAssist personal photo archive which is used as a testbed for all the research reported in this thesis The author was responsible for all aspects of collecting these photos from a number of users and managing these photos within the MediAssist system.

Context-based Photo Analysis

In Section 3.2 we describe the context-based indexing tools used in the MediAssist system. These tools include temporal indexing, location indexing, automatic event-detection and summarisation, light status classification, weather status classification and indoor/outdoor classification. The temporal indexing, location indexing, event-detection and summarisation, light status classification and indoor/outdoor classification modules were all developed by the author of this thesis. Weather classification was implemented by a colleague working on the MediAssist project Integration of all of these analysis tools into the MediAssist system was carried out by the author of this thesis.

Content-based Photo Analysis

In Section 3.4 we describe the content-based analysis tool used in the MediAssist system. These tools include face detection and building detection. Other content-based analysis tools will be described in Chapter 5. The face detection and building detection modules were developed by other researchers working on the MediAssist project. Again, the integration of these analysis tools into the MediAssist system was carried out by the author of this thesis.

Text Search

In Section 3.5 we describe text searching in the MediAssist system. A text search engine was developed by a colleague working on the MediAssist system. Work on creating the text surrogate documents (see Section 3.5) for each photo was carried out by the author of the thesis, as was work on the integration of the text search facility into the MediAssist system.

MediAssist Photo Management System Interface

In Section 3.6 we describe the MediAssist Demonstrator system for personal photo management, which provides tools for browsing, searching and semi-automatic annotation of personal photos. Design decisions relating to user interaction with the system, and the implementation of this demonstrator system, were carried out exclusively by the author of this thesis. Some design decisions were taken in consultation with colleagues working on the MediAssist system, however, and the look and feel of the interface was largely determined by a colleague with expertise in the area of user interface design

3.2 A Context-Aware Personal Photo Archive

It is reasonable to assume that, in the not too distant future, digital photo capture devices will be both location-aware and time-aware. The present batch of commer-

cially available photo capture devices, of course, do not support location-awareness and for correct time and date capture they generally require that the user manually set the correct time and date on the device.

In Chapter 2 we described some of the characteristics of personal photo collections which distinguish them from other image collections. In particular the user, as the creator of the photos, usually has a knowledge of all photos in their collection in addition to knowledge of the context of their capture. In order to create a large, location-aware collection of personal photos in the MediAssist system, we use an interim solution to location-stamping of photos, using World Wide Media Exchange (WWMX) tools proposed by Toyama et al. (2003). Our users carried a separate GPS device whenever they were capturing their digital photos with cameras or with mobile phones. It is then a trivial post-processing task to match photos with GPS tracklog points by matching the timestamps in the EXIF header of the photo from the camera with the timestamps of the GPS tracklog points. For cases where the user did not have a GPS device with them during photo capture, it was possible to retrospectively location-stamp these photos using a map interface. This combination of approaches allowed us to develop a large collection of location-aware photos in advance of the availability of location-aware devices, which we were able to use in the research reported in this thesis.

The users were asked to provide as many of their personal photos as they were willing to donate to the MediAssist system. Photos included photos taken on holidays and business trips, parties, trips to the zoo, photos taken during day to day activities etc. These photos include a mixture of photos taken of scenery and attractions on the end hand, and photos of people, both from events like parties and from day to day activies, on the other hand. The people present in the photos vary from friends and family of the users to their colleagues. Photos of poor technical quality were not removed, meaning that the collection include blurred photos, underexposed photos etc Privacy concerns mean that few of the users actually gave their complete, unfiltered, photo collections for use in the MediAssist system Other

User	Total Photos	Unique Countries	Unique States/ Counties	Unique Towns	Years Spanned
A	5,231	9	25	150	4
B	3,435	13	45	173	4
C	2,672	10	24	90	3
D	2,128	10	35	159	11
E	1,974	6	9	34	3
F	1,860	7	21	69	2
G	1,208	3	8	40	4
H	1,044	4	12	40	2
I	770	9	21	44	4
J	753	5	10	22	1
K	603	5	11	23	2
L	513	5	7	20	1
Total	23,774	32	145	764	11
Avg (top 12)	1,849	7.2	19	59	3.4

Table 3.1: Statistics for the 12 largest user photo collections in the MediAssist photo archive. The second row from the bottom describes the entire collection, which also includes photos from an additional 17 users. The average is calculated for the 12 largest user collections displayed.

than this filtering carried out by the users themselves, however, there was no pre-selection of photos chosen for the MediAssist archive, and we believe the MediAssist photo archive is representative of general personal photo collections.

We collected 23,774 time-aware and location-aware photos from 29 users, taken using 31 different camera and camera-phone models from 14 different manufacturers Of these users, 12 have collections of over 500 photos, with an average of 1,849 photos each for these users. These 12 users' collections cover an average of 7.2 countries each, and span an average of 3.4 years. This average time span is somewhat skewed by the 11 year collection of user D; without this user the average span is 2 7 years. More statistics about the collection can be seen in Table 3.1.

3.3 Context-based Photo Indexing

The basic timestamp, GPS location, and other EXIF (JEITA, 2002) context metadata associated with the capture of each photo is automatically expanded to include

a number of features which enhance the available metadata. In this section we will describe the automatic annotation of photos, based on the context of photo capture, as used in the MediAssist system. We develop the tools described in this section to explore the possibilities for the automatic organisation of personal photo collections based on the analysis of context by integrating these tools into the MediAssist system. This analysis is carried out as the photos are uploaded into the system, and the results are stored in both a relational database and a text search engine, as illustrated in Figure 3.1.

3.3.1 Temporal Indexing

Digital cameras automatically record the time of photo capture in the EXIF header of the photos (JEITA, 2002). Using time for indexing personal photos is commonplace in commercial products. People's memories of dates has been shown to be unbiased in the sense that inaccurately recalled dates are just as often before the correct date as after it (Wagenaar, 1986, Larsen et al., 1995). The time of autobiographical information is not always accurately remembered, however. Wagenaar (1986) shows that 'when' is quite poorly remembered compared with 'who', 'what' and 'where' for autobiographical events.

Work in cognitive psychology suggests that human memory of dates relies on a number of *cyclic temporal schemata*, in particular the year, week and day (Larsen et al , 1995). This theory states that the cognitive topology of time is not linear, like physical time, but based on a number of cyclic schemata. The year and day schemata are related to recurrent patterns in the natural environment. So the year is split into seasons and months based on natural cyclic changes in the climate (the month, of course, is a cultural construct, but months are closely related to seasonal changes). The week schemata, based on the day of the week, differs in that it is not based on cycles of nature, but rather it is a cultural phenomenon based on work patterns and and other cultural conventions in Western societies. The existence of these psychological schemata is supported by experimental evidence that shows that

time of day, day of week and month of the year of autobiographical events are all well-remembered by participants.

The MediAssist system exploits these temporal schemata by indexing time based on these cyclic temporal schemata, in addition to indexing by linear physical time. We expand the time of capture data to include a number of independent time-based fields: year, month, day of month, day of week, hour. Although there is no evidence that a month schemata exists (i.e. people do not have a schema which represents the date within the month (Larsen et al., 1995)) we include day of the month in this prototype because we believe that it can be useful for finding certain 'special events' where the date is more likely to be remembered (e.g the 25th of December for Christmas day, birthdays etc).

By exploiting these temporal schemata and indexing by time along multiple dimensions, users can search for photos when they only remember some of the temporal context of a photo capturing event. For example, they may only remember that an event occurred in the Summer on a certain day of the week, but they may not remember the year By indexing using cyclic temporal schemata the MediAssist system can support searching based on such partial recollection of the temporal context of an autobiographical photo capturing event.

3.3.2 Location-Based Photo Indexing

In order to process the raw latitude co-ordinates supplied by the GPS sensor into a more useful format, we use two publicly available gazetteers to convert these readings into placenames. The Geographic Names Information System (GNIS)[1] from the U.S Geological Survey contains information on places in the United States. The GEOnet Names Server[2] from the National Geospatial-Intelligence Agency and the U S. Board on Geographic Names provides information about places for the rest of the world. Between them, these resources constitute a database of over 7 million

[1] http.//nhd usgs.gov/gnis html
[2] http·//gnswww nima.mil/geonames/GNS/index jsp

locations covering the entire globe. They cover diverse location types, including populated places, rivers, mountains and a number of other location types In order to associate a place-name with a latitude/longitude co-ordinate we calculate the geographical distance, D_G, between two points on the earth's surface, using the following formula (Longley et al , 2001).

$$D_G(A,B) = R\left(\arccos\left[\sin(lat_A)\sin(lat_B) + \cos(lat_A)\cos(lat_B)\cos(lon_B - lon_A)\right]\right) \tag{3.1}$$

This equation assumes that the earth is a perfect sphere of radius R, in whatever unit of measurement we wish to use. We can use this equation to calculate the distance from a candidate photo location to all known locations in the gazetteer, and the place-name associated with a photo is the town/city with the minimum distance to the candidate photo location. We take the name, state/county, and country triplet from that location to provide a place name for that photo. This hierarchical structure allows for hierarchical location-based searching using our system.

This is a simple approach, and it makes simplifying assumptions, and it is possible to use resources that represent locations as a complex shape rather than a latitude/longitude co-ordinate, potentially allowing for more accurate labelling, though such resources are not freely available Our approach, however, uses freely available resources and provides useful location names at low cost for integration into our prototype system.

3.3.3 Event Detection and Summarisation

Event Detection

In Chapter 2 we discussed approaches to event detection in personal photo collections. In the MediAssist system we use the approach proposed by Graham et al (2002), which we will describe briefly here This method analyses the temporal dis-

tance between the capture times of consecutive photos within the user's collection. If this distance is greater than an empirically determined threshold, then a new event is deemed to have begun.

The internal structure of each event is then analysed to look for sub-events within each event. The average distance between consecutive photos within the event is calculated. The method uses the formula $Q_3 + \alpha(Q_3 - Q_1)$ to calculate a threshold for detecting outlier time-distance values within an event. Q_1 and Q_3 are the first and third quartiles, respectively. The first quartile is the value below which 25% of the data falls. The third quartile is the value below which 75% of the data falls. The value of α is determined empirically by optimising performance on a test set. If a time difference is greater than this threshold then it is considered an outlier, and is considered to signal the beginning of a new event. For example, if the average gap between photos in an event is 5 minutes, and we detect a gap of 1 hour, then this gap will be considered the start of a new sub-event.

Event Summarisation

Having detected events in photo collections, we then select five representative photos that can be used to summarise that event in the MediAssist user interface. There is little existing work on summarising events in personal photo collections, and there are a number of strategies that could be used to select the summary photos for an event. For example, information about the presence of faces could be used to ensure that photos of people are included in the summary, or that a balance of photos with and without people are chosen. In this work we take the simple approach of selecting a subset of the photos that gives a good temporal summary of the event. We believe there are avenues for interesting research in exploring alternative summarisation strategies and in determining user expectations in terms of such summaries, although such work is outside the scope of this thesis.

We use the distribution of the photo timestamps to select 5 summary photos for each event. The summary photos are selected as follows:

- If there are less than 5 photos in the event, then no selection is necessary and all photos in the event are included in the summary
- If the event includes more than 5 sub-events than one summary photo is chosen from each of the 5 largest sub-events Because the sub-events reflect the internal structure of an event, and because sub-events with more photos should be more important, we believe this will ensure a reasonable summary based on the temporal structure of the event. For simplicity, we choose the first photo from the event, while acknowledging that alternative approaches could be explored to choose a more representative photo for a sub-event
- For cases where there are less than 5 sub-events, then more photos will be selected from larger sub-events, and we calculate the number of summary photos chosen from each event as the proportion of photos from the event that are in the sub-event. For example, if an event contains 20 photos and a particular sub-event within that event contains 16 photos, then 4 of the summary photos will be taken from that sub-event. When choosing photos from within a sub-event, we take photos evenly spaced temporally throughout the duration of the event. For example, if we were to choose three photos, then we would choose the first and the last photo in the sub-event, along with the photo taken closest to the middle of the sub-event.

The summary photos chosen represent the default summary of an event in the user interface. For calculating the representative place-name for an event, we use the same approach for labelling a photo, with an event place-name corresponding to the place-name of the first photo in the event

Query-Dependent Summarisation

In Section 3.6.2 we will describe search tools in the MediAssist system. For certain types of search query none of the summary photos for an event will be relevant to the query. For example, if a user searches for a particular person who occurs 5

times in an event containing over 100 photos, then it is highly likely that none of the summary photos for the event will contain this person.

To overcome this problem, we can select our representative photos in a query-dependent manner if all photos in the event are not relevant to the query. We do this by replacing each summary photo that is not relevant to the query with the temporally closest relevant photo This ensures that the summary includes photos relevant to the current user search while still ensuring a temporal summary of the event.

3.3.4 Light Status Classification

In Section 3.3.1 we described the day temporal schema used by people for their cognitive representation of time. Closely related to this is the light status at a given time; that is, whether the lighting conditions were daylight, darkness, dusk or dawn when a photo was taken Given that we know the date, time and location of photo capture event it is possible to calculate sunrise and sunset times for the location in question on any given date. In the MediAssist system we use standard astronomical algorithms (Seidelmann, 1992), allowing us not only to calculate sunrise and sunset times for day and night classification, but also to correctly classify dusk and dawn by using standard techniques to calculate the beginning and end of *twilight*, which can last for minutes or hours depending on the location and the time of year Sunrise and sunset are considered to be the times when the upper edge of the disk of the is on the horizon. The periods of twilight are the intervals of time when the sun is below the horizon but natural light from the upper atmosphere is reflected towards the earth's surface, providing illumination *Civil twilight* is the limit at which twilight illumination is sufficient for objects to be clearly distinguished, and this is what would conventionally be considered twilight Civil twilight occurs when the sun is less than 6 degrees below the horizon (Seidelmann, 1992). *Nautical twilight* and *astronomical twilight* provide alternative definitions of twilight, but we do not consider them here. Thus we calculate the times for sunrise, sunset, the beginning

of morning twilight and the end of evening twilight, and it is then a trivial matter to classify the light status at any time as follows:

- *Daylight* begins at the time of sunrise and ends at sunset.
- *Dusk* begins at sunset and ends at the end of the twilight period when the sun falls 6 degrees below the earth's surface.
- *Darkness* begins after twilight ends when the sun is below 6 degrees. It ends when at the beginning of the morning twilight period, when the sun is again 6 degrees below the earth's surface.
- *Dawn* begins with the beginning of the morning twilight period, and ends at sunrise.

Naaman et al (2004a) previously proposed light status classification by querying an online resource for sunrise and sunset times for a given time and location. They consider dusk and dawn to last 1 hour before and after sunrise and sunset, respectively, a figure which is chosen somewhat arbitrarily Our approach has the advantage the it accurately calculates start and end times for twilight, reflecting the fact that the duration of the twilight periods changes depending on the location and the time of the year. Also, because we implement the algorithm directly, we do not need to query an external resource in order to determine light status.

3.3.5 Weather Status Classification

To extract the weather status at the time of photo capture, we use the approach proposed by Naaman et al. (2004a). There are approximately 10,500 international weather stations distributed over the earth which constantly log weather data, and these logs are made available by the Weather Underground website.[3] We find the nearest weather station to each candidate image, and then extract the record from the log of the weather station closest to the time of photo capture. This record

[3]http //wunderground.com/

is selected as the best representative of the image, and the image is annotated as occurring during these weather conditions. We classify the weather as belonging to one of the following four categories: rainy, clear, snow or overcast/cloudy.

3.3.6 Indoor / Outdoor Classification

Indoor/outdoor classification is inferred from EXIF camera setting metadata, stored by the digital camera when capturing a picture, which reflect the ambient light levels when a picture was taken Our classification technique is based on the assumption that, for photos taken during daylight (which we calculate automatically, as described in Section 3.3.4), the sun is a much stronger source of illumination than the artificial lights which light an indoor scene. Since some of the recorded EXIF tags can be seen as functions of ambient light levels it is possible to use them to infer the scene brightness during photo capture. We use 5 EXIF header fields in our indoor/outdoor classification process

- *Brightness* is the brightness level of the subject to be photographed.
- *Shutter Speed* is the length of time the shutter is kept open during the photograph capture.
- The *Aperture* is the size of the opening of the lens. It is measured in f-stops, which is the ratio of the focal length of the lens to the diameter of the lens diaphragm opening. The focal length is the distance from lens to its focal point, which, roughly defined, is the point where the light from the lens converges to form a sharp image.
- The *ISO Speed* is a standard measurement that indicates the sensitivity of film or electronic sensors to light A higher ISO speed means less light is required for an image to be recorded.
- *Flash Fired* indicates whether the flash was fired during photo capture or not.

Since our indoor/outdoor classification is based on environmental light levels, the most useful of these metadata tags is brightness because it measures this directly, although this is unfortunately not recorded by all commercial cameras. On the other hand, all models have shutter speed, aperture and flash, and most have the ISO speed. Using a combination of shutter speed, aperture and ISO Speed, we can calculate a value called the Exposure Value, which is a value given to all combinations of camera shutter speed and aperture that give the same exposure. An equivalent Exposure Value implies equivalent ambient light levels. We use an alternative[4] to the standard formula (Ray, 2000) which takes the ISO speed into account:

$$Log_2\left(\frac{aperture^2}{shutterspeed}\right) + Log_2\left(\frac{ISOSpeed}{100}\right) \tag{3.2}$$

If the ISO Speed is unknown we assume an ISO Speed of 100, which in turn is the ISO rating of 'standard' film. This Exposure Value is a function of the environmental light and the additional light provided by the camera flash when the photograph was captured. After we calculate the exposure value we have two single-valued measures that reflect the ambient light levels at the time of photo capture Brightness and Exposure Value. If Brightness is available we use this value to represent the ambient light levels, otherwise we use the Exposure Value. Whichever value we use, we compare it to an empirically determined threshold to classify an image as indoor or outdoor.

This is a simple approach and alternative approaches may yield improved results Boutell and Luo (2005), for example, combine EXIF metadata information with content-based image analysis for indoor/outdoor classification using a bayesian approach. Their work showed that contextual information alone is better for distinguishing between indoor and outdoor scenes compared with content-based image analysis techniques alone. Combining image content with context gives a small improvement compared with context alone We use the simple approach described

[4]alternative formula. http //en.wikipedia.org/wiki/Exposure_value

here because it gives us a low cost method of classifying images as indoor/outdoor, and because the work of Boutell and Luo (2005) has already shown that EXIF metadata can be used to give good results for indoor/outdoor classification. Our approach also has the advantage that it exploits photographic knowledge by using well understood photographic techniques for measuring exposure based on camera settings, rather than relying on a statistical framework to analyse and combine these camera settings, and it gives us a simple context-based approach that we can easily integrate into our prototype system. An interesting approach could be to combine our context-based processing which calculates Exposure Value with the approaches proposed by Boutell and Luo (2005) in order to compare the performance of their unprocessed representation of camera settings with our Exposure Value-based representation.

3.4 Content-based Photo Indexing

In addition to using tools to analyse the context of each photo, we also have a number of analysis tools which focus on the visual image content of the photos. These tools were developed by other researchers working on the MediAssist system and they also create automatic annotations of all photos in the MediAssist collection. As with context-based analysis, content-based analysis is carried on all photos as they are imported into the system, and the results are stored in both a relational database and a text search engine. The person analysis tools can also be triggered by user input in the form of annotations. This is illustrated in Figure 3.1 and page 26.

3.4.1 Face Detection

The face detection technique used in the MediAssist system was developed by Cooray et al (2006), and extends the Bayesian Discriminating Feature (BDF) model for detecting frontal faces in grey-scale images (Liu, 2003). The extended model exploits

the colour feature using a statistical skin detection model. The skin segmentation model first creates a skin mask which is subject to erosion/dilation morphological operations to remove noisy skin pixels, while expanding the face and cheek areas which are considered regions of interest. Both the original image and the corresponding skin mask are rotated to enable detection of rotated faces. They are then iteratively scaled by a pre-defined factor. The Bayesian decision rule, which was defined in the BDF model based on face/non-face error, classifies a region to be a face or non-face. The exact face location is decided by a modified single response criterion by searching for the best matching sub-image within a pre-defined search area.

3.4.2 Person Analysis

The system uses a number of context-based and content-based tools to facilitate the identification of people in photos, where they have not yet been annotated by the user. The techniques used to support this are based on context-based and content-based analysis, and are described in Chapter 5. These techniques can be used to suggest names for unknown faces encountered by the user as they navigate their personal photo collection. The interactive tools that enable semi-automatic annotation of photos with the MediAssist system are described in Section 3.6.3.

3.4.3 Building Detection

A building detection module, developed by Malobabic et al. (2005), in the MediAssist system detects the presence of a large buildings in photos, where the building is either the single dominant object or one of a few dominant objects. The approach exploits the fact that the presence of a dominant human-made object in a scene generates strong evidence in the form of straight or elliptical line segments and edges. We also assume the presence of contextual information in the form of an indoor/outdoor label. Given that there is huge variation within the building class

in terms of possible shapes of different types of buildings, we take the view that a coarse modeling of building shape/geometric properties is appropriate. Edge orientation histogram-based features are extracted at three different scales, both globally and locally, in order to capture the general shape information of a building. Assuming that semantically important concepts, such as large buildings, occur around the centre of the scene, we also incorporate the localised edge information pertaining to the central 25% of the image thus forming a 24-dimensional feature vector to represent each image. We then train an SVM-classifier (Burges, 1998) to distinguish between the building and non-building classes using this feature vector. This approach to building detection is described in more detail by Malobabic et al. (2005), who developed this approach to building detection for the MediAssist system.

3.5 Text Indexing Using Content and Context

Text-searching is a tool which is widely deployed in popular information retrieval search engines such as Google. Use of text queries for search is intuitive and popular with users, giving a simple way to express an information need. To provide text-based search in the MediAssist system, the automatic context- and content-based features described in Sections 3.3 and 3.4 are mined to construct text surrogates for each photo. A text surrogate is a text file which acts as a surrogate representation of the photo, containing a textual description of the concepts that have been associated with the image by the system. This text surrogate can then be indexed by a search engine to allow for text search. For each feature associated with a photo, a textual representation is written into the surrogate document. An example extract from a text surrogate could be 'dublin ireland september sept weekend afternoon person alan', representing the features location, time, face detection and person annotation. The text search of images we support in this manner is different from existing systems such as Google Images[5] where the search features are based on adjacent

[5] http://images.google.com/

text, rather than analysis of the images and their associated metadata. We also create text surrogates for 'events' to allow for text-based searching of events in the 'Event List' view described in Section 3.6.1.

Our text search engine is based on the standard BM25 text retrieval model proposed by Robertson et al. (1995), which has proven to be an effective text retrieval model. The system has no need to remove stop-words because the surrogate documents are created automatically by the MediAssist system, so the documents contain no stop-words that need to be removed. The system presents a text search box to allow for the quick and easy formulation of text queries based on the automatically extracted content and context features.

In previous work we evaluated the utility of this text search facility using known item search (O'Hare et al., 2007). We showed that the extracted features create a powerful search index, ranking the relevant image in a known item search highly even if we use only a subset of the context-based features.

3.6 The MediAssist Photo Management System Interface

Based on the tools described in the previous sections, the web-based, multi-user, MediAssist prototype system provides tools for browsing and searching personal photograph collections, in addition to tools for semi-automatically annotating photos by confirming and correcting automatically derived annotations. The photo upload tools used in the system, which perform the analysis described in Sections 3.4 and 3.3, are implemented in Java. All camera metadata, and additional automatic annotations are stored in a Microsoft SQL Server Database. Text surrogates for all photos are indexed by a conventional search engine, as described in Section 3.5. In addition, the system uses a number of Java server applications to perform certain tasks in response to user input, for example updating the text search engine after user annotations. The front end is implemented as a HTML application, using

PHP 5 for server-side programming and communication with the SQL database and search engine. Javascript is used for client-side scripting. The system architecture is illustrated in Figure 3.1 on page 26. The main interface screen is shown in Figure 3.2. In this section we describe the search tools and the semi-automatic annotation tools provided by the system.

3.6.1 Photo Browsing Tools

When browsing personal photos using the MediAssist system, a number of views are available to the user In this section we describe each of these views.

Event List View

The *Event List* view is the default view, which a user will see when they log in to the system. It presents an event-based summary of the user's entire photo collection, or of a subset of their collection specified by a user search using the search tools described in Section 3.6.2. This view is enabled by the event detection described in Section 3.3.3. In the *Event List* view, shown in Figure 3.2 on page 45, all events are presented in reverse chronological order. Each event is summarised by a label and five representative thumbnail photos. We use the summary photos selected during the event detection process, with a query dependent set of summary photos chosen if not all photos in the event are relevant to the current query (see Section 3.3.3).

The label for each event is a combination of the location of the event, with the town, county and country displayed, and the date that it occurred. The total number of photos in the event is also shown and, if not all photos are relevant, then the number of relevant photos is also displayed. For example, if an event contains 10 photos, all relevant to the query, then the text will say '10 photos'. If only 5 were relevant then the text would say '10 photos, 5 relevant'. There is also a 'View All' button, which accesses the *Event Detail* view for this event. Clicking on an event label also accesses the *Event Detail* view. Clicking on an image will access the *Photo Detail* view for that photo.

Figure 3.2: The MediAssist Photo Management System Interface, with photos displayed in the Event List view.

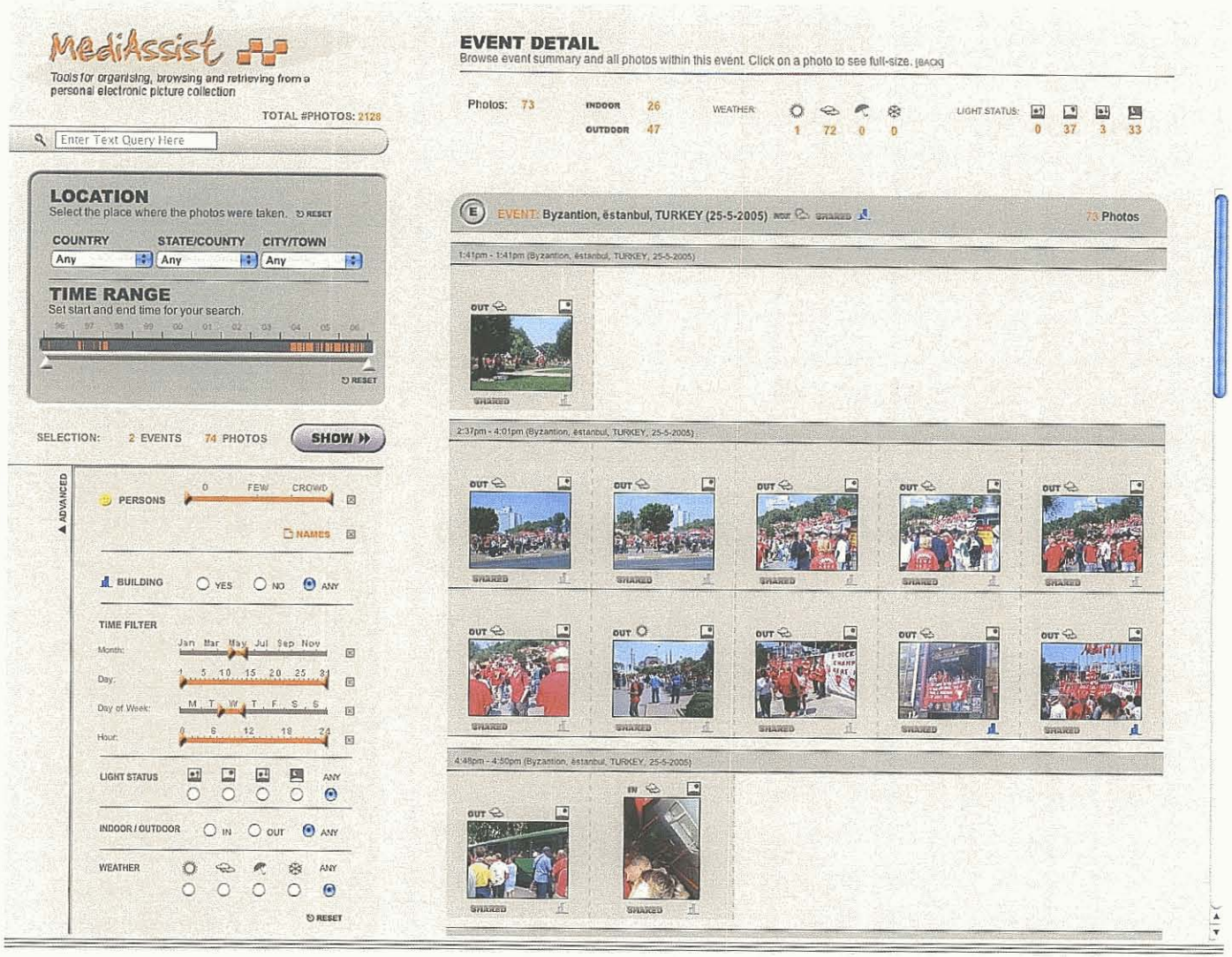


Figure 3.3: Event Detail view in the MediAssist interface.

For large collections of personal photos, we believe that the *Event List* gives an efficient summary of the users collection, allowing them to browse very large collections easily.

Event Detail View

The *Event Detail* view is presented when the user selects an entry in the Event List, and is composed of the full set of photos in the event. This view organises the event into sub-events, with a separate heading, labelled by the start time and location, for each sub-event. This view can be seen in Figure 3.3.

Individual Photo List View

The *Individual Photo List* view is an optional view where thumbnail-size photos are presented without any particular event grouping, but sorted by date/time It can be selected as an alternative to the *Event List* view when a user would prefer to see all photos on screen at the same time rather than being presented with an event-based summary. This view can be seen in Figure 3.8 on page 55

Photo Detail View

The *Photo Detail* view is an enlarged single photo view presented when the user selects one of the thumbnail size photos in any of the above views Arrow buttons allow jumping to previous/next photos in this view. This view also includes tools for semi-automatic annotation of people in photos, as described below in Section 3.6.3. This view can be seen in Figure 3.10 on page 58.

Photo Summary View

In the *Event List*, *Individual Photo List* and *Photo Detail* views there is an area at the top of the screen which summarises the current group of photos. This view is called the *Search Summary* if seen as part of the *Event List* or *Individual Photo List* views. If seen as part of the *Event Detail* view, it is called the *Event Summary*.

In addition to displaying the number of photos and the number of events, this view gives information about the number of indoor or outdoor photos, the number of photos taken during each of the classes of weather conditions, and the number of photos taken during various light status conditions. Giving feedback about the distribution of photos like this should help direct users during search, and can also function as a type of query preview (Doan et al., 1996) as it allows users to see which categories would be most efficient in narrowing their search query. This view can be seen in Figures 3.2 and 3.3: in Figure 3.2 we can see a 'Search Summary' which summarises the results of the current search, while in Figure 3.3 we can see an 'Event Detail' panel which summarises the current event.

3.6.2 Photo Searching Tools

The analysis tools described in Sections 3.3 and 3.4 above create a powerful automatic annotation of the photos in the MediAssist collection, which can then be used for searching for specific photos or specific groups of photos. The MediAssist prototype system provides powerful searching through personal photo collections based on this analysis. The system provides two main search interfaces: a filter-based interface for creating structured queries, and a simple text search interface.

Filter-based Searching

The system provides an interface to allow the user to easily create powerful queries based on all of the features described above. The query interface contains two main sections: a basic query panel and an advanced panel.

The basic search panel includes a text box, location filters and a time range filter, and can be seen in Figure 3.4 on page 49. A location-based search can be quickly formulated using drop-down boxes corresponding to country, state/county and city/town. The drop-down lists are populated with all locations found in the user's collection. The time range filter uses sliders to allow the user to select a time range within their collection. The time range bar segments the collection into

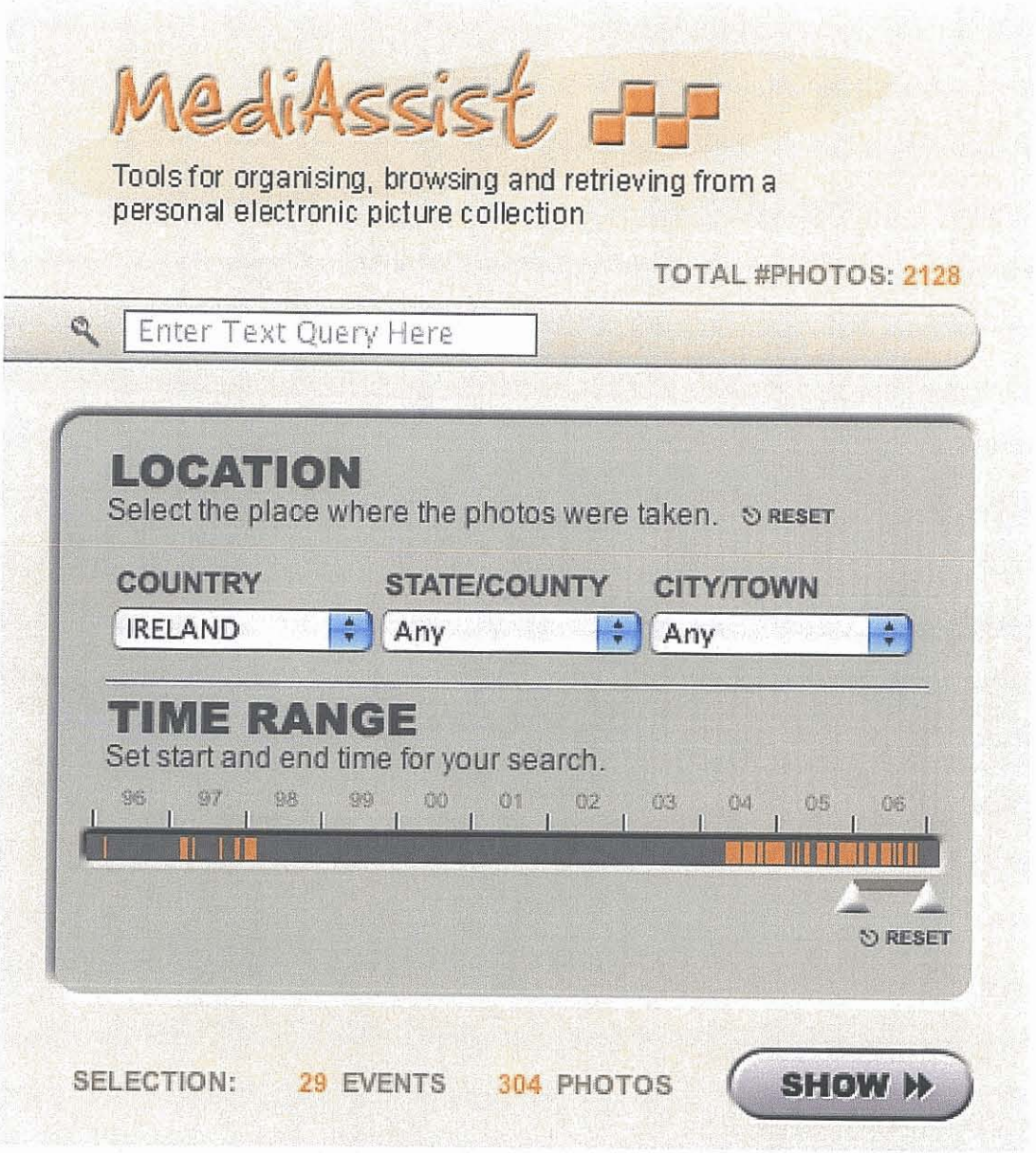


Figure 3.4: Basic search filters in the MediAssist system interface. In this example the user is searching for photos taken in Ireland in 2006. The query preview mechanism indicates that there are 304 photos, from 29 events, relevant to this query.

years, and clicking on a year label automatically restricts the selection to that year. Also, orange strips on the time range bar indicate times where photos were taken, summarising the temporal distribution of photos throughout the collection.

The advanced search panel is hidden by default. Clicking on the 'advanced' button reveals this panel (the 'advanced' button can be found below the basic search filters and is visible in Figures 3 2, 3.3 and 3 5), which features advanced search filters including time filters and filters based on the system's context-based and content-based analysis tools. indoor/outdoor, buildings, people and light status. Slider bars allow for advanced time-based queries to be formulated, based on the *cyclic temporal schemata* described in Section 3.3.1. These slider bars facilitate the easy formulation of time range queries allowing the user to specify, for example, a range of months or a range of days. For each of the filters the bright orange bar between, or outside, the slider icons indicates the selected range. The slider icons are oriented to indicate which side of the slider bar is being selected For the indoor/outdoor, weather and light status features, radio buttons allow the user to choose between mutually exclusive classes Figure 3.5 on page 51 shows the advanced panel with December to February, Saturday to Sunday, darkness and clear weather selected.

The user can also search by specifying the number and identity of faces present in an image. The person identity annotations are created in a semi-automatic fashion, as described in detail in Chapters 4 and 5. We do not discuss the similarity measures used, or the methods for combining these features with other sources of evidence, in this chapter. Such discussion is left until Chapter 5. The interactive person annotation tools used are described in Section 3.6.3. In the system interface, slider bars allow the user to specify the number of faces wanted in a search. Clicking on the 'names' button opens a name query pop-up, which allows the user to tick the names of people they want to search for. The person list can be sorted alphabetically or by frequency of occurrence, to facilitate finding desired people more easily A further option allows the user to specify if 'any' or 'all' of the names ticked should be found in relevant photos. The 'all' option means that only photos containing all of the

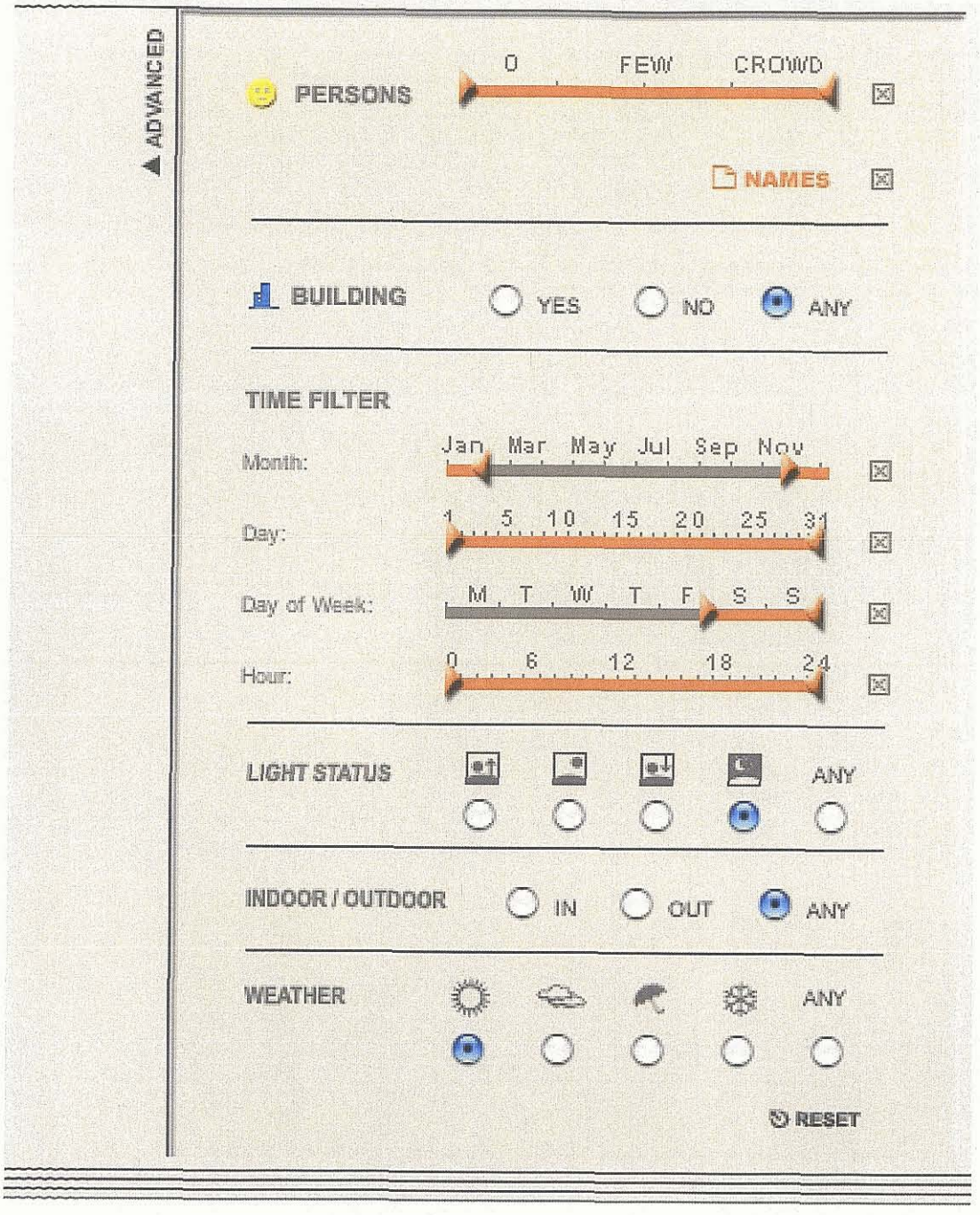


Figure 3.5: Advanced search filters in the MediAssist system interface.

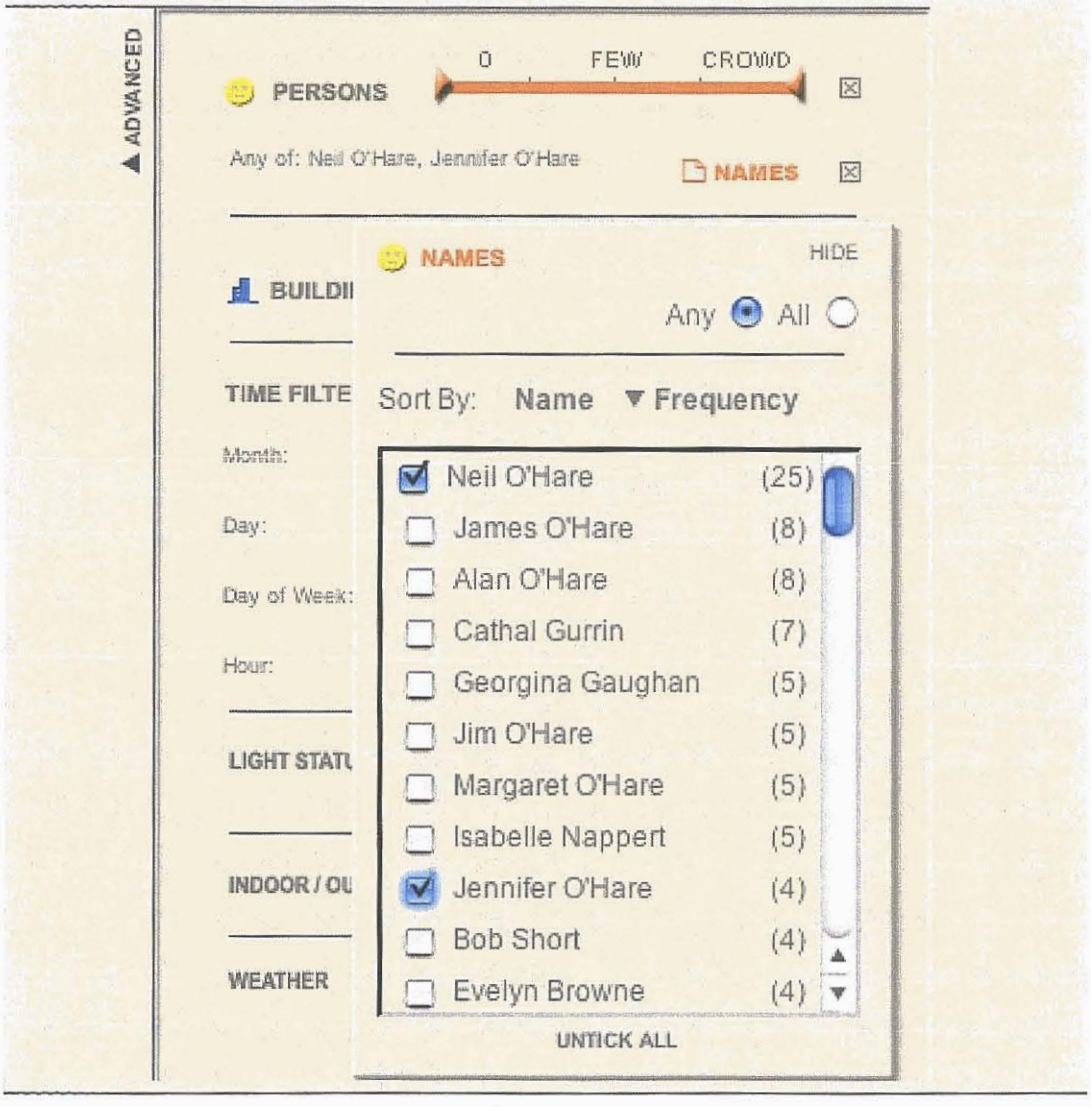


Figure 3.6: People search tools in the MediAssist interface.

selected names should be considered relevant. The 'any' option indicates that if a photo contains any of the selected names then it should be considered relevant. The name query pop-up can be seen in Figure 3.6 on page 52

A simple form of query preview (Doan et al., 1996) is provided to enable the user to see, before the system retrieves any photos, the effect of any alterations to the query. It does this by displaying, beside the 'Show' button, the number of relevant events and photos for the current query. Query preview is especially useful in our system where the total number of photos in the collection is very large (some users have over 3,000 photos), and can be seen in Figure 3 4 on page 49. When the user is happy with the query, the 'SHOW' button is clicked to display the relevant photo set on the right side of the screen, using one of the views described in Section 3.6.1.

Text-based Searching

The system also presents a text search alternative to filter-based searching, enabled by the creation of text surrogates for each image and the use of a standard text search engine, a process described in Section 3.5. The user can type a text query into the query box As with filter-based searching, query preview provides feedback on the number of relevant documents to the current query. The query preview is updated as the user types, so that the user gets immediate feedback about how many images are relevant to the current text query

There are two modes for the text-based search. The default *Text* mode ignores the current state of all other filters in the interface and performs a pure text search *Text + Filters* mode first filters the collection based on the search filters currently in use, and then applies the text search to this subset of the user's collection. Because the text search system uses the BM25 probabilistic approach to relevance, as opposed to the binary relevance model implied by the filter-based search system, it is possible to rank by relevance if a text search is used. The MediAssist system provides the user with the option of sorting text-based searches by relevance or by time. We give the user the option of sorting by time because it has been shown that users

Figure 3.7: Text search in the MediAssist interface.

are comfortable with using time-based sorting to orient themselves within their own personal photo collections (Rodden and Wood, 2003). The text search interface can be seen in Figure 3.7, where an example query is shown with the *text + filters* option selected, so the system will only return photos containing 'Neil' which were taken in Ireland in 2006. If the *text* option was selected then the filters would be ignored and all photos containing 'Neil' would be returned.

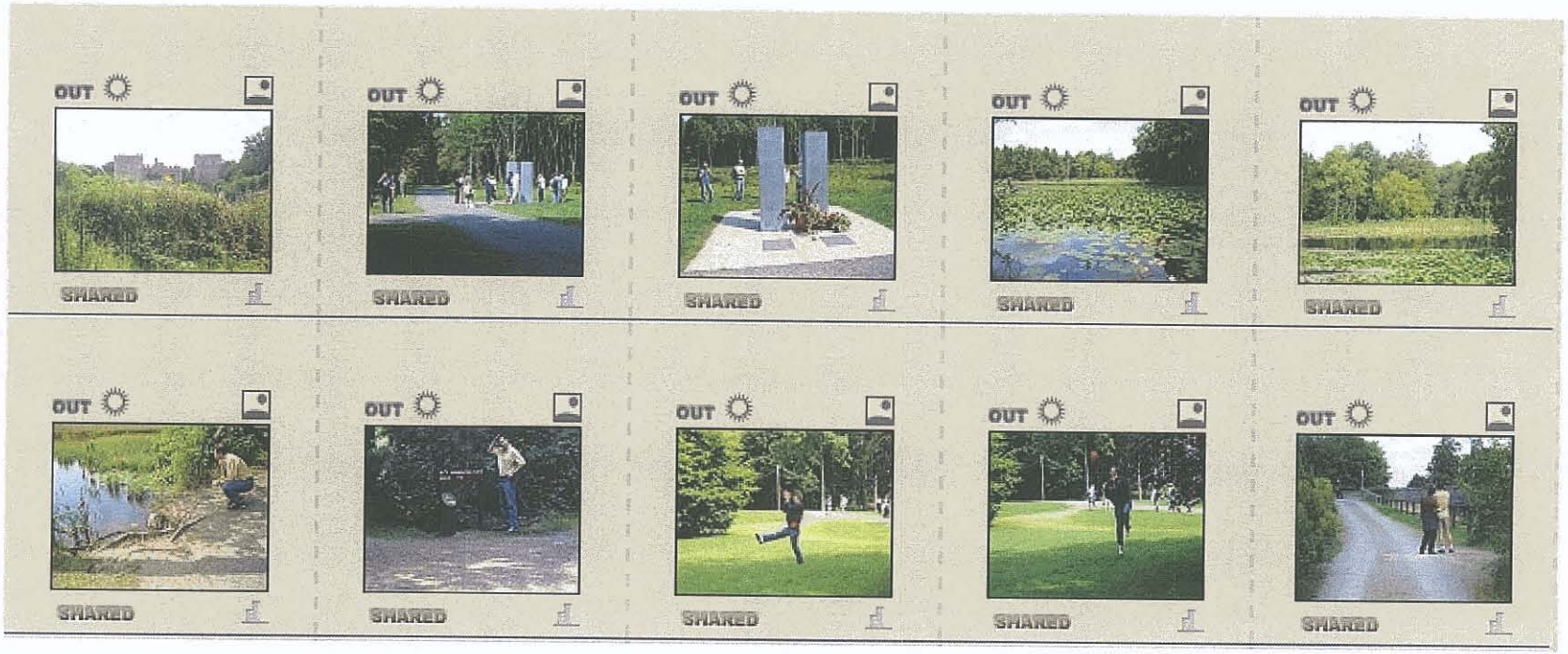


Figure 3.8: Thumbnail images in the Individual Photo List view, surrounded by feedback/annotation icons, in the MediAssist system interface.

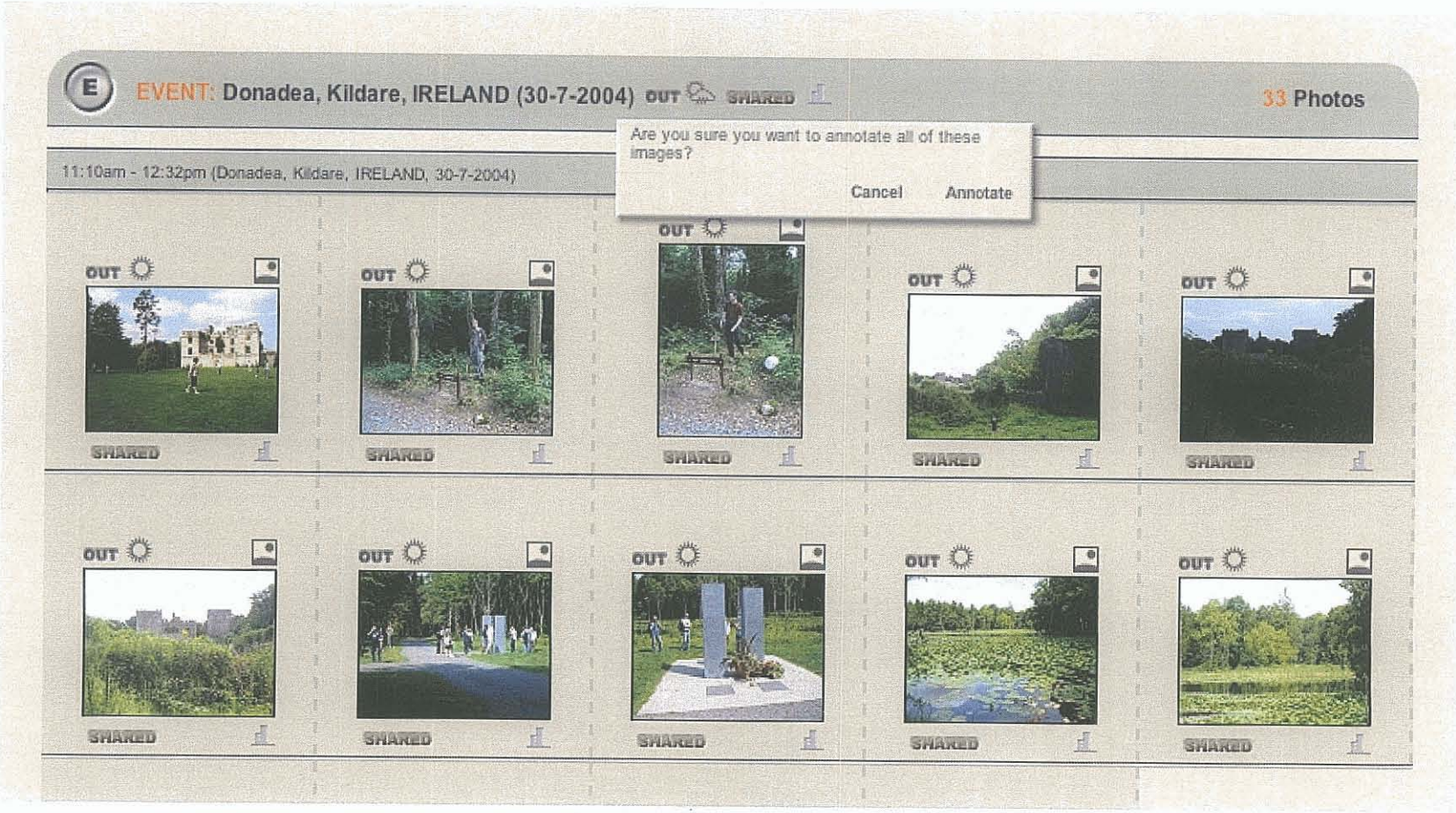


Figure 3.9: Group annotation icons in Event Detail view, in the MediAssist interface. The group annotation icons can be seen in the event heading, beside the time and location of the event.

3.6.3 Semi-Automatic Annotation Tools

In this section we describe the semi-automatic annotation tools available in the MediAssist system. In Chapter 4 we will look in more detail at tools specifically developed for the annotation of people in the MediAssist system.

Icon-Based Annotation

In all views, icons around the photos give feedback about the following automatically extracted features: indoor/outdoor, weather, light status and building detected. There is also a 'privacy' icon indicating whether the user wishes to share this photo or not. These icons can be seen in Figure 3.8. The indoor/outdoor, weather and light status icons can be found above the image: in the example in Figure 3.8 each of the thumbnail photos has an 'out' icon indicating outdoor, a 'sun' icon indicating sunny weather and a 'daylight' icon for the light status. The privacy and building icons are underneath the thumbnails, with all of the photos in Figure 3.8 showing a 'shared' privacy setting and a 'no building' icon. The 'no building' icon takes the form of a grey, inactive, building icon: a blue, active, building icon indicates that a building has been detected.

For the indoor/outdoor, weather and building features it is quite possible that the system annotation is incorrect. Clicking on the icon beside the image will correct these annotations Similarly, in the event view, icons beside the event label indicate the status of each feature for the entire event. For the case of a mixture of indoor/outdoor images, for example, the icon will reflect this using a 'mixed' icon Clicking on the group feedback/annotation icon will change the annotation of this feature for all images in the event, giving the user a very efficient way to annotate a number of images at once. For group annotation, the user must confirm that they want to annotate multiple images at once. These feedback/annotation icons can be seen in Figures 3.8 and the group feedback/annotation icons can be seen in Figure 3 9.

Figure 3.10: Face detection and person-annotation in the MediAssist system interface.

Semi-Automatic Person-Annotation

The system provides semi-automatic tools for annotating the names of individuals present in photographs. In the *Photo Detail* view, the system can highlight all detected faces, drawing a rectangle around the face area. Missed faces can be added by simply drawing a box over the missed face. Drawing a new face in this manner also triggers analysis of the image, using techniques that will be described in detail in Chapter 5. Conversely, incorrectly detected faces can be quickly removed by clicking an 'x' icon under the face. The system will use person classification techniques to suggest names for unknown faces based on comparisons with known, annotated, faces. We discuss the similarity measures used, and the techniques for

combining these features with other sources of evidence for the purpose of identity name suggestion, in Chapter 5.

When a user moves the mouse over a face in the *Photo Detail* view, the top-ranked name suggestion is displayed The user can confirm that this is correct by clicking on the name. If the top-ranked choice is incorrect, the user can choose from a shortlist of 10 names ranked in order of confidence. Otherwise clicking the 'more' button will display all known names. If the current face to be annotated is not known to the system, the user can annotate the face with a new name. Once a face has been annotated with an identity, it becomes one of the known faces, and future name suggestions will reflect this new system knowledge Figure 3 10 shows the semi-automatic annotation interface.

Semi-automtic person-annotation in the MediAssist system will be described in more detail in Chapter 4, where we will also propose enhancements to this interface that will allow for semi-automatic batch person-annotation.

3.7 Summary

We have described the MediAssist prototype personal photo management system. The system analyses personal photographs using a number of context-based and content-based analysis tools. The context-based analysis converts the location co-ordinates to a place-name, and the time is represented both in terms of physical linear time and in terms of the cyclic temporal schemata by which people remember temporal information. The schemata we use for temporal indexing are month of the year, day of the month, day of the week and hour of the day. In addition to these basic contextual features, the system also performs further analysis of the contextual information by performing light status classification in terms of day, night, dusk or dawn, indoor/outdoor classification, and weather status classification. A user's photo collection is also automatically segmented into 'events' based on the contextual information. In addition to the context-based analysis, the system also

indexes photos using content-based analysis tools A face detection engine detects all frontal faces in the photos. These detected faces are then analysed by person classification techniques Finally, a building detection module detects the presence of large buildings in the photos

In addition to being stored in a relational database, the result of this analysis are also written to surrogate text files, with one surrogate for each photo in the collection These text files are indexed by a conventional text search engine to allow for conventional text search of personal photo collections, without the need for user annotation.

The MediAssist prototype system interface provides tools for browsing, searching and semi-automatic annotation of personal photo collections The browsing tools provide a number of views of the photos, including an event-based view which summarises events using a query-dependent subset of the photos in the event Searching tools provide filters to create powerful searches based on the results of the automatic analysis, including the ability to intuitively specify time ranges along a number of independent dimensions simultaneously. The semi-automatic annotation tools allow users to correct any errors in the automatically created annotations, with the option of batch annotation to annotate multiple images simultaneously. Face detection detects faces in photos, while person analysis tools (described in detail in Chapter 5) are leveraged to suggest names for people present in the photos, which the user can then confirm.

We believe that this system represents the state of the art in terms of the use of both content-based and context-based methods for the management and organisation of personal photo collections, and this system represents a platform for our further research on person-annotation in personal photo collections. In the next chapter, Chapter 4, we will propose extensions to the MediAssist interface to allow for semi-automatic batch annotation of people in personal photo collections. In Chapter 5 we will propose techniques for person classification and retrieval in personal photo collections. The MediAsssist system, as described, supports retrieval

of user-confirmed identities The proposed framework also facilitates efficient batch annotation by retrieving unconfirmed annotation suggestions in response to person queries, providing an efficient way for users to annotate people identities in their photo collections in a semi-automatic manner

Chapter 4

Proposed Approach to Semi-Automatic Person-Annotation

In the previous chapter we introduced the MediAssist system for the management of personal photo collections. In this chapter we focus on the annotation of people in personal photo collections, and we look in more detail at semi-automatic person-annotation within the MediAssist system. We propose an extension to MediAssist that allows for semi-automatic batch person-annotation in personal photo collections. We do this by having the system suggest, at appropriate moments as the user interacts with the system, a set of faces to be annotated with a specific candidate person name. We then describe the implementation of the proposed semi-automatic person-annotation approach in the MediAssist user interface.

4.1 Requirements for Person-Annotation

In developing an approach for the semi-automatic annotation we identify some requirements for a suitable approach, which we summarise below.

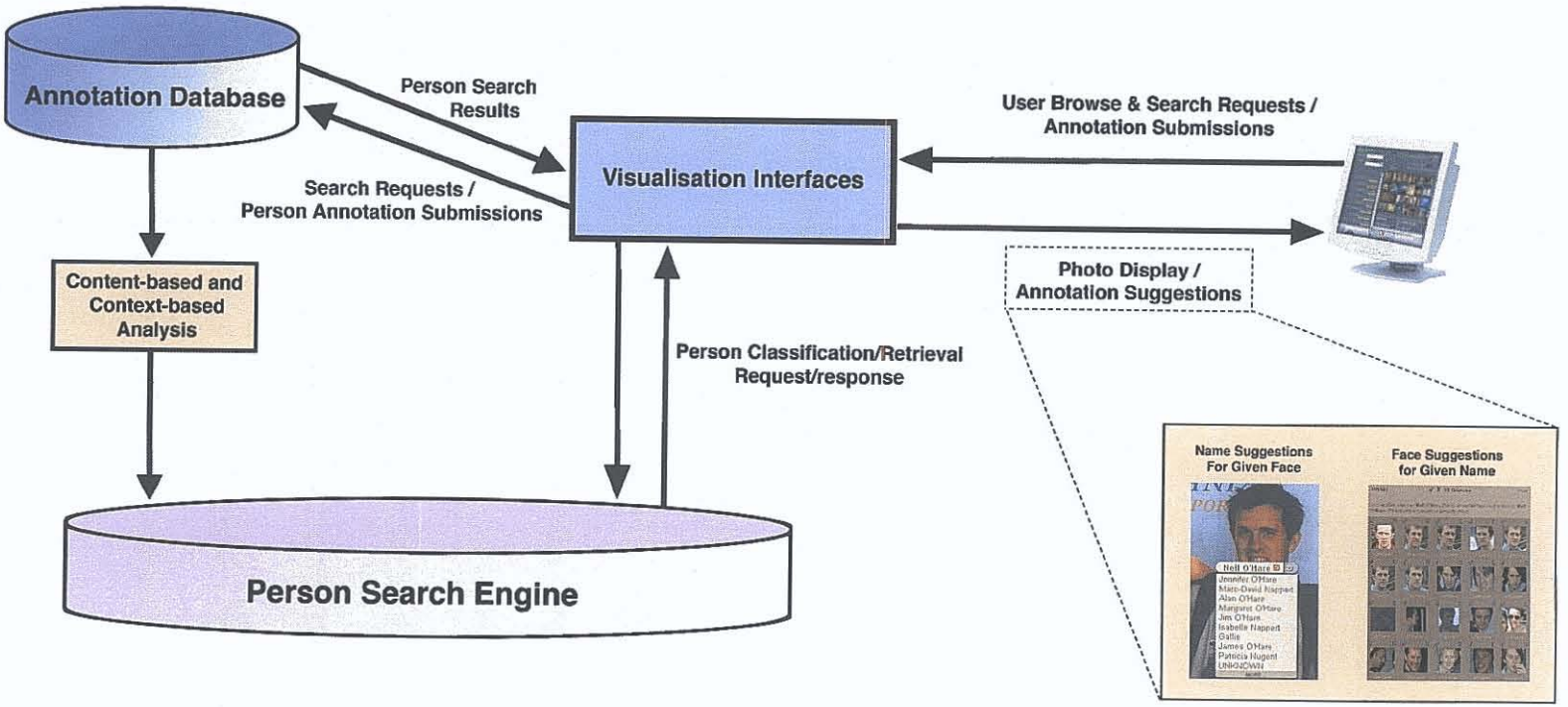


Figure 4.1: Model for semi-automatic person-annotation.

- Annotation tools should be seamlessly integrated into the main user interface In other words, the annotation tools should be naturally available to the user at all times, and the user should not need to switch to a special 'annotation mode' to be able to begin annotating. If this is not possible, then the annotation tools should be made available to the user at appropriate moments as they perform their normal photo browsing and searching tasks
- The interface should not force users to make annotations, but should facilitate such annotations should be user wish to make them.
- Bulk annotation should be facilitated, whereby users are able to annotate a number of faces at the same time in a batch manner
- Negative annotations should be facilitated, whereby users can easily reject suggested annotations if they are incorrect. This will ensure that the system will not continue to suggest the same incorrect faces for a given name if the users ignores them.

4.2 Model for Semi-Automatic Person-Annotation

Figure 4.1 summarises our proposed approach to semi-automatic person-annotation. The user browses through their collection using the MediAssist system interface, as described in Chapter 3. The system will prompt the user to provide names for any of the automatically-detected faces in their collection. When the user submits the annotation for a specific face, this annotation information is stored in an *annotation database*. This will allow the user to find the annotated face the next time they search for that particular person. In addition, the annotations are analysed by a *person search engine*, which will allow the system to suggest annotations for faces that have not yet been annotated. The system interface provides tools for semi-automatic person-annotation, and for semi-automatic batch person-annotation. The semi-automatic batch annotation tools are an extension to the basic MediAssist

system, and are the most important aspect of the semi-automatic person-annotation approach proposed in this chapter. We describe each aspect of the system in more detail in the sections below.

4.2.1 Annotation Database

The *annotation database* stores all person-annotations submitted by the users. The annotations can take two forms:

- An annotation can associate a person name with a face within a photo in the photo collection. If the face corresponds to an unknown person, the generic annotation 'unknown' can be applied. We will show in Section 4.3.2 how it is possible to submit such annotations using the MediAssist system interface. If a face is annotated with a specific name or with the generic label 'unknown' then we say the face is *positively annotated*, or simply that it is *annotated*.
- An annotation can state that a person name does *not* correspond with a specific face. We will show in Sections 4.3.1 and 4.3.2 how annotations of this type can be submitted using the proposed extension to the MediAssist system. If a face is labelled as *not* corresponding to one or more names, but is not annotated with a specific name or as 'unknown', then we say that the face is *negatively annotated*.

We use the term *unannotated* to distinguish between those faces that are *positively annotated* from those faces that are not *positively annotated*. Thus, if we say that a face is *unannotated* we mean that the face is not *positively annotated*. An *unannotated* face may contain *negative annotations*, however, because *negatively annotated* faces are not annotated with a person name. *Unannotated faces*, essentially, are faces that are candidates for annotation suggestions from the *person search engine*.

In the description above we make the assumption that all annotations are face region-level annotations. That is, the annotation describes a face region within a

photo Annotations could also be specified at the level of a photo rather than at the level of a face, where a name is associated with a photo rather than associated with a face region within a photo. We call this photo level-annotation. In our implementation we use the more specific face region-level annotation and, unless stated, all person-annotations in the proposed system in this and subsequent chapters refer to face region-level annotation.

4.2.2 Person Search Engine

In addition to all annotations being stored in an annotation database, the content and context of each annotated face will be analysed by a person search engine. The person search engine models the context-based and content-based characteristics of all annotated faces in the user's collection, and can suggest annotations for unannotated faces in two ways:

- *Person Classification.* Given a specific face, the person search engine can suggest a list, ranked in order of confidence, of suggested names for that face. We call this *person classification* because, given a face, the system attempts to *classify* this face as a specific person We prefer this term to 'face recognition' for two reasons. Firstly, face recognition refers to approaches based on analysis of the face area, while our system uses additional features. Secondly, we prefer the term classification to recognition because it distinguishes better between person classification and person retrieval.

- *Person Retrieval.* Given a specific name, the person search engine can suggest a list of faces corresponding to that name, ranked in order of confidence We call this *person retrieval* because, as in general information retrieval, the task is to *retrieve* documents relevant to a given information need. In the current scenario the information need is a specific person and the retrieved 'documents' are faces which may correspond to that person.

For both classification and retrieval, if we are interested in photo-level annotation rather than face region-level annotation, the person search engine could instead return results at the photo level, suggesting a list of names for a given photo for the classification task, and suggesting a list of photos that may contain the candidate person for the retrieval task.

The person search engine analyses all of the person annotations in the annotation database and, based on these known annotations, it is able to suggest annotations for unannotated faces The search engine differs from the annotation database in that, whereas the annotation database returns faces that have been annotated, the search engine will find names for faces, or faces for names, in cases where the faces have not been annotated by the user. The results from the search engine are uncertain and are based on person classification and retrieval techniques, whereas the results from the annotation database are known to be correct as they are based on annotations submitted by the user After each new annotation has been submitted to the *annotation database*, the person search engine will be updated accordingly, performing any analysis and updates necessary to keep it fully consistent with the classification and retrieval approaches described in Chapter 5.

Search Engine Filters

Before the search engine returns results to classification and retrieval queries, a number of *search engine filters* determine eligible names or faces for a given classification or retrieval query For person classification a list of negative annotations will be used to filter names which are known not to correspond to the candidate face. For person retrieval a list of unannotated faces determines which faces can be returned as annotation suggestions, and again a list of negative annotations is used to remove faces which are known not to be the query person.

The person search engine also uses a *photo presence* filter, which filters people who are known to be present in a specific photo This means that if a person occurs in a photo the system will not suggest that another face in the same photo will be

annotated with the same name.

4.2.3 Semi-Automatic Batch Person-Annotation Approach

If there is a least one annotation in the annotation database, the person search engine is able to suggest an identity for any unannotated face. The system visualisation interface will send person classification name suggestion requests to the person search engine every time an unannotated face is displayed in photo detail view as the user browses through their collection. This classification-based semi-automatic person-annotation can be found in the basic MediAssist system described in Chapter 3.

When the user annotates a face as corresponding to a certain person, we hypothesise that the user is interested in annotation at that moment. Also, when the user searches for a specific person, in addition to displaying all photos known to contain that person, the user may also be interested in seeing additional, unannotated faces, that may correspond to that person.

If this is the case, the system takes the approach of suggesting more faces that may correspond to a given person, where this suggestion is prompted by either a face being annotated with that person name, or by the user searching for that person. To do this, the system will send a person retrieval request to the person search engine, requesting a list of suggested faces for the person name. The system will then return a list of unannotated faces, ranked by the system's confidence that the face corresponds to the query person.

Since it is not known if these suggestions correspond to the candidate person, the system will require the user to either accept or reject these annotation suggestions. Specifically, the user can accept the suggested face, annotating it with the candidate person's name, they can reject it by *negatively annotating* it as not the candidate person, or they can reject the suggested face entirely by annotating it as 'unknown'. When a batch of annotations is submitted to the system, the person search engine can analyse the content and context of these person-annotations, using this information as additional training data to improve the accuracy of the results for future

person classification and person retrieval queries.

We believe that this framework, which suggests person-annotations to the user when the user has expressed an interest in a specific person, will lead to highly efficient annotation of people in personal photo collections In the next section we will introduce the implementation of this annotation framework in the MediAssist system user interface.

Other Approaches to Batch Person-Annotation

The approach presented here is quite similar to that proposed by Girgensohn et al. (2004a), who also rank unknown faces based on their similar to known faces. There are a few crucial differences, however, between their system and ours. While we attempt to integrate the annotation tools naturally into the user interface, making them available to the user as they search and browse their photo collection, their system forces users to first enter a 'face-view mode' that is used for managing and annotating faces. We believe that by naturally integrating our annotation tools into the browsing and search interface users are more likely to use them. Also, their system does not allow users to 'reject' faces for a particular name. This means that if a face is similar to a particular name, then the system will always suggest the face for that name until it has been annotated with the correct name, unlike with our system.

Other approaches to semi-automatic person-annotation create clusters of faces, which the user can then annotate in a batch manner (Cui et al., 2007; Suh and Bederson, 2007). These approaches force the user to firstly browse the face clusters before they can begin annotating faces, unlike our approach which naturally includes the annotation tools in the searching and browsing interface. Also, these approaches do not suggest names for unknown faces but rather they just group a set of faces to be annotated manually by the user. A possible combination of these approaches with our own approach could use our person classification techniques to suggest a name for a cluster of faces.

4.3 User Interface for Person-Annotation

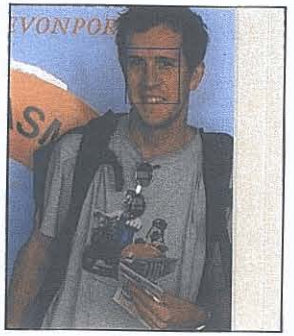
In the previous section we introduced a model for semi-automatic person-annotation in personal photo collections. In this section we describe the implementation of this approach in the MediAssist system. Firstly, we describe the person-annotation tools available in the standard MediAssist system. Then we introduce the enhanced version of the MediAssist system, which facilitates batch annotation of people in personal photo collections.

4.3.1 Semi-Automatic Person-Annotation Interface

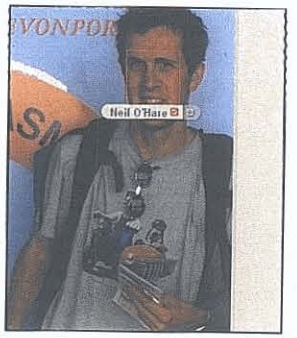
The MediAssist system provides semi-automatic tools for annotating the names of individuals present in photographs. In *Photo Detail* view in the MediAssist system, which is described in Chapter 3, the system can highlight all detected faces, a feature which is activated by clicking a 'highlight faces' button shown in Figure 4.2 (a). When the user clicks this button a rectangle is drawn around all faces in the image (Figure 4.2 (b)). When the user moves the mouse over a face in *Photo Detail* view, the system will display the top-ranked name suggestion for that face, as shown in Figure 4.2 (c). If the top-ranked name suggestion is correct, then the user can confirm the suggestion, and annotate the face with the suggested name, by clicking on the name or by clicking on the red 'confirm' icon to the right of the name.

If the top-ranked name suggestion for a face is not correct then the user can click on the 'show alternative names' button, to the right of the name and the 'confirm' icon. Clicking on this button prompts the system to display a shortlist of 10 name suggestions for this face, ranked in order of confidence, shown in Figure 4.2 (d). If the correct name is displayed in this list then the user can click on the name to annotate the face with that name. Otherwise clicking the 'more' button will display the names of all known people in the user's photo collection (Figure 4.2 (e)).

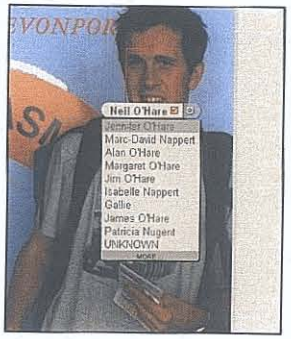
(a) Photo Detail view. User clicks 'highlight faces' button...



(b) Face now highlighted in interface. User moves mouse over face...

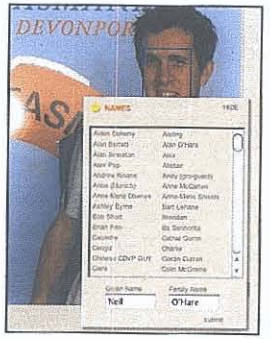


(c) Top name suggestion for face displayed. User can click to confirm, or click 'show alternative names' button to the right of the name...



(d) Top 10 list of name suggestions displayed. User clicks 'More' button...

(e) All known names in users collection are displayed. If this is a new person, user clicks 'Label with new name' button...



(f) 'New Name' text box appears. User types name and clicks submit

(g) Name confirmed. 'Tick' icon beside name now grey, instead of red as in (c)

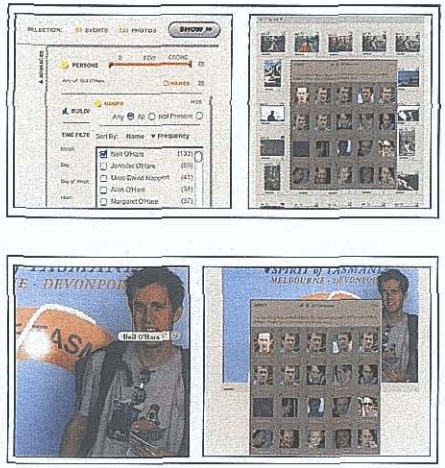
Figure 4.2: Semi-Automatic person-annotation in the basic MediAssist system interface.

If the name of the face to be annotated is not known to the system, then the user can annotate the face with a new name by clicking on the 'Label with New Name' button (Figure 4.2 (e)). The user will then be able to type a new name for this person and annotate the face with this new name. Once a face has been annotated with an identity, the 'confirm' icon to the right of the top name under the face changes colour from red to grey, indicating that it is no longer a name suggestion but a confirmed annotation, as can be seen by comparing Figure 4.2 (g) and Figure 4.2 (c).

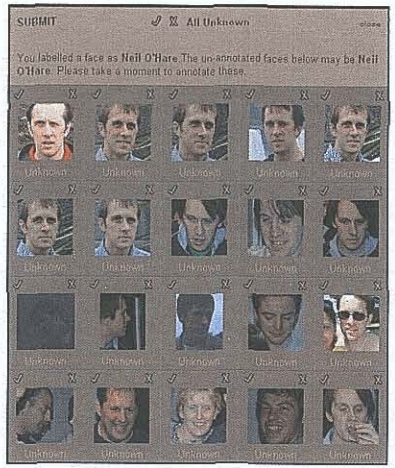
When face is annotated and the annotation is stored in the annotation database the person search engine will subsequently analyse the image content and the context of the newly annotated face, and update the person search engine with this new information.

4.3.2 Semi-Automatic Batch Person-Annotation Interface

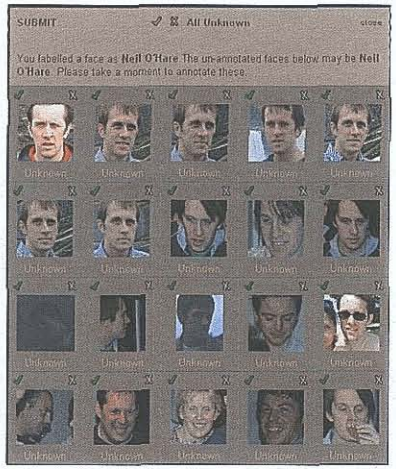
The previous section described the semi-automatic person-annotation tools found in the basic MediAssist system. In this section we describe an extension to the basic MediAssist system that facilitates semi-automatic batch annotation of people in personal photo collections. The proposed enhanced system suggests faces for a given person name at appropriate moments as users browse their photo collections. The display of the batch annotation suggestions in the interface is prompted by two different events. If the user searches for a specific person using the person search tools described in Chapter 3, the system displays photos or events containing this person. In addition, the system will also suggest additional faces that have not been annotated but are likely to correspond to the name being searched for, and will encourage the user to take a few moments to confirm or reject these suggestions The interface is shown in the top panel of Figure 4.3 (a), where it can be seen that the suggested annotations are clearly distinguishable from known annotations.



(a) Top: User searches for 'Neil O'Hare'. Bottom: User annotates an occurrence of 'Neil O'Hare'. In both cases the system suggests potential occurrences of 'Neil O'Hare' for annotation...



(b) Suggested occurrences.The user clicks on the 'confirm all' tick icon, at the top of the picture.



(c) All faces labelled as 'Neil O'Hare'. There are 5 incorrect suggestions here, so the user can click the 'x' icon besdie these faces to reject them...

(d) 4 suggestions have been rejected. An additional indorrect suggestion is an uknown person, so the user clicks the 'unknown' button under this face..

(e) All annotation suggestions are now correctly labelled. The user can click the 'Submit' button...

(f) User clicks 'Submit' button to submit annotations to database. The system will suggest more faces,to begin another iteration of batch annotation.

Figure 4.3: Semi-Automatic batch person-annotation in the enhanced MediAssist system interface.

The second event that initiates batch annotation is when the user annotates a face in *Photo Detail* view using the tools described in the previous section and shown in Figure 4.2. Given that the user has just annotated a face with a person name, we make the assumption that the user would be willing to invest a little more time in confirming some suggested additional annotations of the same person. Accordingly, the system will suggest additional faces that are likely to be the same person, as shown in the bottom panel of Figure 4.3 (a).

Once batch person-annotation has been initiated, whether by a person search or by a person-annotation, the same pop-up window is displayed to the user, and the user interaction within this pop-up window is the same for each case. The batch person-annotation pop-up window displays 20 suggested faces for the candidate name. The system provides 3 types of icons which allow for confirmation of annotation suggestions:

- A 'tick', or *confirm* icon. This icon confirms that a face suggestion is correct and corresponds to the candidate name, and clicking it causes the face to be annotated with the name.

- An 'x', or *reject* icon. This rejects the face as corresponding to the candidate person, and clicking it causes a negative annotation to be applied to this face. The system still does not know the identity of this face, but the face will never be suggested for this name again, and the name will never be suggested for this face. The face may, however, be suggested during future system interaction, for different candidate person names.

- An 'Unknown' icon annotates the face as being an 'unknown' person, and the face will never be suggested for any other candidate person names.

Each of these icons is to be found beside each face, and the colour of the icon indicates if it is active or not. If the icon is monochrome then it is not active and it does not apply to the photo. If it is green or red it is active and applies to the photo. Only one icon at a time can be active with respect to a single photo.

To allow for efficient batch annotation confirmation, the interface also provides batch versions of each of these icons. These batch confirmation icons can be found at the top-centre of the batch annotation pop-up window, and clicking on one of them applies the appropriate label to all faces displayed. They can be seen in Figure 4.3, (b) and (c). So, for example, if all annotation suggestions are correct, the user can click on the confirm all 'tick' icon and all faces will be labelled as correct, indicated by the colour of the 'tick' icon above each face changing its colour to green. This can be seen in Figure 4.3 (c).

Quite often, however, not all of the name suggestions are correct. In the example in Figure 4.3, 15 of the 20 suggestions are correct. So, after a quick scan through the suggested faces, the user could decide to firstly label all faces as correct, and then adjust the 5 incorrect suggestions by using the 'x' and 'Unknown' icons. In Figure 4.3 (d), 4 of the incorrect suggestions have been rejected using the 'x' icon. The other incorrect suggestion is an unknown person, so the user clicks the 'Unknown' icon under this face to label the face as unknown, as shown in Figure 4.3 (e).

Once the user is happy that all of the suggested faces have been labelled correctly, then they can click on the 'Submit' button at the top-left of the batch annotation pop-up window, shown in Figure 4.3 (f). Clicking on this button submits all the annotations to the annotation database. The new annotations will be subsequently analysed by the person search engine. After this process, the system then presents another 20 face suggestions for the candidate name, starting another batch annotation iteration. The user can continue this process for as long as they wish, and when they would like to stop annotating they can click on the 'close' icon at the top-right of the batch person-annotation window (Figure 4.3 (b)).

4.4 Summary

We have proposed an approach to semi-automatic person-annotation in personal photo collections. The proposed approach facilitates batch person-annotation by

using retrieval techniques to suggest faces for a given person name, and presents these suggestions to users at appropriate moments as they interact with their photo collections.

We have also described the implementation of this approach in an enhanced version of the user interface of the MediAssist personal photo management system. In Chapter 7 we will evaluate the proposed approach by conducting experiments with users interacting with this enhanced user interface, annotating faces in their own personal photo collections.

In the next chapter we will propose content-based and context-based approaches to person classification and person retrieval in personal photo collections, as it is these approaches which will enable the semi-automatic person-annotation approach proposed in this chapter.

Chapter 5

Proposed Approaches to Person Classification and Person Retrieval

In the previous chapter we proposed an approach to semi-automatic person annotation which relies on a person search engine to perform automatic person classification and retrieval in order to suggest annotations to the user. In this chapter we will propose approaches to person classification and retrieval that will enable such semi-automatic annotation.

For context-based person classification and retrieval, we propose a language model approach, which extends a previously proposed approach to person classification. The extended approach allows person retrieval in addition to classification, and uses intricate smoothing techniques to improve the estimation of the probabilities for the language model. We also propose a nearest neighbour context-based approach, which classifies and retrieves people based on spatial and temporal distances from annotated faces. We then propose a combined context-based approach which combines the results from individual context-based approaches

For content-based person classification and retrieval, we propose novel *face colour* and *image colour* features, in addition to using *face recognition* and *body patch* features. We also propose a combined content-based approach which combines the results from all four of these features.

We propose combining the results of these context-based and content-based person classification and retrieval techniques to give a combined approach to person classification and retrieval that makes direct use of both context and content

All of the proposed approaches make the assumption that we have a personal photo collection which has been partially annotated in terms of the people present in the photos, as described in Chapter 4

5.1 Context-based Language Modelling Approaches to Person Classification and Retrieval

In this section we will outline a language modelling approach to information retrieval We will then show how this approach can be applied to context-based person classification and retrieval in personal photo collections. After that we describe techniques used to estimate the probabilities for a context-based language model for person classification and retrieval.

5.1.1 Language Modelling in Information Retrieval

A language model is a probability distribution that models the stochastic process behind the generation of a series of tokens in a language, such as words in text Language models have been successfully used in speech recognition, optical character recognition and machine translation (Manning and Schütze, 1999). The probability of a sequence of terms can be expressed as:

$$P(w_1 \ldots w_n) = \prod_{i}^{n} P(w_i | w_1 \ldots w_{i-1}), \tag{5 1}$$

where w_i is the i^{th} term in the text sequence. In practice, different language models make different assumptions about independence between terms and use different estimation strategies to estimate the probabilities of individual terms. The simplest language model, the unigram model, assumes complete independence between terms

and takes the form:

$$P(w_1 \ldots w_n) = \prod_i^n P(w_i) = P(w_1)P(w_2)P(w_3)P(w_4) \ \ldots P(w_n). \tag{5.2}$$

This assumption of independence for the unigram model ignores the relationship between terms but, in spite of this limitation, it has proven to be an effective model in practice. The model has the advantage that is simple and easily understood, and it is straightforward to estimate the parameters of a unigram model from sample data by counting the relative frequency of terms

The use of language models in information retrieval was first proposed by Ponte and Croft (1998). In language modelling approaches to information retrieval, documents are typically modelled using a multinomial unigram model, where the probability distribution of each underlying document is assumed to be a multinomial and the probability of a term is independent from other terms. A separate language model is created for each document in the collection and, given an information need in the form of a query, the *query likelihood* for each document is calculated, which is the probability of the language model for that document creating the query:

$$P(\mathbf{q}|M_d) = \prod_{i=0}^{|q|} P(q_i|M_d), \tag{5.3}$$

where q is the sequence of query terms and M_d is the language model of the document. This *query likelihood* can then be used to rank retrieved documents. This simple model, then, reduces the information retrieval task to the task of estimating the probabilities of the individual terms for each document. We will discuss statistical estimation strategies in Section 5.1 3 below.

A number of alternative approaches to ranking documents in language modelling information retrieval have been proposed. The *document likelihood* approach creates a language model for the query and ranks documents based on the probability that the query created the document. The *relative entropy* approach represents both

queries and documents as language models, and measures the amount of information needed to encode information from one probability distribution into the other probability distribution (Lafferty and Zhai, 2001) In addition to being used in text information retrieval, language models have also been used for multimedia information retrieval (Westerveld et al., 2003a; Mc Donald, 2005; Mc Donald and Smeaton, 2005).

5.1.2 Context-based Language Modelling for Person Classification and Retrieval

It is possible to use a simple and powerful language model approach to information retrieval for both person classification and person retrieval in context-aware personal photo collections Just as we can view text documents and multimedia documents as being created by a stochastic process, it is also possible to view the creation of personal photos and the appearance of people in them in the same way. For the purposes of person classification and retrieval, we can view the vocabulary of the language model as being composed of all the people who can possibly appear in the user's photo collection. As each photo is captured by a user, a stochastic process determines which people will appear in the photo. Intuitively, we can understand this process as being determined by factors such as the people present at the time of photo capture, the location etc.

As with text information retrieval, we create a language model for every photo, representing the probability of occurrence of each person who can appear in the user's collection. In practice, this vocabulary of people is restricted to all the people currently known to occur in the user's collection, based on the person annotations stored in the *annotation database* (see Chapter 4) For person classification and retrieval, we are only interested in one person at a time In other words, a retrieval query will only have one term, namely the person we are searching for, instead of a number of terms, as in the general information retrieval scenario. For this reason

we do not use to *query likelihood* with our approach, but rather we can use *person likelihood*, instead, the probability of a specific person, given the language model

$$P(p|M_d), \tag{5.4}$$

where p is a specific person In this model we treat the photo, rather than a face within a photo, as a document. Since the contextual features of each face within a photo are the same, we use the language model of the photo to represent the person probabilities for each face in the photo.

We use the standard multinomial unigram language model, while noting that this does not model the fact that the same person cannot appear more than once in the same photograph. In practice, we circumvent this problem in a post-processing step by filtering the classification results to remove the names of people already known to be in the photo, and filtering retrieval results to remove faces from the returned list if they occur in a photo known to contain the query person. The nearest neighbour approaches described in Section 5 2 also suffer from this limitation, and we also filter the results from nearest neighbour approaches in the same way The use of this and other filters is discussed in Section 4 2 2 in Chapter 4

Person Classification with Language Models

The language model approach calculates the probability of a specific person occurring for each photo. For person classification, the task is· given a specific face, return the name of the person corresponding to that face. Accordingly, in order to classify a face in a photo, the score for a specific person is the person's probability, given the language model for that photo, $P(p|M_d)$. For classification the document d is fixed, and we rank all possible person names in order of decreasing *person likelihood*, giving a ranked list of suggested names for the query face.

Person Retrieval with Language Models

Person retrieval with this model is very similar to person classification, and we again use *person likelihood* to rank results. For person retrieval, in the scenario where the query is for one specific person, the person is fixed, and we rank all documents according to the probability, $P(p|M_d)$, of creating that person given the language model for the photo.

The difference between *person retrieval* and *person classification* is that for retrieval we use *person likelihood* to rank candidate faces, given a specific person, whereas for the classification task we use the same score to rank candidate names, given a specific face. We also note that, although we propose a scenario where the system would only ever need to process single person queries, this retrieval model is essentially the same as ranking by *query likelihood*, and the model could be used to support queries for photos containing multiple people using Equation 5.3.

5.1.3 Statistical Estimation

The parameters of a language model for text retrieval can be estimated using the relative frequency of terms within each document. For example, if a document contains 1000 terms and the word *photo* occurs 10 times, the probability for that term in the document's language model would be estimated as $10/1000 = 0.01$. In practice, sophisticated smoothing techniques are used to improve the accuracy of these probability estimates, as discussed below.

For standard information retrieval we have an obvious way to estimate the parameters of our language model using the relative frequency of terms within a document, as described above. How do we estimate the probabilities of specific people occurring in personal photo collections? What we do is use the existing annotations in a user's collection to estimate the probabilities of occurrence of each person within the context of a given photo. This approach has previously been proposed by Naaman et al (2005), although they did not propose the approach in terms of a lan-

guage model. They applied their approach to person classification only and not to person retrieval. We extend their approach by associating it with a language model, which gives us a well-understood model we can use for *person retrieval* in addition to *person classification*. In addition, by using this model we can make use of the intricate smoothing techniques developed by the language modelling community to improve the estimates of the individual person probabilities.

We will first outline how to estimate a user collection model for person occurrence, and in the following subsections we will show how we can use the contextual metadata associated with personal photos to create language models for each photo. To estimate the probabilities for the language model for each photo, we make use of the context surrounding the photo. There are a number of context-based features that we can use for this purpose, and we outline each of these in turn below. The *user collection*, *temporal proximity*, *spatial proximity* and *co-occurrence* maximum likelihood language models described below essentially correspond to the estimators proposed by Naaman et al. (2005). After describing these, and our own *Cyclic Temporal Context* language model, we describe proposed smoothing techniques that extend this approach by improving the probability estimates.

Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a simple statistical estimation technique that assigns probabilites such that the maximum possible probability is assigned to the observed data (Fisher, 1922). Below we will outline a number of maximum likelihood estimator language models based on the context of photo capture.

User Collection Language Model. A maximum likelihood estimate of the probability of a person, p, given a user's collection, can be calculated as the relative frequency of annotations of that person in the user's collection:

$$P_{MLE}(p) = \frac{C(p)}{N}, \tag{5.5}$$

where $C(p)$ is the number of annotated occurrences of person p and N is the number of annotations in the user's collection In the personal photo management scenario, all people known to occur in the user's photo collection will have a non-zero probability. This language model is a background model of the user's collection, as opposed to the other models below, which are specific to an individual photo and attempt to estimate probabilities for specific photos So for example, if we have 41 annotations in the users collection, 7 of which correspond to 'Neil', then the estimate for 'Neil' for the user collection language model is $7/41 = 0.171$. This example is illustrated in Figure 5.1 on page 89.

Temporal Proximity Photo Language Model. In order to estimate the probabilities of the language model for a particular photo we can use the relative frequency of that person within a certain *context* of the photo. For the *temporal proximity photo language model* we estimate these by considering the relative frequency of annotations within a certain time window surrounding the time of photo capture, for example within 5 minutes of the time of photo capture, or within the same event. The maximum likelihood estimate of the probability of person p given the temporal proximity of the photo is:

$$P_{MLE}(p|d_{tp}(l)) = \frac{TP(d,p,l)}{TP(d,l)}, \tag{5.6}$$

where $d_{tp}(l)$ specifies a temporal proximity of length l around the photo d, $TP(d,p,l)$ is the number of annotations of person p in this temporal proximity, and $TP(d,l)$ is the total number of annotations within this temporal proximity of d. The temporal proximity used is specified by the parameter l and its duration can vary. This can be a fixed size time window measured in seconds, which corresponds to a time window containing all annotations taken l seconds before and after photo capture. In the example in Figure 5.1 on page 89, if we specify a time window of 5 minutes, containing 4 annoations, none of which correspond to 'Neil', then the estimate for Neil within this time window is $0/4 = 0$.

Alternatively, we can use the automatically detected event hierarchy to specify a temporal proximity window, using either *events* or a *sub-events* as detected using the approach described in Chapter 3 We will evaluate a number of different temporal proximity time windows, including events and sub-events, in Sections 6.4.1 and 6.4.2 in Chapter 6.

Spatial Proximity Photo Photo Model. We can also use the latitude and longitude co-ordinates of the location of photo capture to calculate the geographic distance between photos. The use of this feature is based on the assumption that certain people in personal photo collections tend to occur in certain places, for example in the town or city where they live. Based on this, we can also construct a *spatial proximity photo language model*, which estimates the parameters of a photo's language model based on the relative frequency of annotations within a certain *spatial proximity* of the location of photo capture:

$$P_{MLE}(p|d_{sp}(l)) = \frac{SP(d,p,l)}{SP(d,l)}, \tag{5.7}$$

where $d_{sp}(l)$ specifies a spatial proximity of size l around the photo d, $SP(d,p,l)$ is the number of annotations of person p in this spatial proximity, and $SP(d,l)$ is the total number of annotations within the temporal proximity of d. As with temporal proximity, the parameter l and its size can refer to a fixed size radius around the point of photo capture or, alternatively, it can refer to all photos taken in a certain town, county or country, as determined by the gazetter in the MediAssist system (see Chapter 3). We will evaluate a number of different spatial proximity windows in Sections 6.4.1 and 6.4.2 in Chapter 6. .

Cyclic Temporal Context Photo Language Model An alternative time-based approach is to use the idea of temporal schemata, as described in Section 3.3.1 in Chapter 3. We refer to this approach as *cyclic temporal context* and it is based on recurring temporal patterns such as the time of day, day of week and season of the

year The justification for using these to estimate the parameters of a photo language model is that certain people may have a tendency to appear in photos taken at the weekend, for example, or in the evening. Again, we calculate a maximum likelihood estimate of the probabilities for this model in the same manner as before:

$$P_{MLE}(p|d_{ctc}(c)) = \frac{CTC(d,p,c)}{CTC(d,c)}, \tag{5.8}$$

where $d_{tp}(c)$ specifies a temporal context of type c around the photo d, $CTC(d,p,c)$ is the number of annotations of person p in this temporal context, and $CTC(d,c)$ is the total number of annotations within the temporal context. The cyclic temporal context is specified by the parameter c. We propose using the following cyclic temporal context features: *hour of the day, day of the week, month, year, light status* (day/night/dusk/dawn) We also use two time of the day features called *2-hour slot* and *4-hour slot*. These divide the day into 2-hour and 4-hour time slots, beginning at 12:00 am, and consider each time window as a separate temporal context. As an example, using *4-hour slot*, the 4-hour time window between 12:00 am and 04:00 am would be considered a separate temporal context. Although these 2-hour and 4-hour time slots are somewhat arbitrary, we evaluate them in order to determine how their performance compares to hour of the day, in order to see if such arbitrary time division can be useful.

Co-occurrence Photo Language Model. It is also possible to model the social co-occurrence patterns of people in photo collections. That is, given that *Neil* occurs in a particular photo or event, then we say that say a certain other person, *Isabelle* for example, is also likely to occur in the same photo or event because they tend to occur together. Naaman et al. (2005) propose a measure they call *PeopleRank* to model this. Since this measure creates a probability distribution across all people in the user's collection we can also create an estimate of the parameters of a language model using this feature. The maximum likelihood estimate of the probability of

person p for the co-occurrence photo language model is:

$$P_{MLE}(p|d_{cooc}(t)) = \frac{CL(d,p,t)}{CL(d,t)}, \tag{5.9}$$

where $d_{cooc}(t)$ specifies co-occurrence type t for photo d. The type of co-occurrence can either be *event co-occurrence* or *photo co-occurrence*. For photo co-occurrence, two people are considered to have a *co-occurrence link* if they are known to occur in the same photo. For event co-occurrence, two people are considered to have a co-occurrence link if they are known to occur in the same event. One could also imagine co-occurrence based on spatial or temporal proximity, for example, *week co-occurrence* if two people occur within a week of each other, or *city co-occurrence* if they occur in the same city. In this work we only explore photo co-occurrence and event co-occurrence. $CL(d,t)$ is the total number of co-occurrence links, across the user's entire collection, to any of the the people present in the current context (photo or event). $CL(d,p,t)$ is the number of co-occurrence links from person p to any of the the people present in the current context.

All photos in the same event will have the same *event co-occurrence language model*. All photos known to contain the same set of people will have the same *photo co-occurrence language model*

Smoothing Techniques for Probability Estimation

Maximum likelihood estimates are problematic for estimating probabilities because they do not deal well with sparse or missing data. In text-based information retrieval the problem manifests itself when a term is missing from a document, and is given a probability of zero. This means that the document's language model could never create that term, so the document would have a *query likelihood* of 0 for any query containing that term. In language modelling for information retrieval, where a language model is created for each document, this problem is exacerbated because documents can be quite small and so are likely to be missing terms that are pertinent

to the topic of the document.

In the person classification and retrieval scenario, the same problem can be found if a person does not occur in the context of a given photo This will particularly affect proximity-based estimators which use narrow windows For example, using a *5-minute temporal proximity language model* will lead to a probability of zero for many people for many photos, because only a small number of people will be present during such a short time span. A solution could be to use a coarser estimator, for example *1 week*, but then the information regarding the immediate photo capture context is lost.

The language modelling community has addressed this problem of sparse and missing data by developing techniques which adjust a probability distribution by assigning a portion of the distribution to terms with zero frequency, a process known as *smoothing*. A wide range of smoothing techniques have been proposed, and they generally fall into two main categories: discounting-based methods and combination-based methods (Manning and Schütze, 1999). Discounting methods work by taking a small portion of probability away from observed events so that this probability can then be assigned to unseen events. Combination-based methods involve combining the document language model with a background model that is less prone to sparse and missing data. In information retrieval, this background model is usually a language model for the entire collection

For our work in context-based person classification and retrieval, the *user collection language model* can be used as the background model. We propose smoothing using linear interpolation, also known as *Jelinek-Mercer Smoothing* (Lafferty and Zhai, 2001), to improve the probability estimates for context-based photo language models. *Jelinek-Mercer Smoothing* compared well with other smoothing methods in a large-scale evaluation in the field of video retrieval (Mc Donald, 2005; Mc Donald and Smeaton, 2005). This smoothing approach calculates the smoothed estimate of the probability of a term, w, as a linear interpolation of the document model, D,

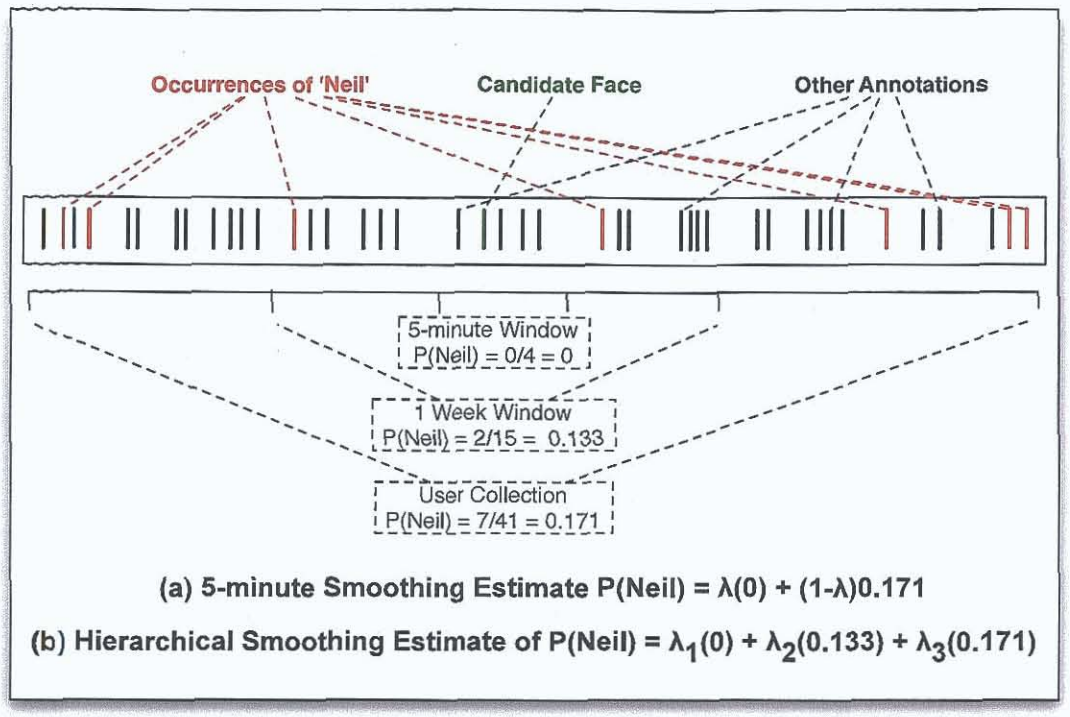


Figure 5.1: Smoothing and Hierarchical Smoothing for temporal proximity language models. (a) 5-minute MLE estimate smoothed by the user collection estimate. (b) 5-minute estimate hierarchically smoothed by the 1 week estimate and the user collection estimate.

with a background model, represented by the collection model, C:

$$P_{SM}(p) = (\lambda)P_D(p) + (1 - \lambda)P_C(p). \tag{5.10}$$

It is necessary to choose a suitable value for the parameter λ, and it is customary to learn this value empirically using a test collection.

Hierarchical Smoothing Techniques for Probability Estimation

Westerveld et al. (2003b), working in the field of video retrieval, proposed a hierarchical smoothing technique for estimating term probabilities. A video can be understood as having a hierarchical structure, as it is composed of scenes, which are in turn composed of shots. With this in mind, they propose a hierarchical

Jelinek-Mercer language model which takes advantage of this structure:

$$P(w|SHOT, \lambda_{shot}, \lambda_{scene}, \lambda_{vid}, \lambda_{col}) = \lambda_{shot} \times P(w|shot) + \lambda_{scene} \times P(w|scene) + \lambda_{vid} \times P(w|vid) + \lambda_{col} \times P(w|col). \quad (5.11)$$

The parameters are subject to the constraint that $\lambda_{shot} + \lambda_{scene} + \lambda_{vid} + \lambda_{col} = 1$, and are tuned on an appropriate collection. This model can be understood as being hierarchical because the maximum likelihood estimate for a *shot* is first smoothed by the estimate for the *scene* that contains it, then by the *video* which contains the scene, and finally by the *collection*. Hierarchical language models have also been proposed for other smoothing techniques (Mc Donald, 2005; Mc Donald and Smeaton, 2005).

We believe a similar hierarchical model can be applied to context-based photo language models. Temporal proximity, for example, can easily be structured hierarchically, with years composed of months, which are composed of days, which are composed of hours. We can also easily imagine a spatial proximity hierarchy constructed from cities, counties and countries. We propose *3-layer hierarchical smoothing* and *4-layer hierarchical smoothing* language models to exploit this hierarchy and improve the probability estimates of language models. We can define a 3-layer hierarchical smoothing scheme for temporal proximity, spatial proximity and co-occurrence language models as follows.

$$P(p|photo, \lambda_1, \lambda_2, \lambda_{col}) = \lambda_1 \times P(p|context_1) + \lambda_2 \times P(p|context_2) + \lambda_1 \times P(p|col). \quad (5.12)$$

where $context_1$ is the most specific context in the hierarchy and $context_2$ is a wider context. This model assumes that the language models being combined are indepen-

dent of each other but, although this is clearly not the case, the model still performs well in practice.

We see an example of hierarchical smoothing in Figure 5.1, with a *5 minute* $\rightarrow$ *1 week* hierarchy, with *5 minutes temporal proximity* corresponding to $context_1$ and *1 week temporal proximity* corresponding to $context_2$. In this example the MLE estimate for *5 minute* for is 0 because there are no occurrences of 'Neil' within 5 minutes of the candidate face. The MLE estimate for *1 week* for 'Neil' is $2/15 = 0.133$ because there are 15 annotations within 1 week of the candidate face, 2 of which are 'Neil'. The MLE collection for *user collection* is $7/41 = 0.171$ because 7 out of 41 annotations in the users collection are 'Neil'. Hierarchical smoothing techniques work by combining all of these estimates.

It is also possible to construct hierarchies for spatial proximities and co-occurrence. For co-occurrence, the hierarchy can be constructed using co-occurrence photo as $context_1$ and co-occurrence event as $context_2$.

For temporal proximity and spatial proximity we can also construct a 4-layer hierarchy:

$$\begin{aligned} P(p|photo, \lambda_1, \lambda_2, \lambda_3, \lambda_{col}) = \; & \lambda_1 \times P(p|context_1) \\ & + \lambda_2 \times P(p|context_2) \\ & + \lambda_3 \times P(p|context_3) + \lambda_{col} \times P(p|col). \end{aligned} \tag{5.13}$$

For both 3-layer and 4-layer hierarchical language models, the values for λ must add to 1, and we will tune these parameters on a test collection. In Sections 6.4.1 and 6.4.2 in Chapter 6 we will evaluate a number of combinations for hierarchical smoothing techniques to enable us to discover which values of $context_1$, $context_2$ and $context_3$ give the best probability estimates, for both temporal proximity and spatial proximity approaches. We will also compare 3-layer and 4-layer hierarcharies to evaluate if the added complexity of 4-layer hierarchies also leads to improved performance.

Work in the automatic organisation of personal photos, for example that of Pigeau and Gelgon (2005), automatically create hierarchies for the purpose for navigating and summarising photo collections This is somewhat different from the hierarchies presented here, which are created to improve the probabilities of language models and are not intended to be used directly in the organisation of personal photo collections. Automatically created hierarchies such as those proposed by Pigeau and Gelgon (2005), however, could also be used in hierarchical language models where they could possibly even further improve the probability estimates.

Combined Context Photo Language Model

It is possible to use linear interpolation to combine temporal, spatial and co-occurence hierarchcial smoothing approaches to create a combined language model. We do not repeat the equation here because it is essentially the same as the equations for linear interpolation smoothing given above (equations 5.10, 5.12 and 5.13). This will give us probability estimates based on all of the context-based features. We will evaluate a combined context photo language model based on temporal proximity, spatial proximity and co-occurrence hierarchical smoothing language models. Because location information is not currently available from all photo capture devices, we also combine temporal proximity and co-ccurrence hierarchical smoothing language models, to evaluate the performance of context features without using location information.

5.2 Context-based and Content-based Nearest Neighbour Approaches to Person Classification and Retrieval

An alternative approach to context-based person classification and retrieval is to use the "distance" between annotated faces and unannotated candidate faces. We

propose to apply standard *nearest neighbour* classification techniques (Duda et al., 2000) to context-based person classification and retrieval. We outline below how this approach can be applied to both person classification and person retrieval. Then we describe the context-based features and distance measures used for a context-based nearest neighbour approach, before describing the content-based features and distance measures used for a content-based nearest neighbour approach.

5.2.1 Nearest Neighbour Classification

The nearest neighbour approach to classification assigns a test point to the class of the closest labelled point to the test point (Duda et al., 2000). In our classification task we are interested in ranking suggested person names, given a specific face, instead of simply assigning it to a single class. We can do this by ranking person names based on their distance from the query face. In practice, there are numerous annotated occurrences of each candidate name, so the score assigned to a name, given a face, is the minimum of these. Let $\mathbf{b} = \{b_1 \ldots b_i\}$ be the set of all faces known to contain person p. The nearest neighbour score, $NN_{score}(a, p)$, for person p and face a, is:

$$NN_{score}(a, p) = \min_i [D(a, b_i)]. \tag{5.14}$$

$D(a, b_i)$ is a function which returns the distance between a and b_i. We will specify different nearest neighbour approaches by defining different distance measures between two faces. For nearest neighbour person classification, we will rank candidate names in increasing order of this nearest neighbour score. We will also use this score when we combine context-based nearest neighbour classification with content-based nearest neighbour classification, as described in Section 5.3.3.

5.2.2 Nearest Neighbour Retrieval

In order to apply the nearest neighbour approach to retrieval, we give each unannotated face a score, given a query for person p, based on the distance from

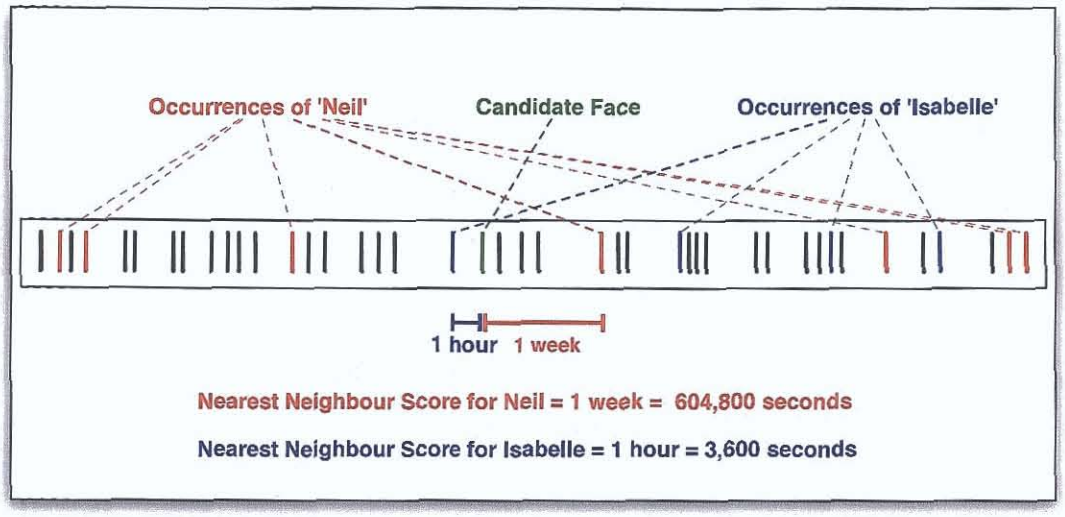


Figure 5.2: Nearest Neighbour classification using temporal proximity. The nearest neighbour score for a person is the minimum distance from the candidate to that person.

the candidate face and the nearest known occurrence of p. In the retrieval context however, we rank suggested faces based on their minimum distance from a given person name, calculated using Equation 5.14, rather than ranking names given a face as in the classification scenario. So, for example, nearest neighbour temporal proximity retrieval will rank all unannotated faces based on the temporal distance, in seconds, between each unannotated face and a known occurrence of the query person.

5.2.3 Context-based Approaches

Nearest neighbour approaches rely on a measure of distance, which restricts its use, in terms of context-based features, to temporal proximity and spatial proximity context-based features. Below, we define distance measures for temporal and spatial proximity. Using these temporal and geographic distance measures shown below allows us to use two separate context-based nearest neighbour approaches to person classification and retrieval.

Temporal Proximity Distance

For temporal distance between two faces, we simply measure the absolute difference in seconds between the capture times of the photos containing the faces. The temporal distance in seconds, D_T between photo A and photo B is:

$$D_T(A, B) = |Time_A - Time_B|. \tag{5.15}$$

For example, if the temporally closest occurrence of 'Neil' to a candidate face is 1 week away then the score for 'Neil' for this face is 1 week or 604,800 seconds. This example is illustrated in Figure 5.2.

Spatial Proximity Distance

We can measure the geographic distance, D_G, between two photos, A and B, using the following formula (Longley et al., 2001):

$$D_G(A, B) = R\left(\arccos\left[\sin(lat_A)\sin(lat_B) + \cos(lat_A)\cos(lat_B)\cos(lon_B - lon_A)\right]\right) \tag{5.16}$$

This equation assumes that the earth is a perfect sphere of radius R, in the unit of measurement we wish to use. In our implementation, we use kilometers, so $R = 6378.7$. Lat_A and lon_A are the latitude and longitude, respectively, of photo A, while lat_B and lon_B are the latitude and longitude of photo B.

5.2.4 Content-based Approaches

It is also possible to use nearest neighbour classification and retrieval with content-based features, extracted based on analysis of the visual image content of the photo. In the sections below we outline the features that we propose to use with this approach, followed by the distance measure that we will use.

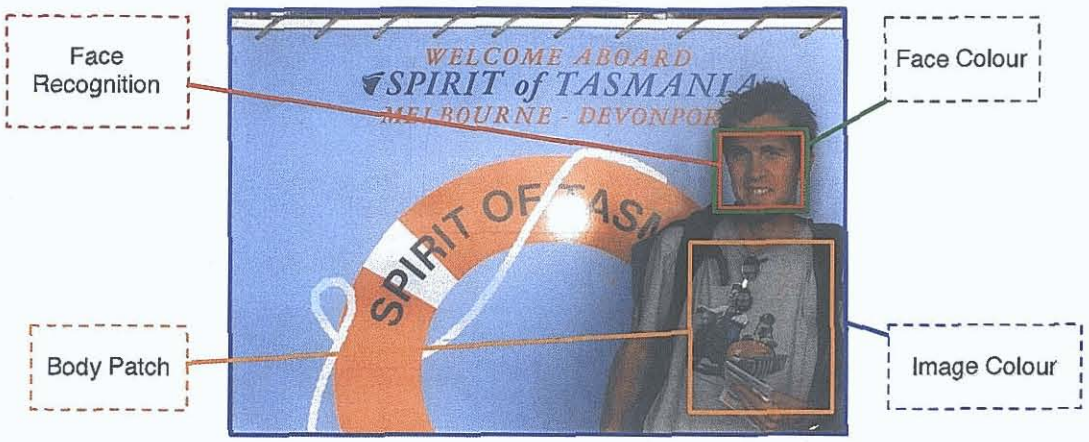


Figure 5.3: Content-based features for person classification and recognition.

Content-based Features

In this section we describe the content-based features that we use for person classification and retrieval, shown in Figure 5.3. In Section 3.4.1 we described the face detection approach used in the MediAssist system. Once faces have been detected in photos we can then analyse the visual image content of the face region and other regions in the photo-image to facilitate content-based person classification and retrieval. We outline below the 4 content-based features that we use for content-based person classification and retrieval. Work on the extraction of *face recognition* and *body patch* features was carried out by colleagues working on the MediAssist project. The *face colour* and *image colour* features are proposed as novel features for the task of person classification and retrieval in personal photo collections.

Face Recognition Feature. Face recognition is the traditional approach to identifying people in images and video, and works by creating a model for the face of each person in the database. The face recognition approach used in this work first analyses each face to determine the location of the eyes. This allows us to position, scale and rotate the face to create a normalised face image, which is then used for extracting facial features for recognition. The eyes are located using principal com-

ponent analysis (PCA) projections of candidate eye regions, similarly to Talmi and Liu (1999). A number of candidate eye regions are examined, and the region which is closest to the centre of the PCA eigeneye subspace is regarded as containing true eye region. The eye centres are located in a fixed position within the chosen eye region The face recognition technique used is the ICA (independent component analysis) subspace method based on the *Architecture II* method proposed by Bartlett et al. (2002). Each face is represented as a 48-feature vector.

Body Patch Feature. It is reasonable to assume that, within a given event, a person will not change the clothes they are wearing. The implication of this, in a context-aware system like MediAssist which detects events automatically, is that if we can detect the torso or 'body patch' area underneath a face then it should be a useful feature for person classification and retrieval.

We extract body patches corresponding to detected faces, with a body patch considered to be double the width and double the height of its corresponding face (Cooray et al., 2006). The body patch is located below the lower level of the face, and so the extracted area is an area below the face. Not all faces will have a corresponding body patch region, however: very large faces and faces towards the bottom of the photo-image will not have a corresponding body patch and it will not be possible to extract a body patch region for these faces. The MPEG-7 scalable colour descriptor (SCD), which is an encoded HSV histogram-like description of the colour of an image or region (Manjunath et al., 2001), is used as a representation of the extracted body patches. The scalable colour descriptor is *scalable* in the sense that the number of bins used in the feature vector can be scaled. Cooray et al. (2006) show that this feature outperforms other MPEG-7 colour features for body patch matching. In this work we use a 256-bin scalable colour descriptor

Face Colour Feature. We propose, in addition to using conventional face recognition techniques, to use a colour-based feature to represent faces in personal photo collections Our justification for this is that personal photo collections will often

create highly constrained environments for photo capture, with similar lighting conditions and similar photo capture situations This suggests that a simple, colour-based face representation could be useful for person classification, because faces of the same person, within the same context (e.g. the same event), should exhibit similar colour characteristics. In addition to being able to model skin tone under consistent lighting condidtions, we believe that this feature will be able to model other colour-based variations between the faces of people, for example the colour of hair occluding a face, or sunglasses. Again, we use the MPEG-7 scalable colour descriptor, with 256 bins, for this feature.

Image Colour Feature. In personal photo collections, photos which contain the same people are often taken from the same camera set-up; with the photographer often taking multiple near-duplicate photos in quick succession. In addition, photos that are not near duplicates, but taken at the same event, are likely to share visual characteristics due to the fact that they were all taken in the same location under similar lighting conditions This means that photos which contain the same people will sometimes contain the same image-level visual characteristics. Based on these observations, we propose an *image colour* feature for person classification and retrieval, which represents each face using the global, colour-based characteristics of the image containing the face. As with the *body patch* and *face colour* features, we use a 256-bin scalable colour descriptor to represent this feature.

Like the *face colour feature*, this feature is not applicable to the general face recognition task, but under the specific constraints of personal photo collections we believe that both of these features can be useful for content-based person classification and retrieval.

Distance Measure for Content-based Features

Each of the content-based features described above is stored as a separate feature vector for every face. We use Manhattan distance, also known as L1 distance (Duda

et al., 2000), to calculate the distance between two feature vectors, A and B:

$$d_{L1}(A, B) = \sum_i |a_i - b_i|. \quad (5\ 17)$$

To use these content-based features for nearest neighbour person classification and retrieval, we substitute this distance measure, using the feature vector for a given feature, into Equation 5 14

As we are primarily interested in the relative performance of different features for person classificastion and retrieval, and in improvements to be gained from combining these features, we do not explore alternative distance measures or, indeed, alternative colour representations for colour-based features, in this work. Manhattan distance, however, has been shown to give good results for general video retrieval, performing much better than Euclidean distance, for example (Mc Donald, 2005; Mc Donald and Smeaton, 2005).

Event-filtering

The *body patch*, *face colour* and *image colour* features all make the assumption that they will be used in a highly constrained environment, such as within an event, in order to be useful for face classification and retrieval. Accordingly, to take advantage of this assumption, we filter by event when using content-based features. The *event-filtered* content-based approaches only compare candidates with known annotations within the same event. We use a variation of equation 5.14 to calculate an event-filtered nearest neighbour score, replacing the set $\mathbf{b}$ of all faces known to contain person p with the set $\mathbf{c}$ of all faces within the same event as face a known to contain person p:

$$NN_{score}(a, p) = \min_i [D(a, c_i)]. \quad (5.18)$$

For classification, the event-filtered content-based approaches can only suggest names already known to occur in the event, and for retrieval the event-filtered approaches will only suggest faces from events known to contain the query person. We will

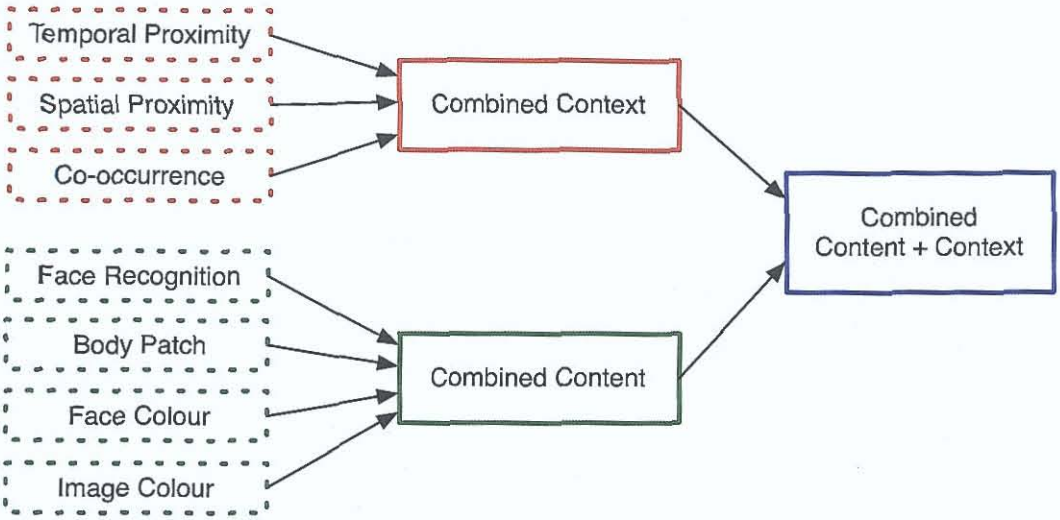


Figure 5.4: Combining individual context-based and content-based approaches to person classification and retrieval.

compare the performance of this *event-filtered* approach with an approach which is unfiltered, or filtered by the user's entire collection, and for this reason we call it *user-filtered.*

5.3 Combined Approaches to Person Classification and Retrieval

We will investigate fusion techniques to combine the results of various context-based approaches to person classification and retrieval, various content-based approaches, and finally we will combine context-based and content-based approaches. This is illustrated in Figure 5.4. We use simple fusion approaches originally developed to combine the results of multiple text search engines, and which have been previously evaluated by Fox and Shaw (1994) and Lee (1997). We will use the *CombSum* fusion technique, which works well in both studies. *CombSum* works by simply summing the normalised scores from each system. We will also use a *Weighted CombSum* approach, which multiplies the normalised score from each system by a weight before summing. Because each system outputs scores with a different range of values, it

is necessary to normalise them so that their scores fall within the same range. We normalise the scores from each individual approach as follows (Lee, 1997):

$$nscore_p = \frac{score_p - score_{min}}{score_{max} - score_{min}}, \tag{5.19}$$

where $score_p$ refers to the score for person p in the classification scenario, and to the score for face p retrieval scenario. This normalisation strategy, then, simply sets the maximum score to 1, the minimum score to 0, and adjusts all other scores to fall between 0 and 1.

5.3.1 Combined Context-based Classification and Retrieval

For combined context-based approaches we will combine approaches based on *temporal proximity*, *spatial proximity* and *co-occurrence* features.

Combined Context-based Nearest Neighbour Classification and Retrieval

For combined context-based nearest neighbour classification and retrieval we combine *temporal proximity* and *spatial proximity* context-based approaches. The distribution of scores for context-based *nearest neighbour* approaches is skewed, with, for example, some photos separated by seconds and others separated months or years. Before combining, if we normalise using equation 5 19, then small differences in the order of seconds or minutes will be rendered insignificant by the normalisation process For example, if the minimum distance is 10 seconds, and the maximum is 1 year (31,536,000 seconds) we will normalise by dividing by the difference between these values. If we divide by 31,535,990 then the difference between 20 seconds and 30 seconds, for example, will be rendered insignificant This will affect combined results using context-based *nearest neighbour* approaches. Also, it should be intuitively clear that the difference between two short time intervals, 1 day and 2 days for example, is not as important, in terms of identity classification, as the difference between two larger time intervals, 11 months and 12 months for example. To

counteract this we can can take the log of the *temporal proximity nearest neighbour* and *spatial proximity nearest neighbour* distances before normalising and combining them with other approaches. In Chapter 6 we will compare combined context-based nearest neighbour approaches which use the raw temporal proximity and spatial proximity distance with approaches that use the log of the distance.

Combined Context-based Language Model

In Section 5.1.3 we described how it is possible to use linear interpolation to create a combined context-based approach to estimating the probability of a person given a specific photo. This language model takes a different approach to combination than that used for nearest neighbour approaches, and no normalisation is carried out on the scores as they represent probabilities in this approach. We treat this language model as a combined approach to facilitate comparison with combined nearest neighbour approaches.

5.3.2 Combined Content-based Classification and Retrieval

For combined content-based approaches we investigate whether it is possible, using simple fusion techniques, to improve content-based classification and retrieval by combining the results of the 4 content-based features that we use. Again, we use that standard *CombSum* and *Weighted CombSum* approaches.

For combined content-based approaches we also compare *user-filtered* and *event-filtered* combined approaches to examine whether the difference between these approaches is consistent with the difference between *user-filtered* and *event-filtered* individual approaches.

5.3.3 Combined Context-based and Content-based Classification and Retrieval

We again use *CombSum* and *Weighted CombSum* to combine context-based person classification and retrieval approaches with content-based approaches When we combine the probabilistic language model approaches with distance-based content-based approaches it is necessary to carry out some pre-processing on the data. This is because, with the language model approach, a higher score indicates higher confidence. With content-based approaches, a lower score indicates higher confidence. In order to make the two approaches compatible, we simply subtract the language model probability from 1. This changes the score from being the probability of a person to being the the probability of 'not the person', making it compatible with the distance-based content-based approaches.

For combined context-based and content-based approaches we will also evaluate whether *user-filtered* or *event-filtered* content-based approaches work best in combination with context-based approaches. Although we expect *event-filtered* content-based approaches to outperform *user-filtered* content-based approaches, it is possible that this difference will be lost when we combine them with context-based approaches. This is because the *event-filtered* approaches already make indirect use of context, making the context-based information somewhat redundant when we combine *event-filtered* content-based approaches with context-based approaches The *user-filtered* approaches are not context-aware, on the other hand, and we would expect them to benefit more from the new context information.

In cases where the best content-based or context-based approach is an individual, rather than a combined, approach, then we will use the individual run in the combination of context and content. For example, this means that we might use *body patch* and *combined context* as a combined context-based and content-based approach.

5.4 Photo-level vs. Face Region-level Classification and Retrieval

So far in this chapter we have assumed that the user wishes to annotate the people in their collection at the level of the face region. In other words an annotation is associated with a face region within a photo. This means that the classification and retrieval approaches that we propose all work at the granularity of the face region. The person classification approaches proposed, given a face, will suggest a list of names, and the person retrieval approaches will suggest a list of faces, given a name

An alternative approach could allow the user to annotate at the level of the photo, associating an annotation with a photo rather than a face region within the photo. Person classification and retrieval systems to support this type of annotation would no longer need to return face region-level results but could instead return photo-level results, as the user would only be interested in annotating at that level. So, for person classification, such a system would take a photo rather than a face as input, and return a list of suggested names for that photo. For retrieval, the system would return a list of suggested images, given a query name Such systems, though, could not make use of face-region level features such as face recognition and body patch, as the system would not have knowledge of names for each face.

Such photo-level classification is naturally supported by the approaches described in this chapter. For context-based features, for example, the score for each face is taken from photo-level features. The same score can be used to rank photos rather than faces, and we can use this photo-level score for photo-level classification and retrieval. If we wish to use region-level features for photo-level classification and retrieval, then for each person name or face to be ranked we can use the minimum score (assuming we are using a distance measure) for that person or face as the photo-level score. As noted above, however, we would not expect such a system to have region-level features available to it.

In our evaluation of classification and retrieval, we will in general evaluate the

accuracy of approaches that use photo-level features at the level of the photo, and we will evaluate the accuracy of region-level approaches (face recognition, body patch, face colour) at the level of the face region To enable comparison with photo-level approaches we will evaluate region-based approaches at the level of the photo in addition to evaluating them at the level of the face region.

5.5 Summary

In this chapter we have proposed a number of context-based approaches to person classification and person retrieval in personal photo collections. For context-based person classification and retrieval we propose two approaches We propose a language model context-based approach to person classification and retrieval which extends an approach which estimates the probability of a person given a specified temporal or spatial context. The extended approach allows for person retrieval in addition to person classification, and also proposes using smoothing and hierarchical smoothing techniques to improve the probability estimates of the language model. The second context-based approach we propose is a nearest neighbour approach, which uses a standard nearest neighbour approach with context-based features, and we use this approach for both person classification and person retrieval.

We also propose using the same nearest neighbour approaches for content-based person classification and retrieval, using features automatically extracted from the visual image content of photos We propose using four different content-based features in this framework, two of which are novel features for this task: *face colour* and *image colour*, both of which we expect to be effective in the highly constrained environment of personal photo collections.

We propose combining various context-based and content-based approaches using standard data fusion techniques based on summing the scores from the individual approaches, and on the weighted sum of the individual approaches.

In the next chapter we will evaluate these proposed approaches, and in Chapter

7 we conduct user experiments with a system powered by these person classification and retrieval techniques.

Chapter 6

Evaluation of Context-Aware Person Classification and Retrieval

In this chapter we evaluate automatic person classification and retrieval in partially annotated, context-aware personal photograph collections. We evaluate the use of contextual features such as temporal proximity, spatial proximity and co-occurrence for person classification and retrieval. We also explore the use of cyclic temporal context, such as the day of the week, as distinct from temporal proximity. We compare the proposed nearest neighbour classification and retrieval to the proposed language model approaches.

We then evaluate content-based approaches, based on analysis of the image content of photos. We evaluate content-based approaches using four different features: face recognition, face colour, body patch and image colour.

Finally, we evaluate techniques for combining both content-based and context-based features, showing how combining the two can improve performance compared with using either in isolation

Throughout this chapter, for brevity we well refer to maximum likelihood estimate (MLE) language models, smoothing language models, and hierarchical smoothing language models as MLE, smoothing and hierarchical smoothing respectively.

User	Total Photos	Faces	Known Faces	Photos With Faces	Photos With Known Faces	Distinct People	Retrieval Queries
A	5,231	655	505	419	341	50	5
B	3,435	2,418	1,774	1,197	1,007	71	17
C	2,672	368	285	240	194	15	6
D	2,128	2,425	2,116	1,139	1,080	148	31
E	1,974	609	397	314	258	23	6
G	1,208	1670	1554	893	885	45	12
H	1,044	847	705	435	392	40	13
J	753	1202	992	488	445	62	17
L	513	687	519	289	246	45	8
Total	18,958	10,881	8,847	5,414	4,848	344(499)	115
Avg	2106.4	1209	983	601 6	538.7	(55.4)	12.8

Table 6.1: Statistics for the presence of faces in the collections of all users included in evaluation. User labels are consistent with Table 3.1. The total number of distinct people is not equal to the sum across all users: this is because some people are present in multiple collections (the number in brackets is the sum across all users).

6.1 Test Collection

The MediAssist personal photo archive, described in Section 3 2 in Chapter 3, contains 23,774 photos from 29 users. Of these, 9 users have collections that we consider suitable for evaluation of automatic person classification and retrieval approaches We consider other users' collections unsuitable because they do not contain enough known people.

As the focus of this evaluation is on person classification and retrieval we do not want the effect of imperfect face detection to add noise to our results Accordingly, the author of this thesis manually annotated the presence of all faces in the collection, which effectively assumes that our system has perfect face detection This gives us a richer set of faces for evaluation, and allows our evaluation to focus exclusively on person classification and retrieval, rather than having to cope with noise from incorrectly detected faces. After manually annotating the presence of all faces in the 9 test collections the author consulted each of the users to elicit names for these

manually annotated faces. Based on these consultations, the author then manually annotated the names of all of the faces in these 9 user collections, and it was this annotation that was used as a ground truth for evaluation.

Table 6.1 summarises the 9 personal photo collections used in the evaluation in this chapter. These 9 user collections contain a total of 18,958 photos, 5,414 of which contain faces. These contain a total of 10,881 faces, of which 8,847 are faces of known people. There are an average of over 55 distinct known people per user collection.

As discussed in Section 3.2, we believe that the MediAssist photo archive is broadly representative of personal photo collections as we requested that the users give us any personal photos that they were willing give give, without pre-selecting certain types of photos or specifically requesting certain types of photos from the users. The people in the photos will vary in terms of their relationship to the user, from friends and family to colleagues and even strangers. The locations of the photos vary from photos taken at home to parties to photos taken on holidays or on business trips. Low quality photographs like blurred photographs and overexposed photographs have been left in the collection, making the collection even more challenging for person classification and retrieval.

For evaluation purposes, each collection is split into a training set and a test set of equal size. We are modelling the situation where the user has 50% of their collection already annotated and the system is using this information to learn suggested annotations for the remaining 50% of the collection. Although this assumes a high level of prior annotation, we believe this is necessary for evaluation purposes. If we were to use a smaller training set, which more accurately models the starting point of a real user, it is likely that while the proposed techniques would still work, the lack of training data would minimise the differences between system variants. Given this, we prefer a rich training set that allows the system to learn properly from all available features and which therefore allows us to explore the utility of all features.

6.2 Experimental Methodology

6.2.1 User Annotation Model

We use a simple approach to modelling the user annotation process, essentially assuming that the user annotates all identities in their collections, including unknown people, in a random order. To model this our system creates a random partition of all annotations into training and test sets, modelling the situation where the user has 50% of their collection randomly annotated. All of the results reported in this chapter use this model of user annotation. We include annotation for unknown people because our interactive system also encourages the user to label unknown faces as 'unknown', as described in Chapter 4. We encourage users to label unknown faces because this information can be used by our retrieval system to filter the results of queries for people, since faces labelled as 'unknown' cannot be relevant to queries for specific individuals Our results from interactive user experiments in Chapter 7 will show that, given the appropriate user interface, users can be encouraged to annotate unknown people in their personal photo collections.

6.2.2 Evaluation Measures

In this section we outline the evaluation measures used for person classification and person retrieval.

Evaluation Measure for Person Classification

The person classification task takes a face as input and returns a candidate list of suggested names for this face as output, as described in Chaper 4. The assumption is that the user will be satisfied if the correct identity is present towards the top of this list. Chen et al. (2003b) propose a measure called *H-hit rate* to evaluate this scenario. This measure takes a list of H suggested names for a given face, and if the correct name is present in this list then this is considered a 'hit'. The *H-hit rate* is simply the proportion of H-hits within the collection. If F is the set of known faces

to be evaluated, and N is the number of elements in this set, then:

$$H\text{-Hit} = 1/N \sum_{f \in F} \text{hit}_H(f) \tag{6.1}$$

where $hit_H(f)$ is 1 if f is present in the list of H suggested names, and 0 otherwise.

We follow Naaman et al. (2005) by using a value of 5 for H in our evaluation. We believe that this is a reasonable size for a list of name suggestions to return to a user, although we recognise that user trials would be needed to determine the optimum length of a suggestion list.

Photo-level vs Face Region-level Evaluation. The evaluation of person classification is carried out at the face region-level, meaning that we are concerned with classifying faces, and a face label is a *hit* if it correctly labels a specific face region An alternative would be photo-level classification, where the task would be to correctly identify people occurring in a photo, rather than to label specific faces within a photo, which could be considered suitable if we are using photo-level features. We note here that, if we were to use photo-level evaluation to evaluate approaches based on photo-level features (e g context-based approaches), then the results will actually be the same as if we have used face region-level evaluation This is because, with these systems, all unannotated faces in the same photo will have the same classification results. a face region-level evaluation will count how many of these faces are correctly classified, while a photo-level evaluation, using *H-hit rate,* will count, out of the top h name suggestions, how many are actually in the photo. Accordingly, while we restrict our evaluation of person classification to face region based evaluation, for the photo-level approaches the evaluation measure can also be understood as a measure of photo-level accuracy.

Evaluation Measures for Person Retrieval

For the person retrieval task, the system takes a query for a given person and returns a ranked list of photos, or face regions within photos, that should contain the query person. It is essentially an information retrieval task, with the information need corresponding to photos or face regions of a specified person. Precision and Recall are the traditional measures used in information retrieval. Precision is the proportion of retrieved documents that are relevant. Recall is the proportion of relevant documents that are retrieved. These measures must be reported together because a result set giving a high precision could have low recall, and vice versa, so it is the combination of precision and recall that is important.

Precision at document cut-off is a single-valued measure of retrieval performance with *Precision at X* equal to the precision after X documents have been retrieved. Because our system encourages users to find relevant documents towards the top of the result set for the purposes of confirming annotations, and because many queries have less than 30 relevant documents, we use *Precision at 10*, *20* and *30* (*P10*, *P20* and *P30*) in our evaluation.

Average Precision is another single-valued measure of the effectiveness of a retrieval run. It is the average of the precision when each relevant document is found in the ranked list. Non-retrieved relevant documents are given a precision of zero. *Mean Average Precision (MAP)* is used to measure performance over a number of queries, and is calculated by averaging the *Average Precision* for each query.

Photo-level Evaluation. For the context-based approaches we evaluate *MAP* and *Precision at 10*, *20* and *30*. We evaluate at the photo level, considering a photo to be our unit of retrieval, with a photo labelled as relevant if it contains the query person. For a context-only system we believe that photo-level annotation and retrieval is sufficient, since all faces in an image have the same context and a user should be happy if an image is returned in response to a query, with little added incentive to specify the region within the image containing the person.

Face-Region level Evaluation. For content-based approaches, regional information within the photo-image about the face and body patch are required. As previously described in Chapter 6, the system needs to know the location of a face within a photo to enable it to analyse the face and body patch regions, facilitating the content-based approaches to person classification and retrieval described in Chapter 5 Because of this it is necessary to annotate images at the level of a specific face within a photo, not just at the photo level, if we are to benefit from content-based approaches. For this reason we expect a content-based system to suggest a specific face region within a photo in response to a query, not just a photo To enable comparison of the content-based results with the context-based results, we evaluate the content-based results at the photo level. However, we also evaluate our content-based systems at the face region-level, treating a face region, not a photo, as the unit of retrieval and labelling a face region as relevant if it corresponds to the query person.

6.2.3 Cross-Validation

Rather than create one partition between training and test sets, we use cross-validation to create a number of independent partitions. We use *hold-out* cross-validation, where a subset is randomly selected as the training data and the remainder of the data is used in testing (Devijver and Kittler, 1982), to create 5 independent partitions of our data in training and test sets. We evaluate separately over each partition and then average the results over each partition. This avoids any accidental bias in the creation of the training and test partitions. In all the results reported in this chapter, *H-hit rate*, *MAP*, and *Precision at 10, 20* and *30* (*P10*, *P20* and *P30*) are averaged over the results from 5 different partitions of the data into training and test sets

6.2.4 Statistical Tests

Since we calculate our evaluation measures for each user, and then average over users, it is quite possible that outlier results from a minority of users, caused by chance, could bias the results. For this reason it is necessary to perform statistical tests, which test the hypothesis that two results are significantly different to the null hypothesis that the difference between the results is due to chance. Statistical tests are run at a specified significance level, which is the probability that the difference between the two runs is due to chance. For example, if a test is deemed significant at a significance level of 0 05, it means that, according to the test, there is a probability of 0.05 that the difference is due to chance, or that we are 95% confident that there is a significant difference between the runs.

In this thesis we use randomization testing (Kempthorne and Doerfler, 1969) to test for statistical significance. The input into the tests are the *H-hit rate*, *MAP*, *P10*, *P20* or *P30* results for each user for the runs to be compared, so we are effectively testing if the differences between runs are consistent across users, or if they are due to chance variations between users. We test at two different significance levels, 0 05 and 0 01 If a run is better than another run at a significance level of 0.05 we will say it is *significant* If a run is better than another run at a significance level of 0 01 then we will say it is *highly significant.*

6.3 Outline of Experiments

In this section we will give an outline of the evaluation results reported in this chapter.

Context-based Person Classification Experiments

In Section 6 4 1 we will evaluate context-based approaches to person classification. We firstly evaluate the impact of smoothing approaches for statistical estimation and how they improve performance compared with MLE. These results are shown for

temporal proximity in Figure 6.1 on page 122, for spatial proximity in Figure 6.2 on page 123, for co-occurrence in Figure 6.3 on page 124 and for cyclic temporal context in Figure 6.4 on page 125 From these figures we can compare the performance of different spatial and temporal windows for spatial and temporal proximity. In addition, we compare the results from fixed time windows with those from automatically detected events and sub-events.

In Figures 6.5 and 6 6 on pages 128 and 129 we look at the improvements to be gained from 3-layer hierarchical smoothing techniques for spatial proximity, temporal proximity and co-occurence features, comparing each with the best-performing smoothing run, and comparing the performance of different hierarchical structures. Then in Figure 6.7 on page 130 we compare the best MLE, smoothing, 3-layer hierarchical smoothing and 4-layer hierarchical smoothing approaches with the nearest neighbour approach for both the spatial proximity and temporal proximity features. In Figure 6.7 we also compare hierarchical smoothing results using spatial proximity with the results using temporal proximity, to determine which is the best feature for person classification.

Context-based Person Retrieval Experiments

In Section 6.4.2 we evaluate context-based approaches to person retrieval as opposed to person classification. In Figures 6.8 and 6.9 on pages 134 and 135 we examine the improvements to gained from smoothing compared with MLE for temporal proximity, spatial proximity and co-occurrence using the MAP evaluation measure, and we compare this improvement to the improvement gained in the classification scenario.

In Figures 6.10, 6.11, 6.12 and 6.13, on pages 136, 138, 139 and 139, we examine the retrieval results for temporal proximity, spatial proximity, co-occurrence and cyclic temporal context smoothing approaches for the evaluation measures MAP, P10, P20 and P30. We compare the relative performance of different system variants for each evaluation measure, in particular noting that approaches that work well for MAP are not always the best-performing approaches for P10.

In Figures 6.14 and 6 15 on pages 142 and 143 we examine 3-layer hierarchical smoothing results for person retrieval, noting a similar improvement to that found in the classification scenario. The results for 4-layer hierarchical smoothing in addition to those for nearest neighbour retrieval are examined in Figures 6.17 and 6.18 on pages 145 and 146 They show an improvement over 3-layer hierarchical and confirm the strong performance of temporal proximity nearest neighbour shown for the person classification scenario. These figures also allow us to compare the relative performance of temporal proximity and spatial proximity, confirming the better performance of temporal proximity.

Content-based Person Classification Experiments

In Section 6 5.1 we evaluate content-based approaches to person classification. In Figure 6.19 on page 149 we examine the 5-hit rate for each of the 4 content-based features used: image colour, body patch, face recognition and face colour. We compare the performance of user-filtered and event-ftilered approaches, with *user collection MLE* and *event MLE* language models used as a baseline. In Figure 6.20 on page 150 we take a closer look at the event-filtered content-based results, comparing the *H-hit rate* performance of each of the content-based features for various values of H, showing that while body patch is the best-performing feature for *1-hit rate*, image colour is the best feature for *5-hit rate*.

Content-based Retrieval Experiments

We then, in Section 6.5.2, examine the performance of content-based features for person retrieval. In Figure 6.21 on page 154 we examine the content-based results for the photo-level retrieval scenario, where the task is to retrieve a photo containing the query person. We show that, for this scenario, the best-performing user-filtered feature is image colour, while the best-performing event-filtered feature is body patch, with event-filtering outperforming user-filtering In Figure 6 22 on page 155 we present the content-based retrieval results for the face region-level

retrieval scenario, where the task is to retrieve a face region corresponding to the query person Again, the best-performing user-filtered feature is image colour and the best-performing event-filtered feature is body patch, and event-filtering outperforms user-filtering.

Combined Context-based and Content-based Classification Experiments

In Section 6 6.1 we evaluate combined approaches to person classification

Combined Context Classification. We firstly evaluate combined context-based approaches, which combine the results of various context-based features. In Figure 6 23 on page 159 we evaluate combined approaches based on language models with combined approaches based on nearest neighbour. In each case we compare the relative performance of unweighted and weighted combinations to determine if it is necessary to learn combination weights to achieve an improvement. We show that the while weights are necessary when combining language models, it is possible to achieve improved performance with an unweighted combination of nearest neighbour approaches. For hierarchical language models we also evaluate a combined approach that does not use location information in order to determine how well context-based classification can perform if location information is not available

Combined Content Classification. We then evaluate combined content-based approaches to person classification, which combine the results from each of the 4 content-based features that we use. In Figure 6.24 (a) and (b) on page 161 we compare the performance of combined content-based classification approaches, both for user-filtering and event-filtering, showing the improvement given if we combine features. We also compare the performance of weighted and unweighted combination approaches, showing that for content-based classification there is little benefit to be gained from learning combination weights. In Figure 6.25 on page 162 we compare user-filered and event-filtered combined content-based approaches, showing that while event-filtering performs slightly better for *1-hit rate* user-filtering

performs better for *5-hit rate*

Combined Context and Content Classification. After evaluating combined context-based classification and combined content-based classification, we then evaluate classification based on a combination of context and content In Figure 6.26 on page 165 we evaluate combined approaches that use language models as the context-based approach. For the language model approaches we only evaluate weighted approaches, as we previously showed that unweighted combination is ineffective for language models. We also compare the results of combined approaches that use user-filtered content-based approaches with those that use event-filtered content-based approaches, showing that if we use event-filered content-based approaches the combination is better. In this figure we also evaluate combined approaches based on the use of image colour only as the content-based feature to evaluate the performance that is achievable without region-level features

Figure 6.27 on page 166 evaluates the performance of combined approaches which use nearest neighbour as the context-based approach, and in this figure we also evaluate unweighted combined context-based and content-based approaches, showing that it is not necessary to learn weights to improve performance by combining context and content.

In Figure 6.28 on page 167 we compare the *H-hit rate*, for various values of H, of combined context-based, combined content-based and combined context-based and content-based approaches to person classification, in each case also comparing the performance of the best unweighted and the best weighted approach, showing that content-based approaches perform better than context-based approaches for *1-hit rate* while context-based approaches perform better for *5-hit rate*. Combined context-based and content-based approaches outperform context-only and content-only approaches for all values of H, with little difference between weighted and unweighted approaches that combine context and content.

Combined Context-based and Content-based Retrieval Experiments

In Section 6.6.2 we evaluate combined approaches to person retrieval.

Combined Context Retrieval. Combined context-based retrieval approaches are evaluated in Figure 6.29 on page 171. As with the classification scenario, combined approaches based on language models are compared with those based on nearest neighbour, and we again show that we can improve performance with unweighted combined nearest neighbour approaches. We also evaluate the performance of combined approaches without the location information.

Combined Content Retrieval. In Figure 6.30 on page 174 we show combined content-based results for the photo retrieval scenario. We compare the results given by user-filtering with those given by event-filtering, showing that event-filering performs better for person retrieval for all evaluation measures. We also show that unweighted combined approaches improve performance compared with individual approaches. In Figure 6.31 on page 175 we evaluate combined content-based approach for the face region retrieval scenario, showing similar results to the photo retrieval scenario.

Combined Context and Content Retrieval. Finally, in Figures 6.32 and 6.33 on pages 178 and 179 we evaluate approaches to person retrieval based on the use of both context and content, for the photo retrieval and face region retrieval scenarios. We again compare weighted approaches with unweighted approaches, and we compare approaches which use location information with those that do not use location information. We show that for the retrieval scenario, in addition to the classification scenario, unweighted nearest neighbour combinations of context and content improve performance. We also compare the best photo-level retrieval results based on photo-level features with the best face-region level results based on region-level features in order to determine if there is any real benefit in annotating at the region level, and we show that there is little real difference between the two.

6.4 Evaluation of Context-based Approaches to Person Classification and Retrieval

6.4.1 Context-based Person Classification

For the evaluation of person classification, we evaluate all faces of known people in the training set. There are a total of 8,847 known faces across all 9 collections, which is an average of 983 known faces per collection. The test partition contains half of these faces: approximately 4,400 faces over 9 collections, an average of over 480 faces per collection to be evaluated.

For our evaluation of person classification, the weights for all language model smoothing approaches are learned on the test set by optimising the evaluation criterion, *5-hit rate*, and they so are biased, 'oracle' weights and cannot be said to represent weights that we could expect a system to learn automatically. However, by learning these oracle weights we can discover which smoothing schemes are the most powerful for person classification and retrieval, giving a useful upper bound on performance. In this section, and elsewhere in this chapter, the smoothing method used for estimating the probabilities of the language model is Jelinek-Mercer smoothing (Zhai and Lafferty, 2001), which is described in Chapter 5

For this evaluation we used two versions of our system. The first version considers all identities in the user's collection as possible candidates for classification, and so ranks all identities for each query. The second version, following Naaman et al. (2005), makes the assumption that the user is only interested in annotating the most popular identities in their collection and so only considers the most frequently occurring identities for classification We use the top 20-most popular people in this version, meaning that only faces belonging to the 20-most frequently occurring identities in each user collection are evaluated, and the list of identity suggestions is filtered to only rank these 20 identities instead of ranking all identities

Figures 6 1, 6.2, 6.3 and 6.4 show *user collection MLE* results as a baseline against which other approaches are compared. They show the results based on the global, background probabilities of occurence for identities given the user's collection, and these provide the background probabilities used for the smoothing techniques. For classifying all identities, the *user collection MLE* approach achieves a 5-hit rate of 0.528 for classifying all documents and 0 61 for classification of the top 20 documents. These results can be viewed as a baseline against which other approaches are compared.

Figure 6 1 shows classification results using temporal proximity, giving results for *temporal proximity MLE* and *temporal proximity smoothing*, both of which calculate the probability of a person being present, given a temporal window around the photo to be classified. *Smoothing* always gives an improvement over *MLE*, the improvement being similar for top 20 classification and classification of all identities For all time windows with the exception of *1 year* this improvement of smoothing over both *MLE* and *user collection* classification is highly statistically significant for *1 year* the improvement is statistically significant rather than highly significant. The summarise top 20 classification *temporal proximity smoothing* approach gives a 5-hit rate of 0 835 (using the *2 day* time window, which marginally outperforms time windows between *Event*, *6 hours* and *1 week*). For top 20 classification *2 day smoothing* is significantly better than all other *temporal proximity smoothing* runs, and highly significantly better than the following runs *10 seconds*, *60 seconds*, *5 minutes*, *subevent* and *1 year* Statistical significance tests on the results of classifying all identities give similar results. The fine-grained classifiers benefit most from smoothing, which should be expected since it is these which are most likely to suffer from missing or sparse data with *MLE*.

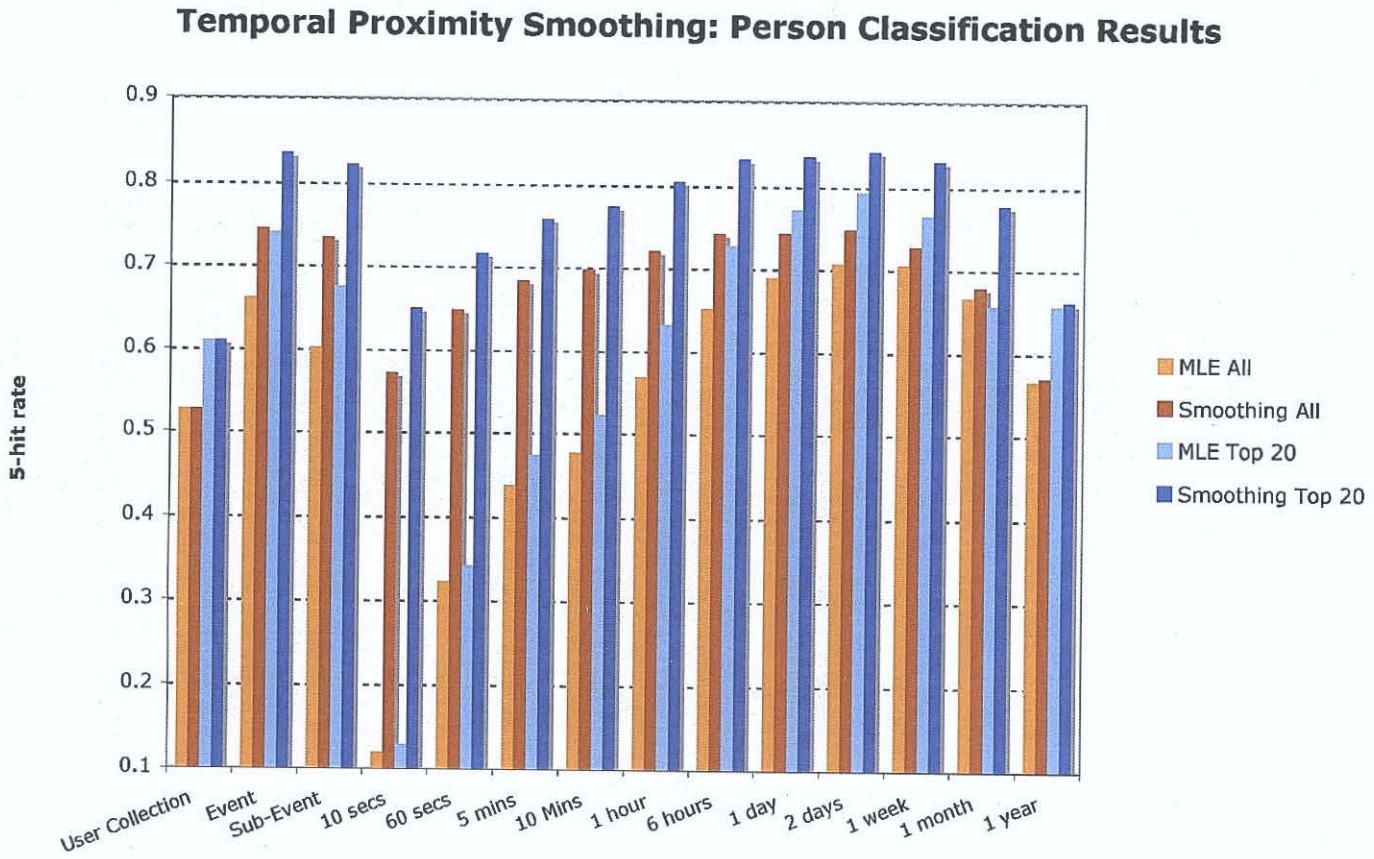


Figure 6.1: Context-based person classification results using Temporal Proximity MLE and Temporal Proximity Smoothing.

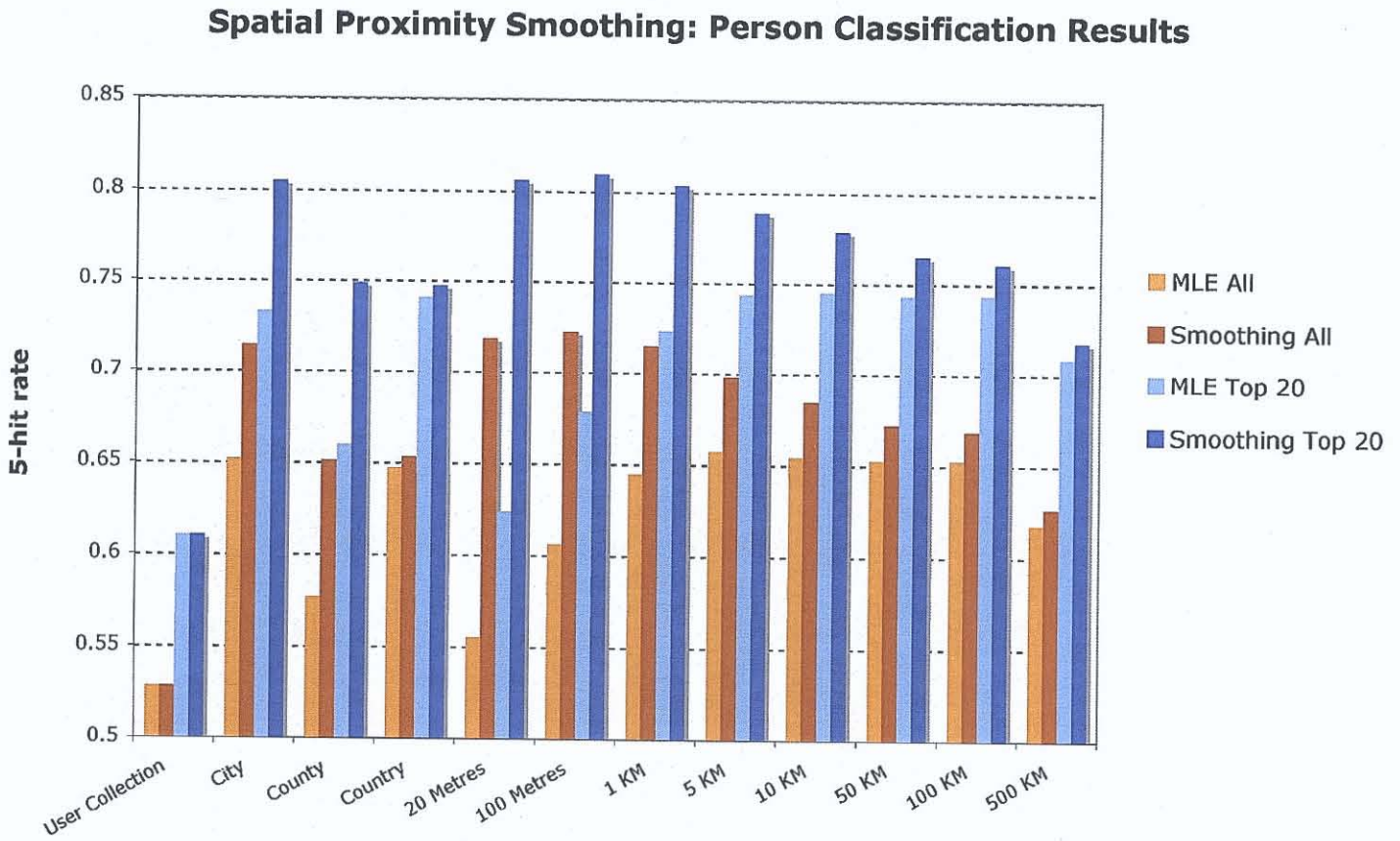


Figure 6.2: Context-based person classification results using Spatial Proximity MLE and Spatial Proximity Smoothing.

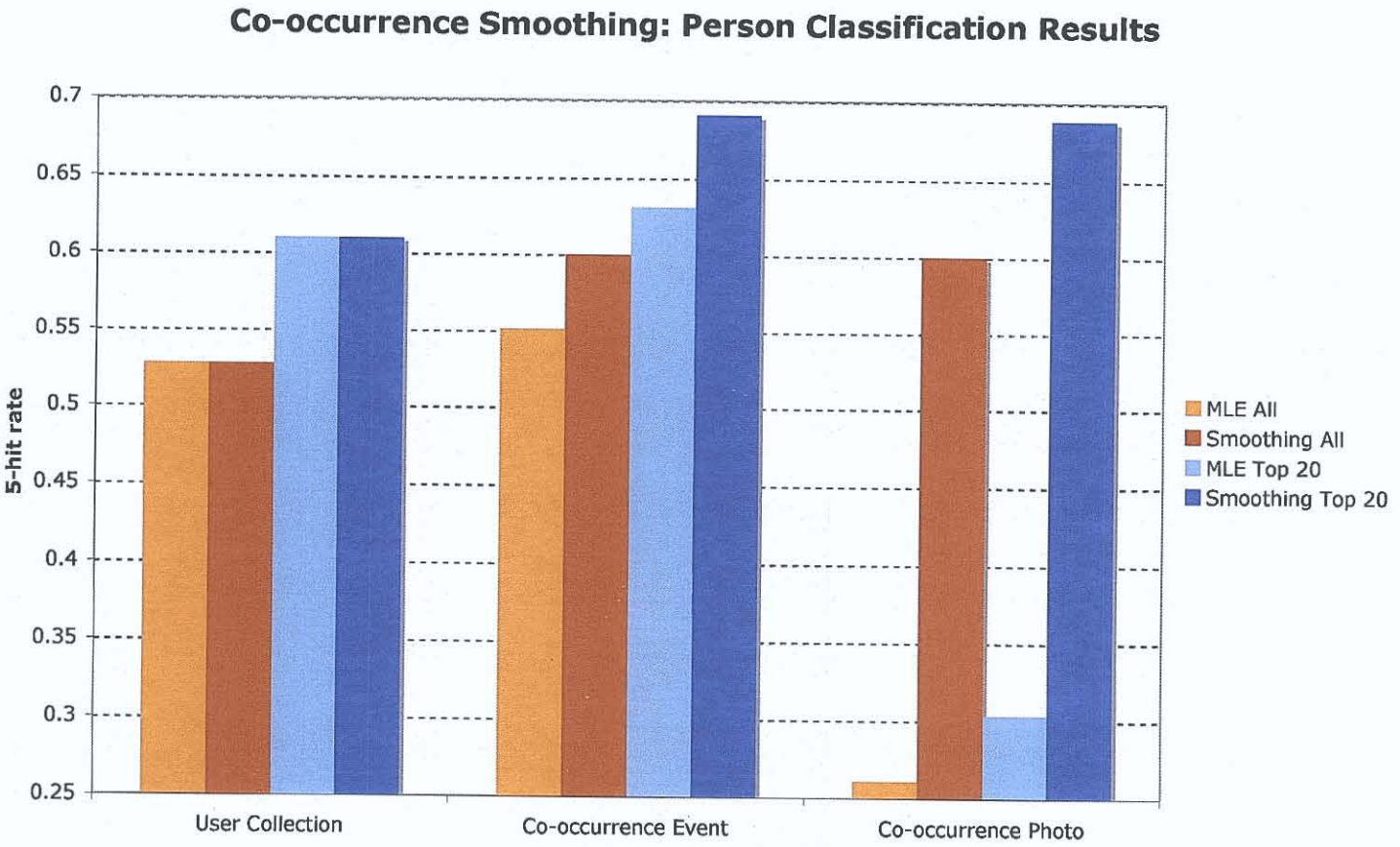


Figure 6.3: Context-based person classification results using Event Co-occurrence and Photo Co-occurrence.

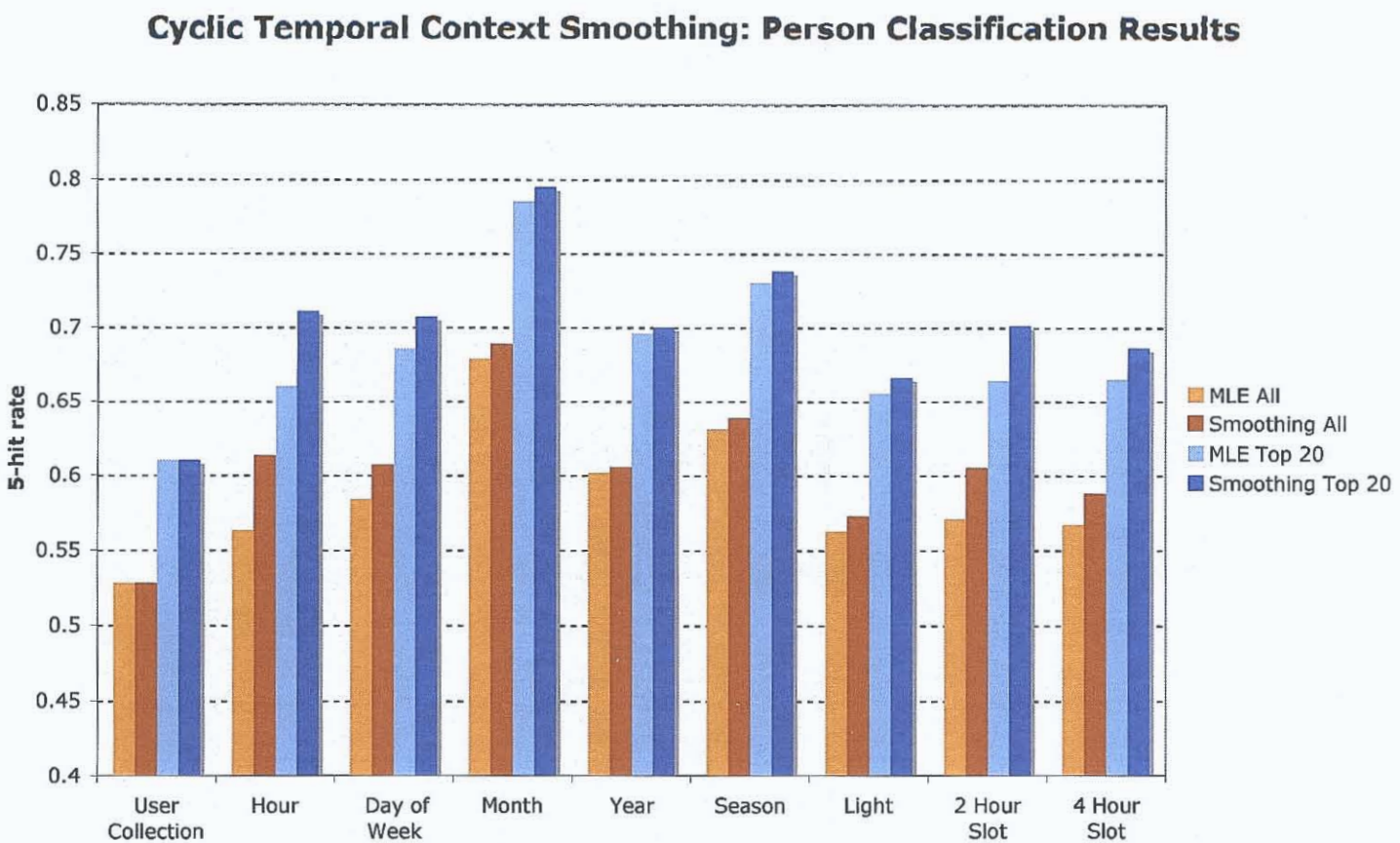


Figure 6.4: Context-based person classification results using Cyclic Temporal Context MLE and Cyclic Temporal Context Smoothing.

Figure 6.2 shows results for location-based classification, giving results for *spatial proximity MLE* and *spatial proximity smoothing* Again, all *smoothing* runs are highly significantly better than *user collection. Smoothing* is highly significantly better than *MLE* for all sizes of location windows.

Comparing *spatial proximity* with *temporal proximity*, the best *spatial proximity smoothing* result is 0 81 for top 20 classification, compared with a best result of 0.836 for *temporal proximity smoothing* This difference is highly significant for both *MLE* and *smoothing*, and for both top 20 classification and classification of all identities.

Results of co-occurrence-based classification are shown in Figure 6 3 *Event co-occurrence MLE* outperforms *photo co-occurrence MLE*, which is not surprising because missing and sparse data is a much greater problem for *photo co-occurrence*, given that many photos contain only one face, in which case photo-based co-occurrence is useless. This result is highly statistically significant and reinforces the results reported by Naaman et al (2005). Again the results show that smoothing gives a highly significant improvement in performance There is no significant difference, however, between *event co-occurrence smoothing* and *photo co-occurrence smoothing.* We will see later, however, that if we use hierarchical smoothing to combine these two types of co-occurrence then performance is significantly improved, showing that two approaches complement each other.

Figure 6 4 shows the results using *cyclic temporal context*, such as the day of the week or the month of the year. There is, of course, an overlap between these approaches and *temporal proximity* approaches. For example, if a photo was taken in the middle of December and we use two week *temporal proximity*, then we will use appearance frequencies for December to calculate the probabilities of each individual appearing. Using cyclic temporal context we calculate probabilities based on any December, and when a collection does not span many years the probabilities for each system could be quite similar. The performance of *cyclic temporal context* is in general much lower than *temporal proximity.* For example *1 day temporal proximity smoothing* gives a 5-hit rate of 0.832, whereas *day of week cyclic temporal*

context smoothing gives a 5-hit rate of 0.707, a highly significant difference. *Month cyclic temporal context smoothing*, however, outperforms *1 month temporal proximity smoothing*, suggesting that this is a useful feature, perhaps because people tend to re-appear on an annual basis for occasions such as Christmas and birthdays. This result is not statistically significant, however, suggesting that this type of re-occurence only applies to the collections of a minority of users.

Context-based Hierarchical Smoothing and Nearest Neighbour Person Classification

Figures 6.5 and 6.6 show results for hierarchical smoothing language models using a 3-layer hierarchy. The results for time, location and co-occurrence approaches are compared to the best-performing smoothed approach in each case. We restrict our analysis to top 20 classification because we have shown in the previous section that the same patterns hold for top 20 classification as with classification of all identities. For all approaches *3-layer hierarchical smoothing* gives a large improvement over *smoothing*, improving *temporal proximity* from a best 5-hit rate of 0.835 to a *3-layer hierarchical smoothing* best 5-hit rate of 0.858 and improving *spatial proximity smoothing* from 0.81 to 0.833 co-occurrence is improved from 0.691 to 0.748. In all cases, the best *3-layer hierarchical smoothing* approach performs significantly better than the best *smoothing* approach.

Looking at the best-performing runs reveals that hierarchical smoothing does indeed benefit from the fine-grained proximity windows by smoothing their results with coarser windows. For example, the best *temporal proximity* approach, *user→day→60 secs*, uses *60 seconds* at the fine end of the hierarchy, even though *60 seconds temporal proximity smoothing* performs poorly in isolation

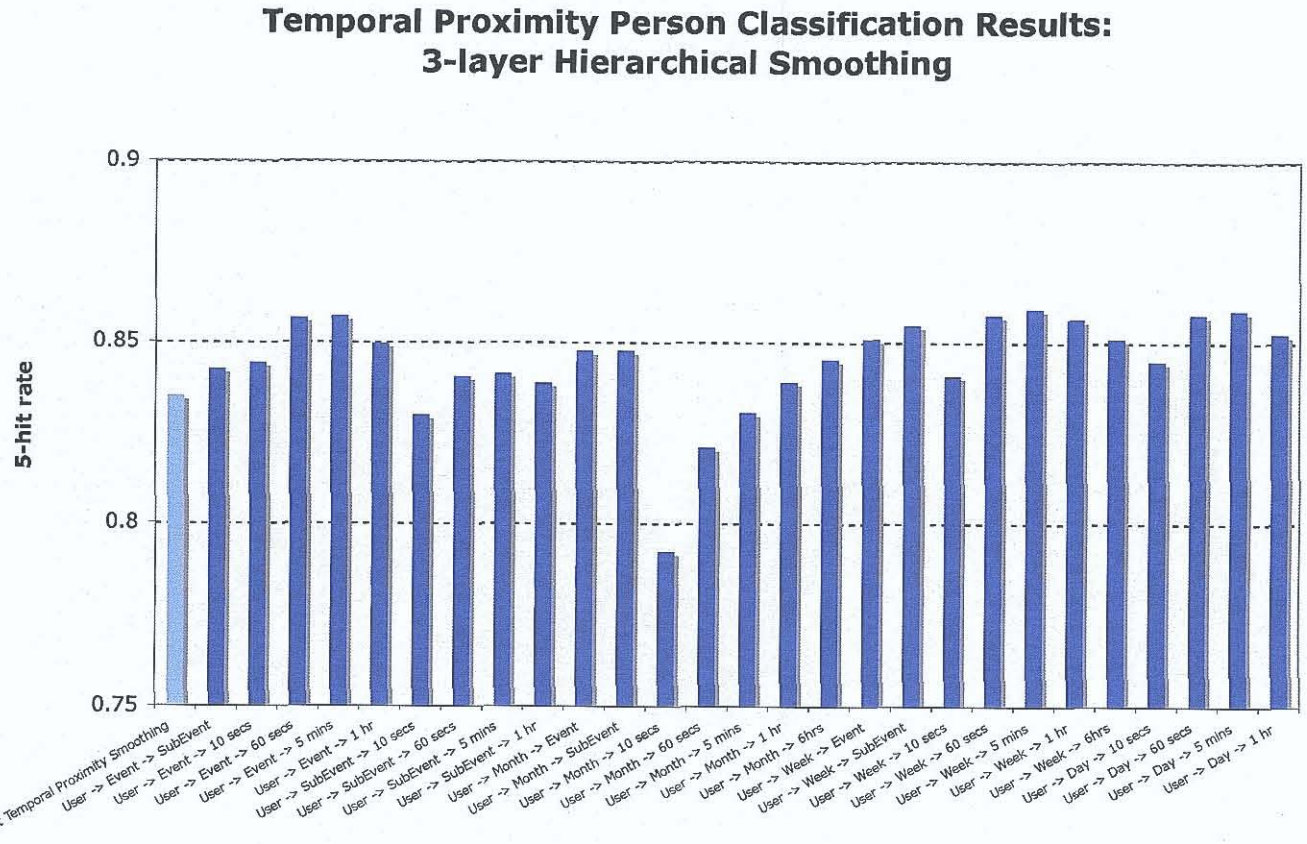


Figure 6.5: Context-based person classification results using Temporal Proximity 3-layer Hierarchical Smoothing.

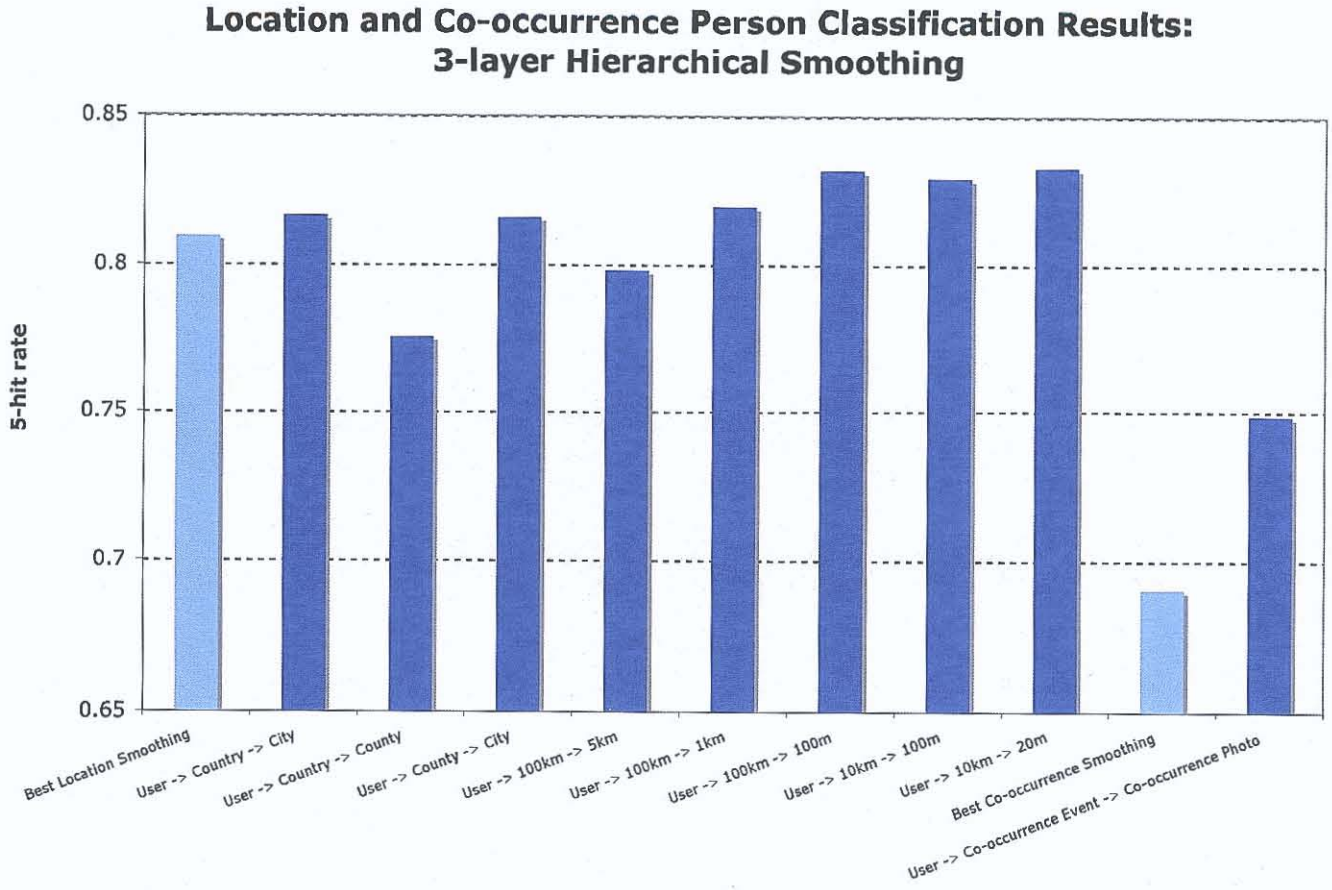


Figure 6.6: Context-based person classification results using Spatial Proximity and Co-occurrence with 3-Layer Hierarchical Smoothing.

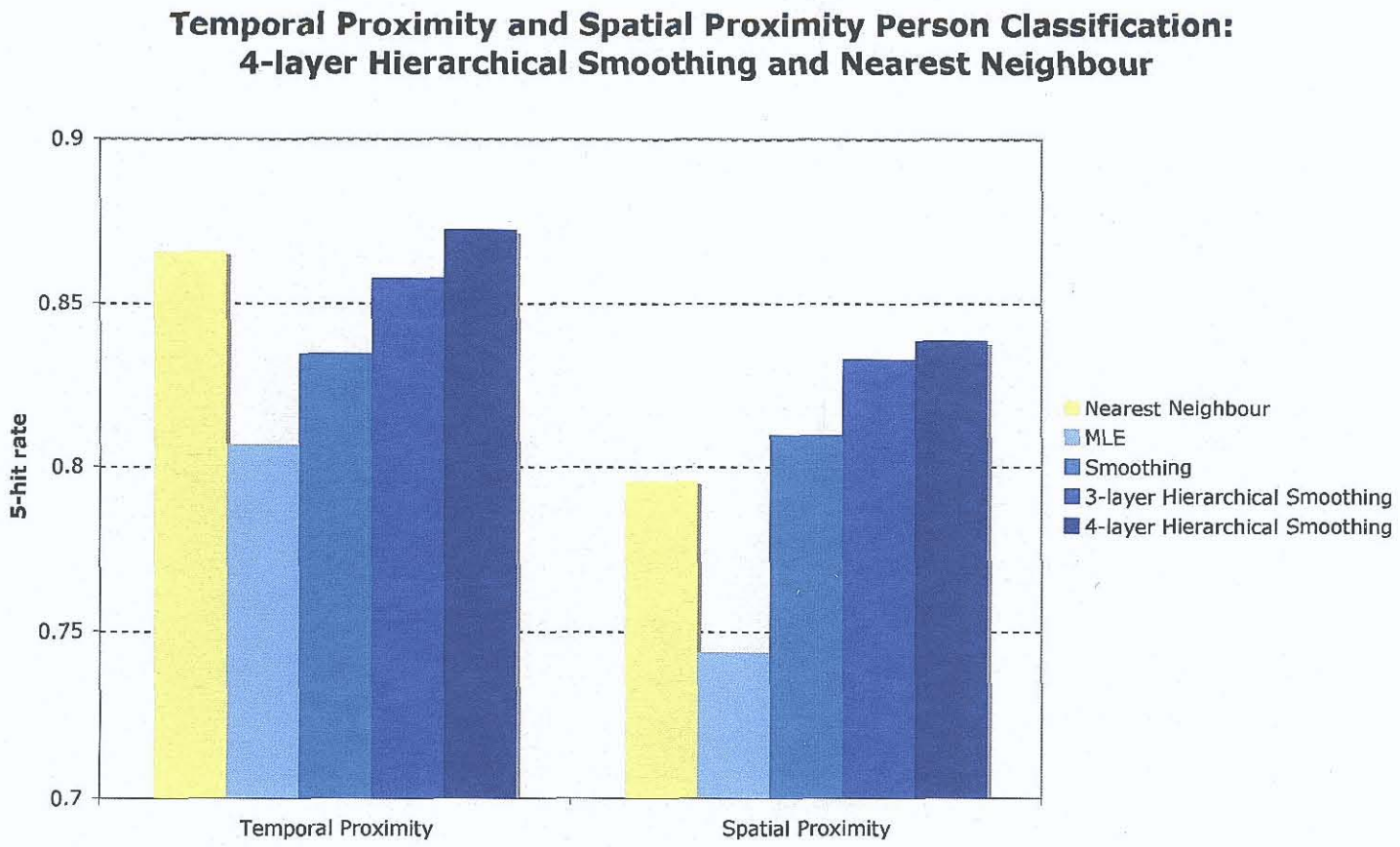


Figure 6.7: Context-based person classification results for Nearest Neighbour, MLE, Smoothing, 3-Layer Hierarchical Smoothing and 4-Layer Hierarchical Smoothing using Temporal Proximity and Spatial Proximity.

Figure 6.7 shows results for *temporal proximity* and *spatial proximity* using a 4-layer hierarchy, in addition to showing the results for *nearest neighbour classification*. For both *temporal proximity* and *spatial proximity*, the best-performing *4-layer hierarchical smoothing* approach outperforms all other approaches. We can see a steady increase in performance as we move from *smoothing*, and on to *3-layer hierarchical smoothing* and *4-layer hierarchical smoothing*. For both *temporal proximity* and *spatial proximity*, the summarise *4-layer hierarchical smoothing* approach performs significantly better than the best *3-layer hierarchical smoothing*, *smoothing* and *MLE* approaches. *4-layer hierarchical smoothing* gives a best 5-hit rate of 0.872 for *temporal proximity* and 0.839 for *spatial proximity*. The best-performing 4-layer hierarchies are *user→week→subevent→300secs* and *user→100km→1km→20m*, again showing how hierachical smoothing makes use of fine-grained approaches that do not perform well in isolation.

From Figure 6.7 we can also see that the relatively simple *nearest neighbour* approaches perform surprisingly well *Temporal proximity nearest neighbour*, with a *5-hit rate* of 0.867, outperforms all *temporal proximity smoothing* approaches except *4-layer hierarchical smoothing*, with a *5-hit rate* of 0.872. It performs significantly better than *temporal proximity MLE* and *temporal proximity smoothing*, and there is no significant difference between *nearest neighbour temporal proximity* and both *temporal proximity 3-layer hierarchical smoothing* and *nearest neighbour 4-layer hierarchical smoothing* There is no need to learn optimal weights for any parameters for the *nearest neighbour* approach, which gives it an advantage over *smoothing* approaches.

Spatial proximity nearest neighbour does not perform quite so well, however, although it does perform significantly better than *spatial proximity MLE*. All other *spatial proximity* approaches outperform *nearest neighbour*, and this difference is statistically significant for both *spatial proximity 3-layer hierarchical smoothing* and *spatial proximity 4-layer hierarchical smoothing* We believe the reason for the poorer performance of *spatial proximity nearest neighbour* is that many photos will have

identical or near-identical locations, meaning that when we rank by distance from a given location we will have a lot of known identities with identical locations and therefore identical distance, and the system is not able to rank these based on location. This problem is exacerbated by the asynchronous nature of the GPS data capture, as described in Chapter 3, which means that the location information is not as accurate as it would be if it was captured synchronously. The time-based system does not suffer from the same problem as every image has a unique timestamp In addition, it is intuitively more reasonable to assume that a person will appear in a photo if he or she appeared 1 minute previously than it is to assume that that person will appear in a photo if he or she was once in a photo taken 10 metres away.

6.4.2 Context-based Person Retrieval

For the evaluation of face retrieval we chose a number of people as queries from each user's collection. Any person with an occurrence frequency of 20 or more was chosen. This means that for the average partition between training and test sets, there will be at least 10 relevant images for each query. We considered people with a lower frequency than this to be unsuitable for evaluation. This gives us a total of 115 queries from 9 users, an average of 12.8 queries per user, as shown in Table 6.1. The person with the greatest occurrence frequency occurred 539 times, with an average frequency of 57.4 for the query people. This means that, for the average partition between training and test sets, there is an average of about 29 relevant photos per query, with a maximum of 270 relevant photos for one query. We evaluate context-based approaches at the photo level, the task being to retrieve photos in that user's collection containing the query person.

We only rank photos containing faces, assuming that our system has perfect face detection. We can see from Table 6.1 that the average collection has 601 photos containing faces. Half of these will belong to the test set for any partition, giving an average of 300 images containing faces in each user's test collection, which is the average number of images to be ranked in the person retrieval task.

For our evaluation of person retrieval, the weights for all smoothing approaches are learned on the test set by optimising the evaluation criterion, *MAP*. As with the classification task the learned weights are biased, but using biased weights allows us to discover which smoothing schemes are the most powerful for face retrieval.

Context-based Smoothing Approaches to Person Retrieval

Figures 6.8 and 6.9 compare *MLE* and *smoothing* retrieval *MAP* results for *temporal proximity*, *spatial proximity* and *co-occurrence*. As with the classification scenario, smoothing improves performance in all cases. However, for retrieval the difference is not always statistically significant. For *temporal proximity*, *smoothing* is significantly better than *MLE* for all time windows shorter than a day, but not for time windows longer than a day. For *spatial proximity*, *smoothing* is significantly better than *MLE* for all runs except *country* and *500km*.

We believe that the reason that *smoothing* is not always significantly better than *MLE* is that, when ranking photos for a query identity, the background probability for this identity is the same for all photos. When we apply smoothing, therefore, the photos to be ranked are smoothed with the same background probability For fine-grained runs that suffer from missing data this will have the effect of appending any 'missing' photos to the end of the ranked list. For the coarser runs this smoothing will not alter the ranking as all photos already have a non-zero probability. This is different from the classification scenario, where we are ranking identities for a given image, and so the background probability of each identity to be ranked is different, which explains why smoothing always, with the exception of the *1 year* time window, gives a highly significant improvement in the classification scenario The fact that smoothing still improves the average performance, however, shows that even for these coarser runs, smoothing gives an improvement for some users. In the next section, we will see that by using the more powerful hierarchical smoothing techniques we can achieve significant improvements.

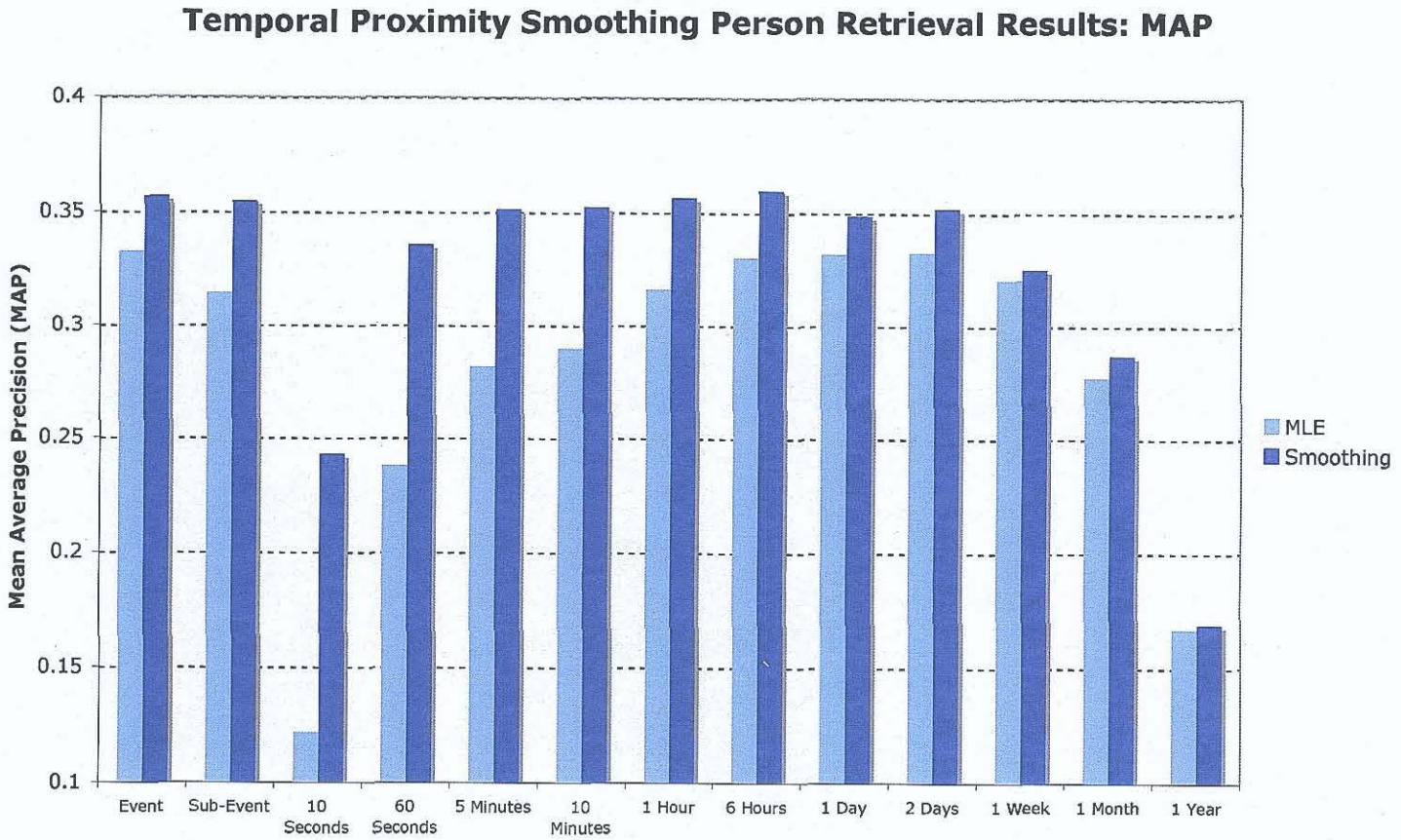


Figure 6.8: Context-based person retrieval results for Temporal Proximity using MLE and Smoothing.

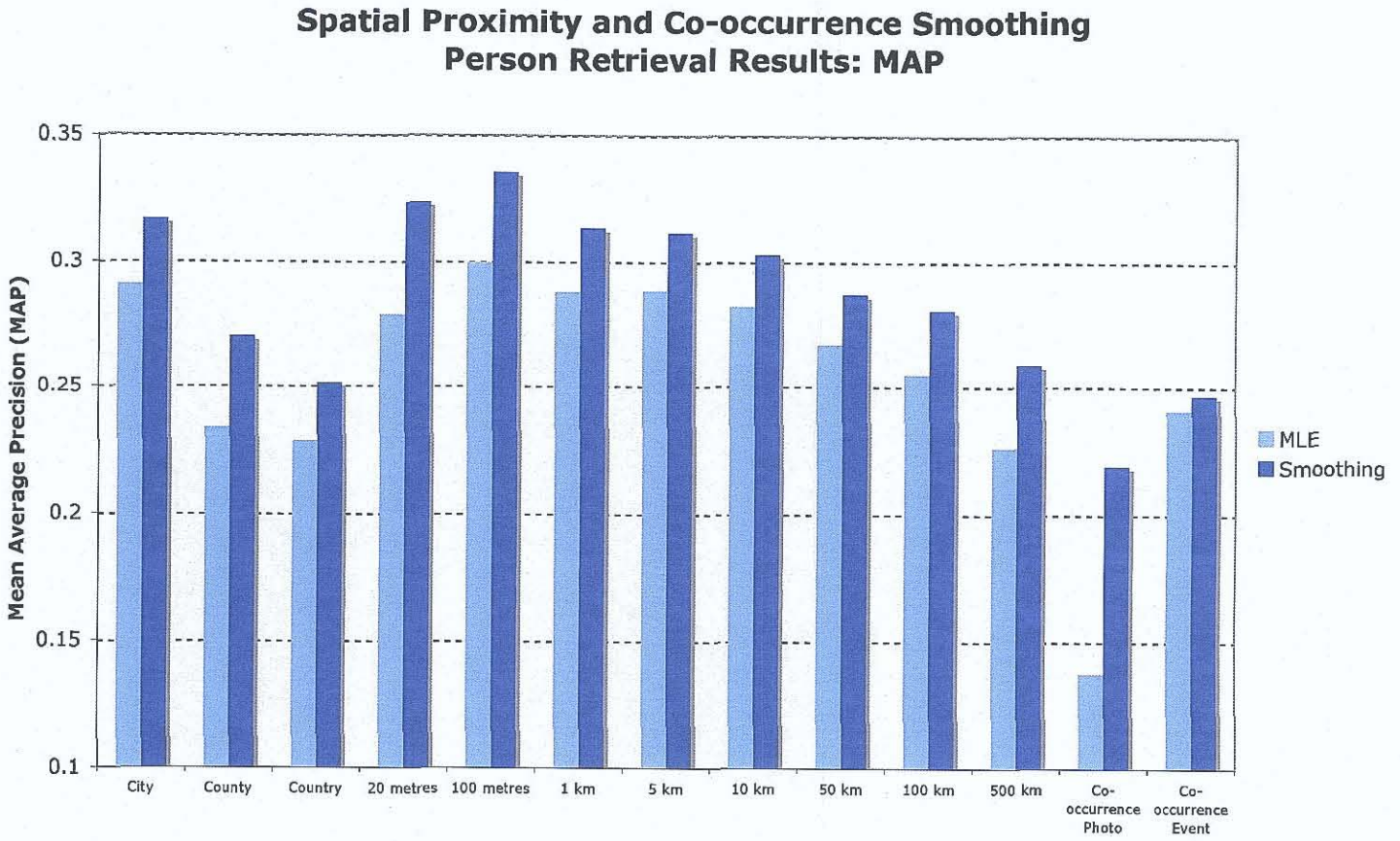


Figure 6.9: Context person retrieval results for Spatial Proximity and Co-occurrence using MLE and Smoothing.

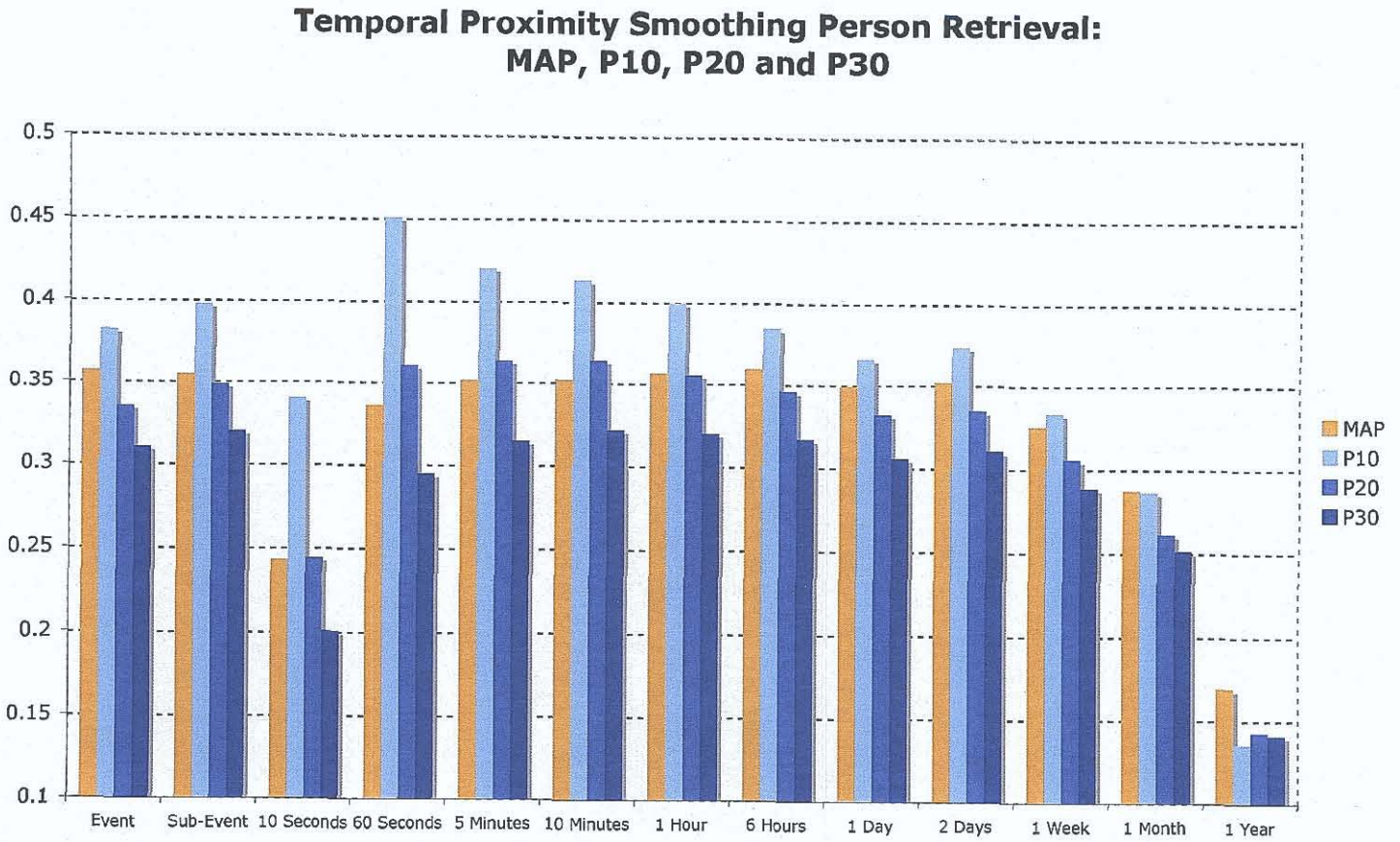


Figure 6.10: Context-based person retrieval results for Temporal Proximity Smoothing: MAP, P10, P20 and P30.

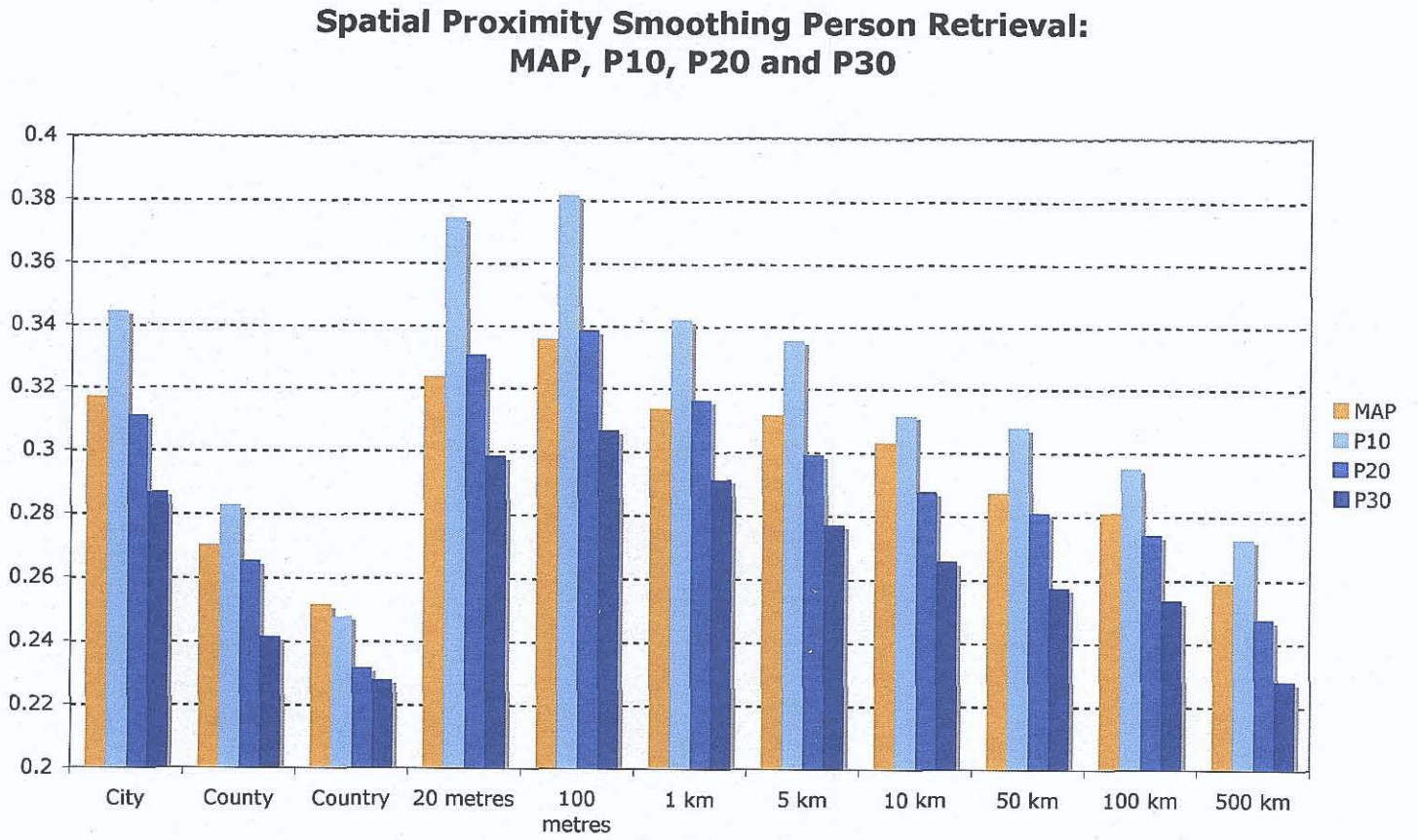


Figure 6.11: Context-based person retrieval results for Spatial Proximity Smoothing: MAP, P10, P20 and P30.

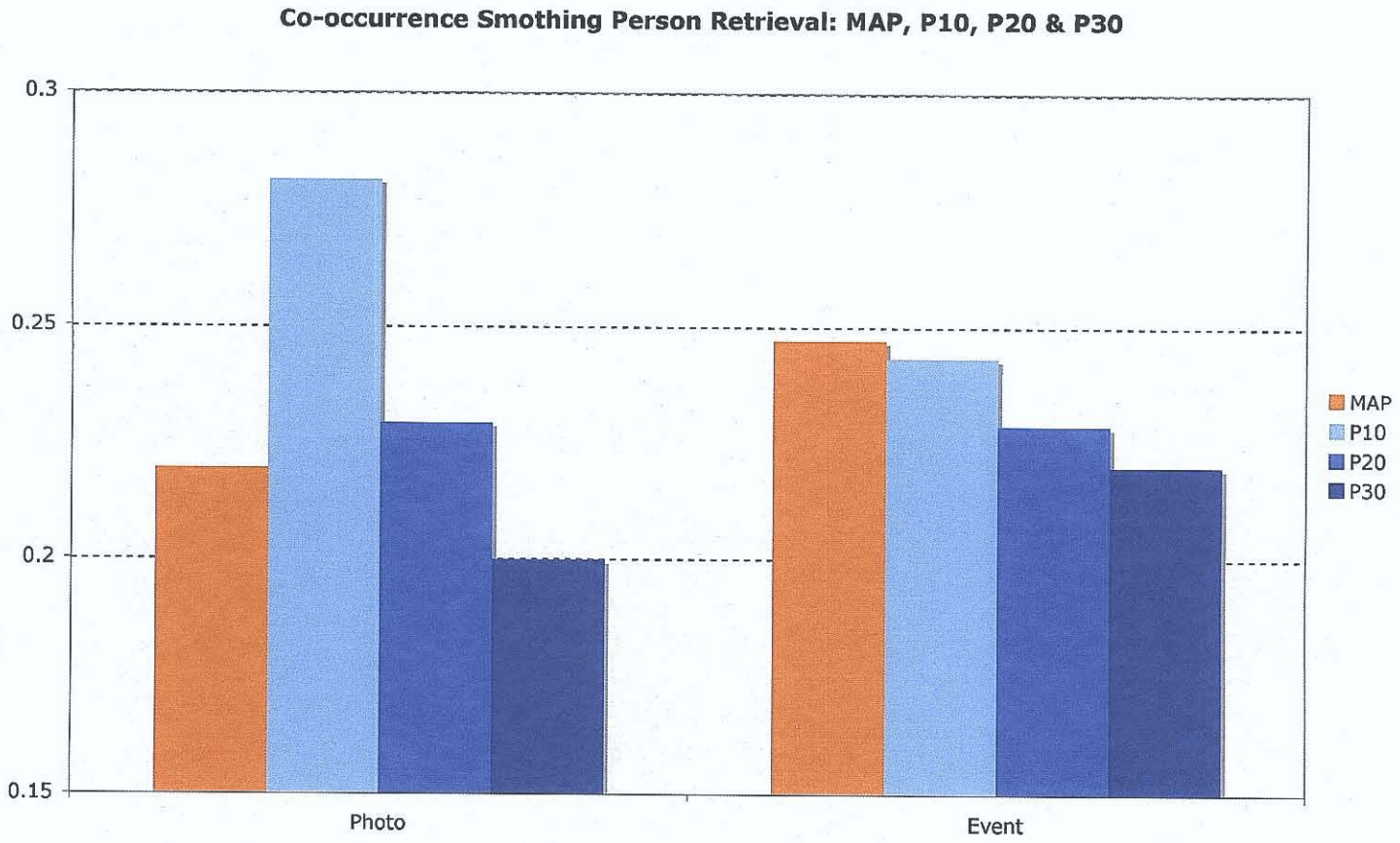


Figure 6.12: Context-based person retrieval results for Co-occurrence Smoothing: MAP, P10, P20 and P30.

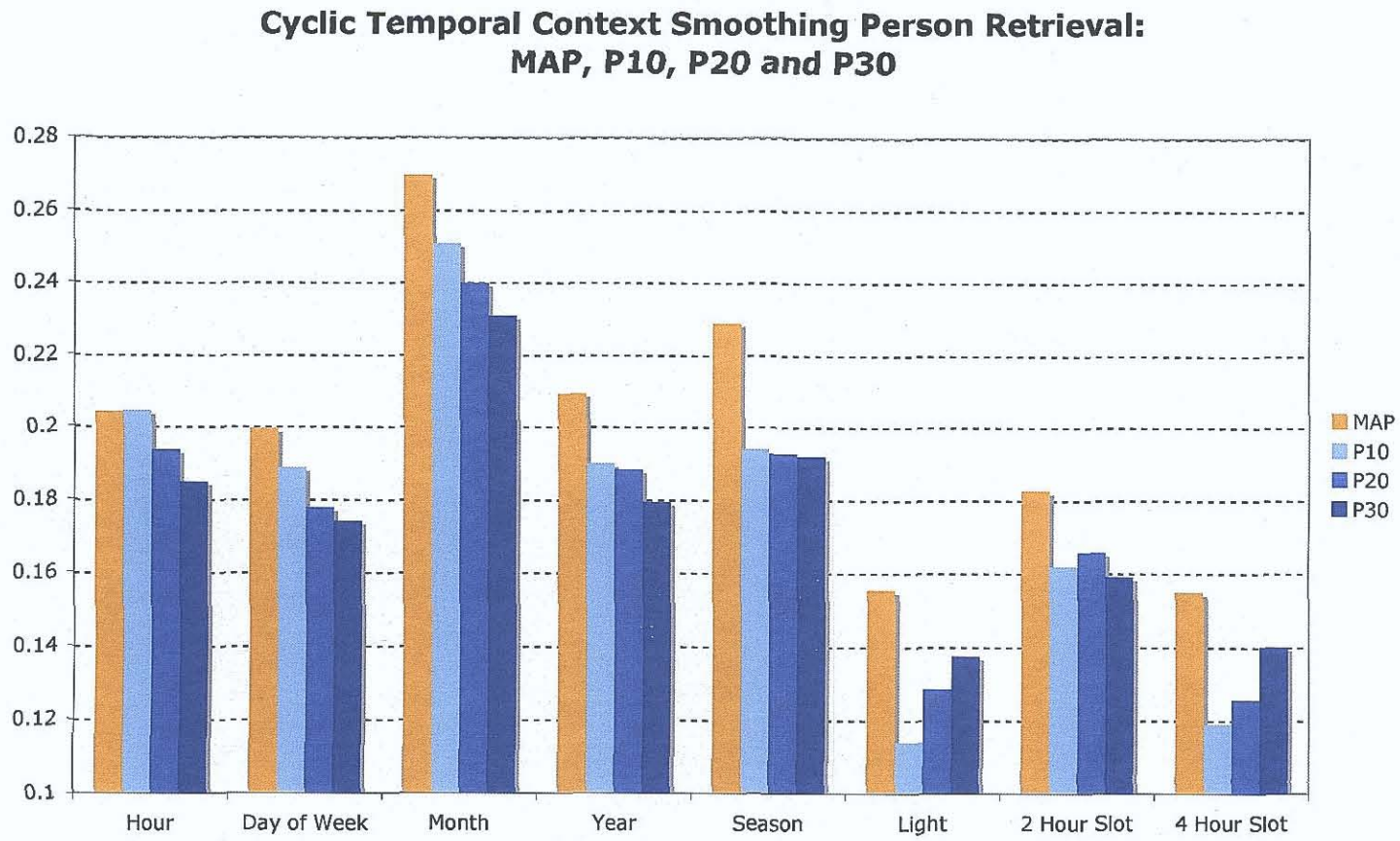


Figure 6.13: Context-based person retrieval results for Temporal Context Smoothing: MAP, P10, P20 and P30.

In Figure 6.10 we take a closer look at results from *temporal proximity smoothing*, comparing the results using each of our evaluation measures *MAP*, *P10*, *P20* and *P30*. We can see from this that the same runs do not perform consistently well for different evaluation measures. The best *MAP* comes from *6 hours*, with a score of 0.36, while *60 seconds* performs relatively poorly with a score of 0.336. *60 seconds*, however, has a *P10* of 0.451 compared to *P10* of 0.384 for *6 hours*, and this *P10* score is significantly better than all runs except *sub-event*. We believe the explanation for this is that the fine-grained *60 seconds* more accurately determines whether a person is likely to be present in a photo and so provides high quality results at the beginning of its ranked list, but missing data means that the performance drops off as we go down the ranked list, meaning that it performs well for *P10*, but suffers for other evaluation measures The coarse-grained *6 hours* is not quite as accurate as *60 seconds*, but suffers less from missing data and so performs better as we go down the ranked list.

The results for *spatial proximity smoothing*, shown in Figure 6.11, show that the fine-grained runs perform better for all evaluation measures. The best run is *100 metres*, which significantly outperforms all other runs except *1km* for *P10*. For *MAP* it is significantly better than *county*, *country*, *50 km* , *100 km* and *500 km*.

Figure 6.12 shows the results for *co-occurrence smoothing*. *Photo co-occurrence* performs much better than *event co-occurrence* for *P10*, a result which, although not statistically significant, deserves comment We believe that the reason that *photo co-occurrence* performs relatively well for this measure is that, for some users at least, *photo co-occurrence* is a better indicator of relationships between people than *event co-occurrence* and so it gives very good performance towards the top of its ranked list. *Photo co-occurrence*, however, as mentioned in Section 6.4.1, suffers more from missing data than *event co-occurrence* does, meaning than the ranked list from this system is likely to be truncated and so the other evaluation measures, which are influenced by performance further down the ranked list, suffer.

Figure 6.13 shows the results for *cyclic temporal context smoothing*, confirming

the results from the classification scenario that *month* is the summarise run. *Month* is significantly better than all other *cyclic temporal context* runs for *MAP*, *P20* and P30, and it performs significantly better than all runs except *hour* for *P10*. *Cyclic temporal context month* is outperformed by *temporal proximity month*, however, for all evaluation measures, although this result is only statistically significant for *P10*. This lack of statistical significance again suggests that *cyclic temporal context month* works well for some users but not for others

Hierarchical Smoothing and Nearest Neighbour Person Retrieval

Figure 6.14 shows the retrieval results for *temporal proximity 3-layer hierarchical smoothing*. We can see that the better-performing hierarchical systems outperform the best *smoothing* run. It is also noteworthy that the best-performing runs perform well for all evaluation measures, meaning that we do not need to choose between runs depending on what evaluation measure we deem most important, as we would need to do for *smoothing*. The best-performing run is *user→day→60secs* with a *MAP* of 0 443 and *P10* of 0.506. *User→day→60secs* performs significantly better than all other *temporal proximity 3-layer hierarchical smoothing* runs except *user→event→60secs* and *user→week→60secs* for *MAP* and *P10*. For *P20* and *P30* it is significantly better than most other runs *User→day→60secs* is also highly significantly better than the best-performing *temporal proximity smoothing* run for all four evaluation measures. All of the best-performing runs have *60secs* at the fine end of the hierarchy, confirming the good performance of *60secs* observed in the previous section.

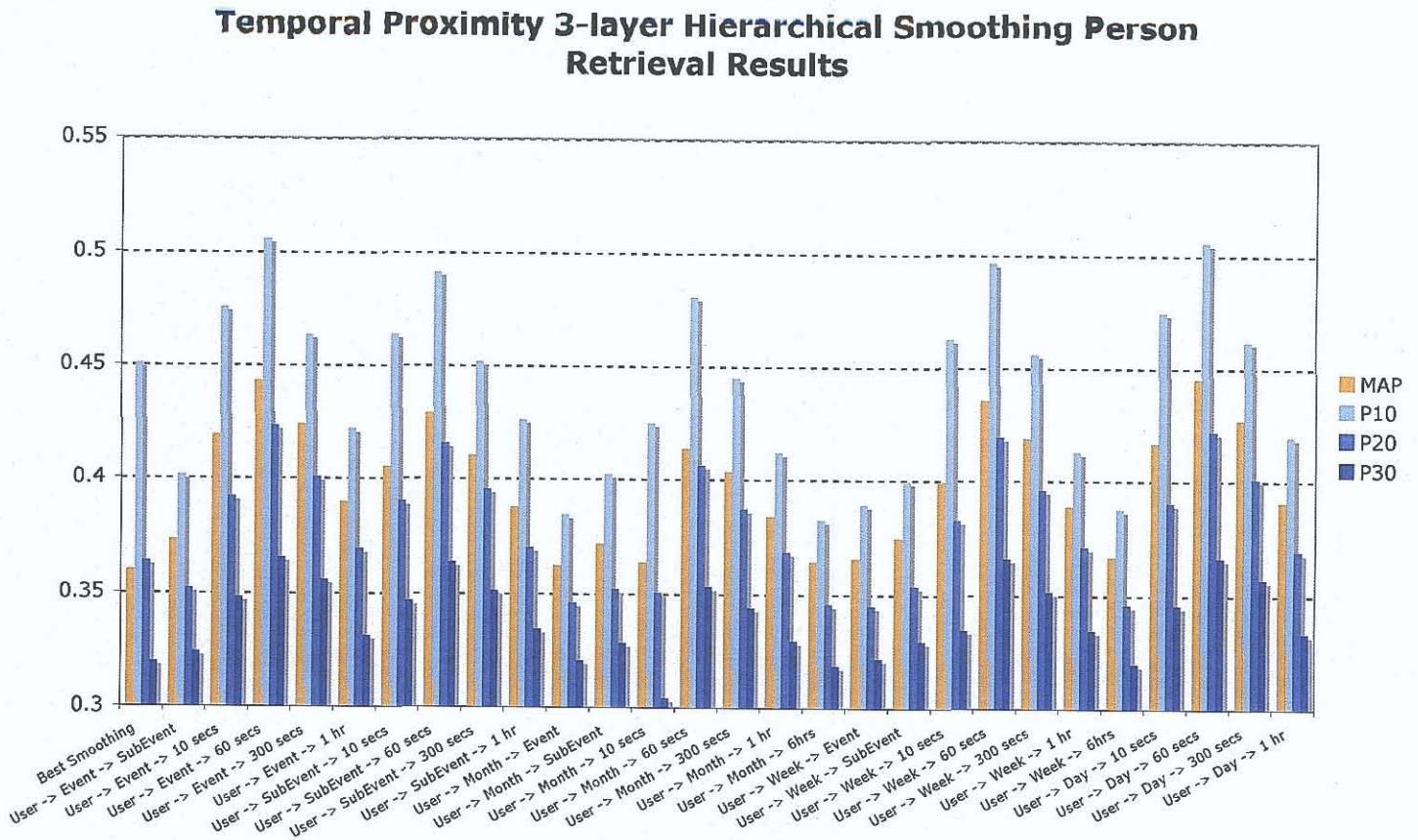


Figure 6.14: Context-based retrieval results for Temporal Proximity 3-layer Hierarchical Smoothing: MAP, P10, P20 and P30.

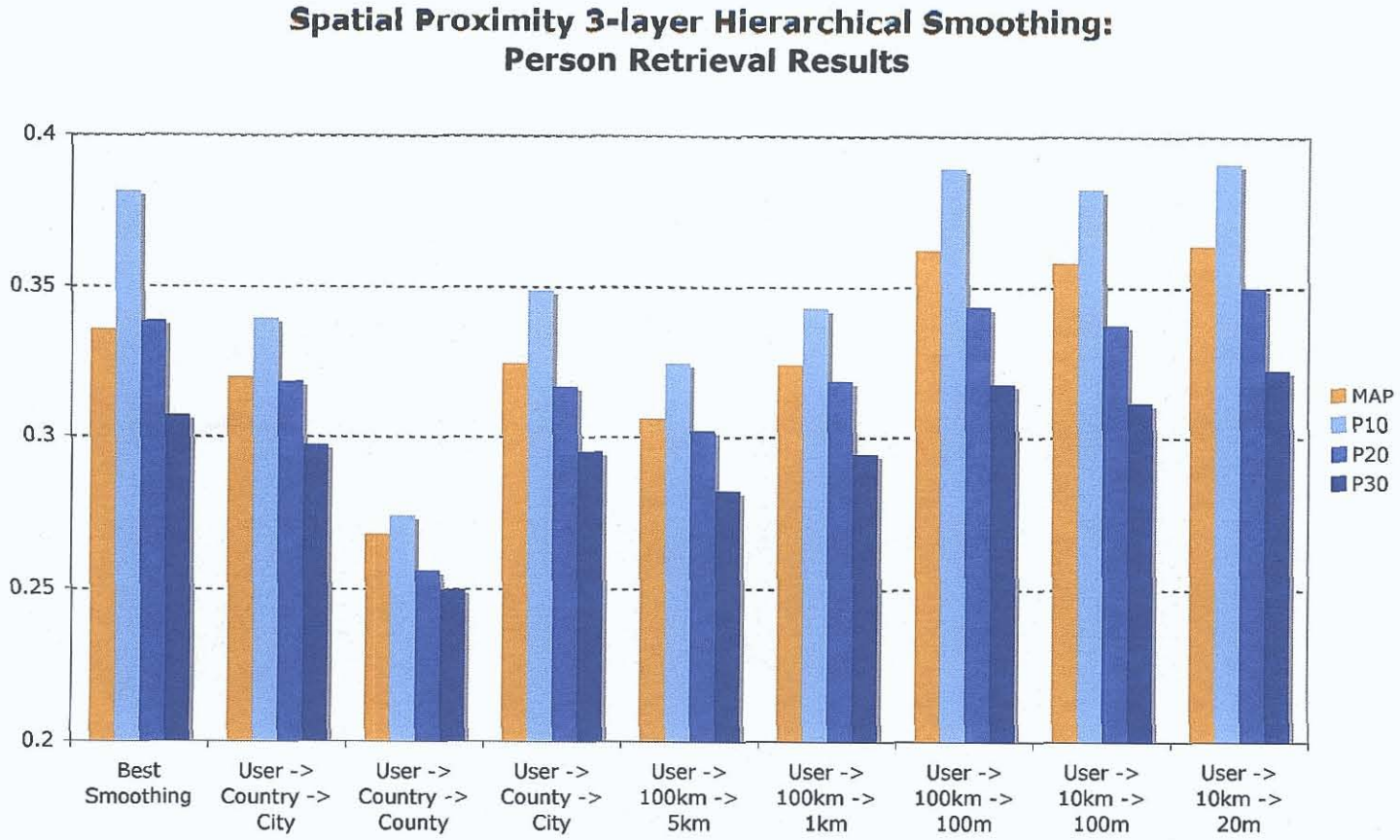


Figure 6.15: Context-based retrieval results for Spatial Proximity 3-layer Hierarchical Smoothing: MAP, P10, P20 and P30.

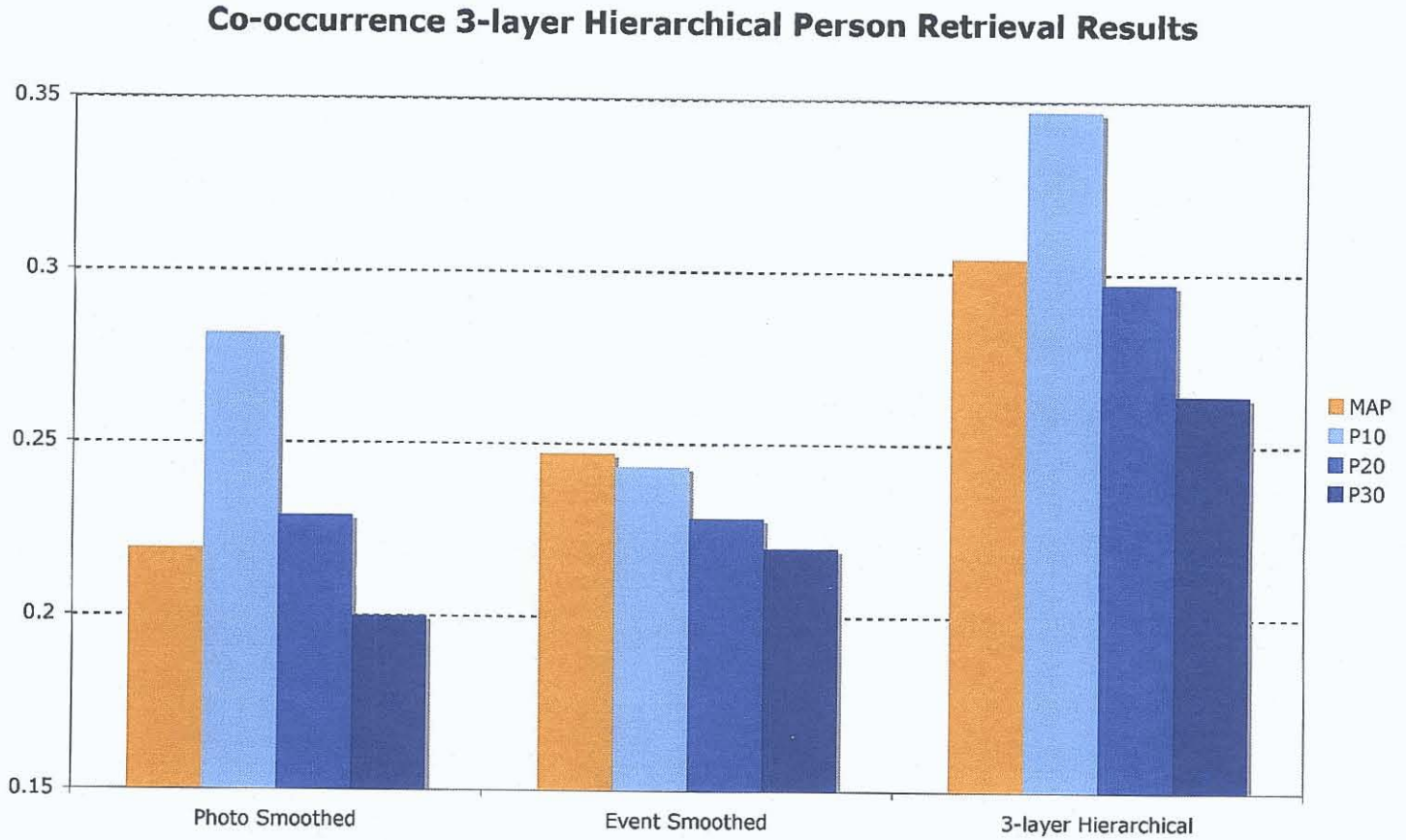


Figure 6.16: Context-based retrieval results for Co-occurrence 3-layer Hierarchical Smoothing: MAP, P10, P20 and P30.

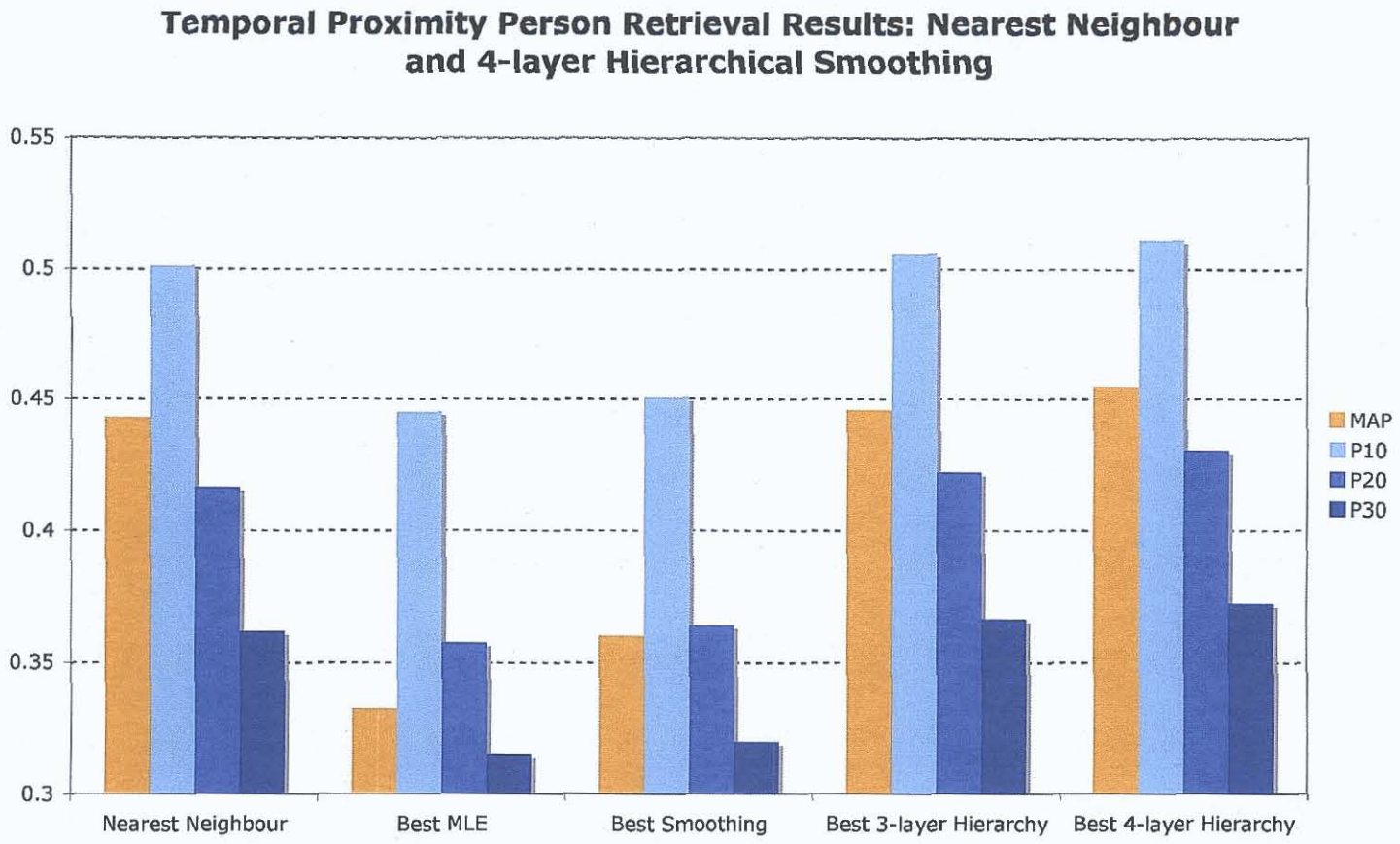


Figure 6.17: Context-based retrieval results for Temporal Proximity 4-layer Hierarchical Smoothing and Temporal Proximity Nearest Neighbour: MAP, P10, P20 and P30.

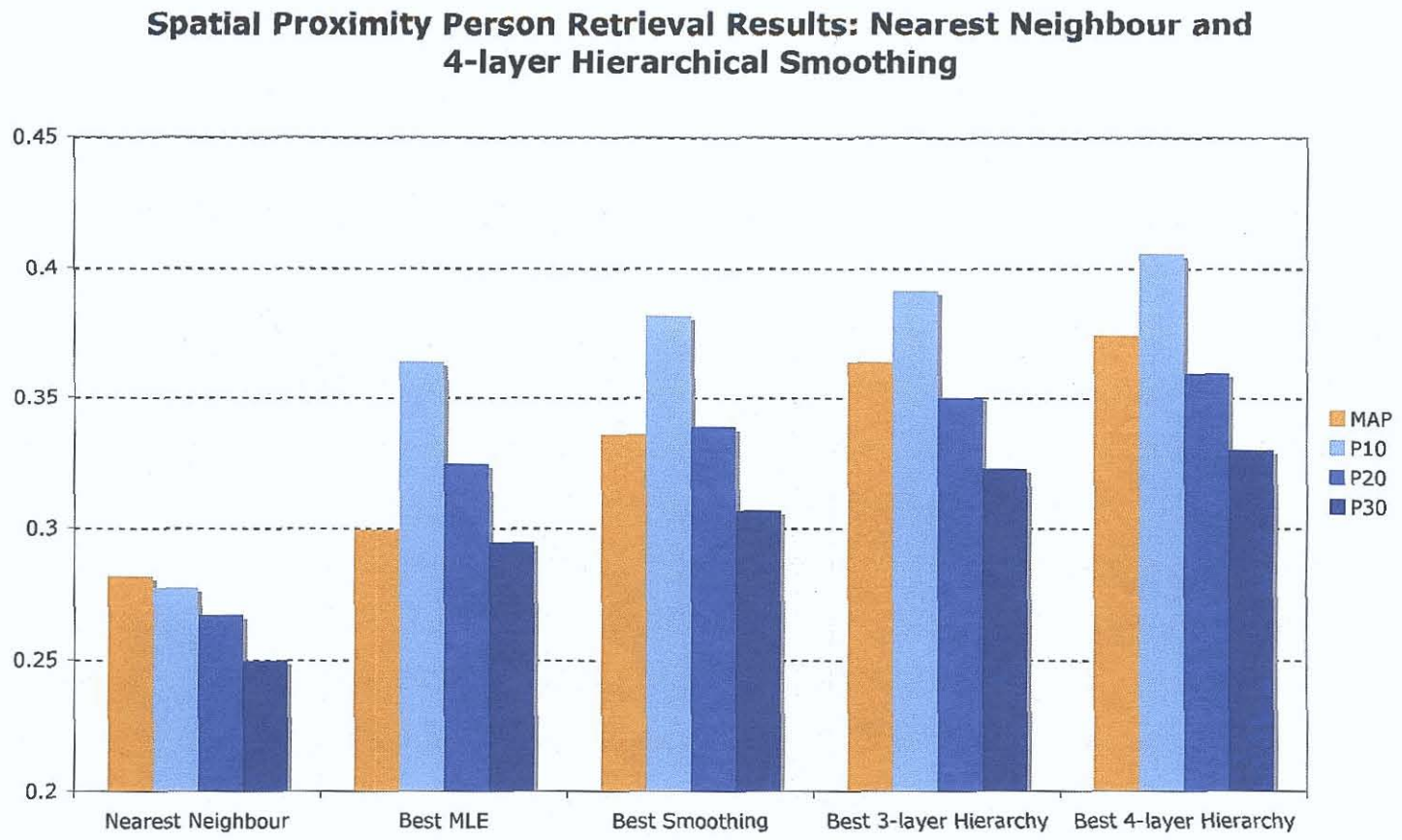


Figure 6.18: Context-based retrieval result for Spatial Proximity 4-layer Hierarchical Smoothing and Location Nearest Neighbour: MAP, P10, P20 and P30.

The results for *spatial proximity 3-layer hierarchical smoothing* show a similar improvement of hierarchical approaches over smoothing approaches (Figure 6.15). The best-performing *spatial proximity 3-layer hierarchical smoothing* run, *user→10km→20m*, is highly significantly better than the best *spatial proximity smoothing* run for *MAP* and significantly better for *P30*. The results for *P10* and *P20* also show an improvement over *spatial proximity smoothing*, although this is not statistically significant It also performs significantly better than all other *spatial proximity 3-layer hierarchical smoothing* runs except *user→100km→100m* and *user→10km→100m* for *MAP*, *P10* and *P20*.

3-layer hierarchical smoothing also gives a large improvement in performance for *co-occurrence* retrieval, as shown in Figure 6 16. *3-layer hierarchical smoothing* gives a large, and statistically significant, improvement over both *co-occurrence photo smoothing* and *co-occurrence event smoothing* for all four evaluation measures, improving *MAP* from 0.247 to 0.304, and improving *P10* from 0.281 to 0.347.

In Figure 6.17 we see the results for *temporal proximity 4-layer hierarchical smoothing*, alongside *nearest neighbour* retrieval and other *smoothing* approaches. *4-layer hierarchical smoothing* gives an improvement over *3-layer hierarchical smoothing* for all evaluation measures, and this difference is statistically significant for *MAP*, *P20* and *P30*. *Nearest neighbour* also performs very well, with a *MAP* of 0.442 and *P10* of 0.50. This compares with a *MAP* of 0.455 and *P10* of 0.511 for the best-performing *temporal proximity 4-layer hierarchical smoothing* approach. This relatively small difference, which is not statistically significant, in addition to the fact that it significantly outperforms *temporal proximity MLE* and *temporal proximity smoothing*, confirms the strong performance of *nearest neighbour temporal proximity* seen in the classification scenario

Figure 6 18 shows the results for *spatial proximity 4-layer hierarchical smoothing* retrieval. The best-performing *spatial proximity 4-layer hierarchical smoothing* run, *user→ 100km→1km→20m*, gives a significant improvement over all other approaches for all evaluation measures *Nearest neighbour* performs poorly for *spatial proximity*,

being significantly outperformed by all other approaches for *P10*, *P20* and *P30*, and significantly outperformed by all approches except *spatial proximity MLE* for *MAP*.

Comparing the results given by different modalities, *temporal proximity* retrieval, with a best *MAP* of 0.455 and *P10* of 0 511, performs highly significantly better than *temporal context*, *spatial proximity* and *co-occurrence*. We believe the retrieval results in this section are encouraging, and that the *P10* score of *0.511*, which means on average 5 out of the first 10 ranked photos are relevant, should be good enough to encourage real users to use the annotation tools provided by our system, as described in Chapter 4. We will report the results of a real user study of the system in Chapter 7. In the next section we examine content-based approaches to person classification and retrieval.

6.5 Evaluation of Content-based Approaches to Person Classification and Retrieval

6.5.1 Content-based Person Classification

Figure 6.19 shows the *5-hit rate* results for content-based person classification In this section we restrict our evaluation to the ranking of the top 20 most popular people in each user's collection, and we evaluate content-based rankings of candidate identities filtered by user and by event The event-filtered systems are all context-aware since they make use of context-based event information, and this constitutes a simple approach to combining context-based and content-based techniques We use this simple event-filtering approach here to test the usefulness of content-based techniques in an environment which is highly constrained by context, where lighting and camera conditions are the same for all photos, and where we can assume that the people in the photos will be wearing the same clothes for the duration of an event. We will examine techniques for combining content and context in more detail in Section 6.6.

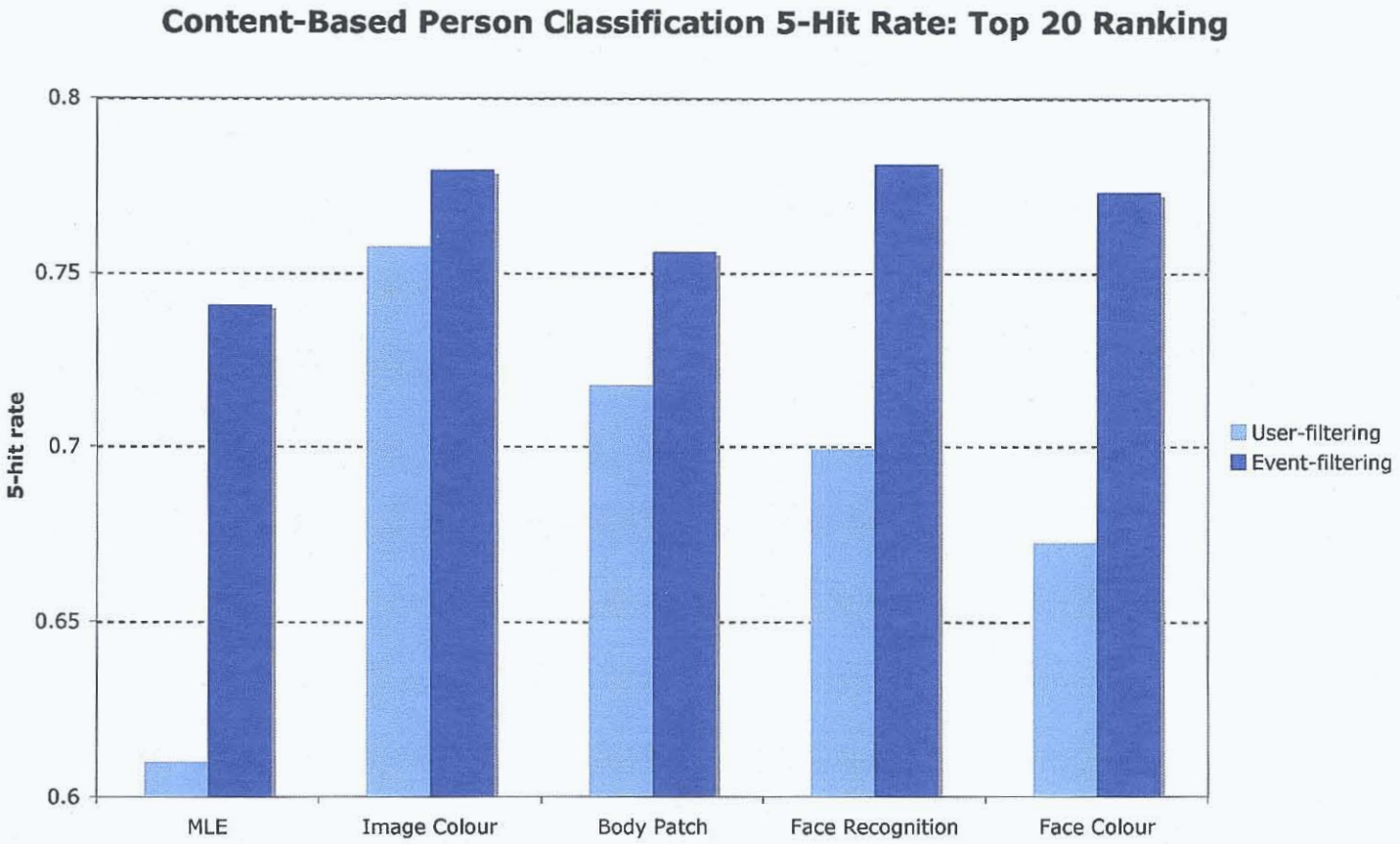


Figure 6.19: Content-based person classification results: 5-hit rate.

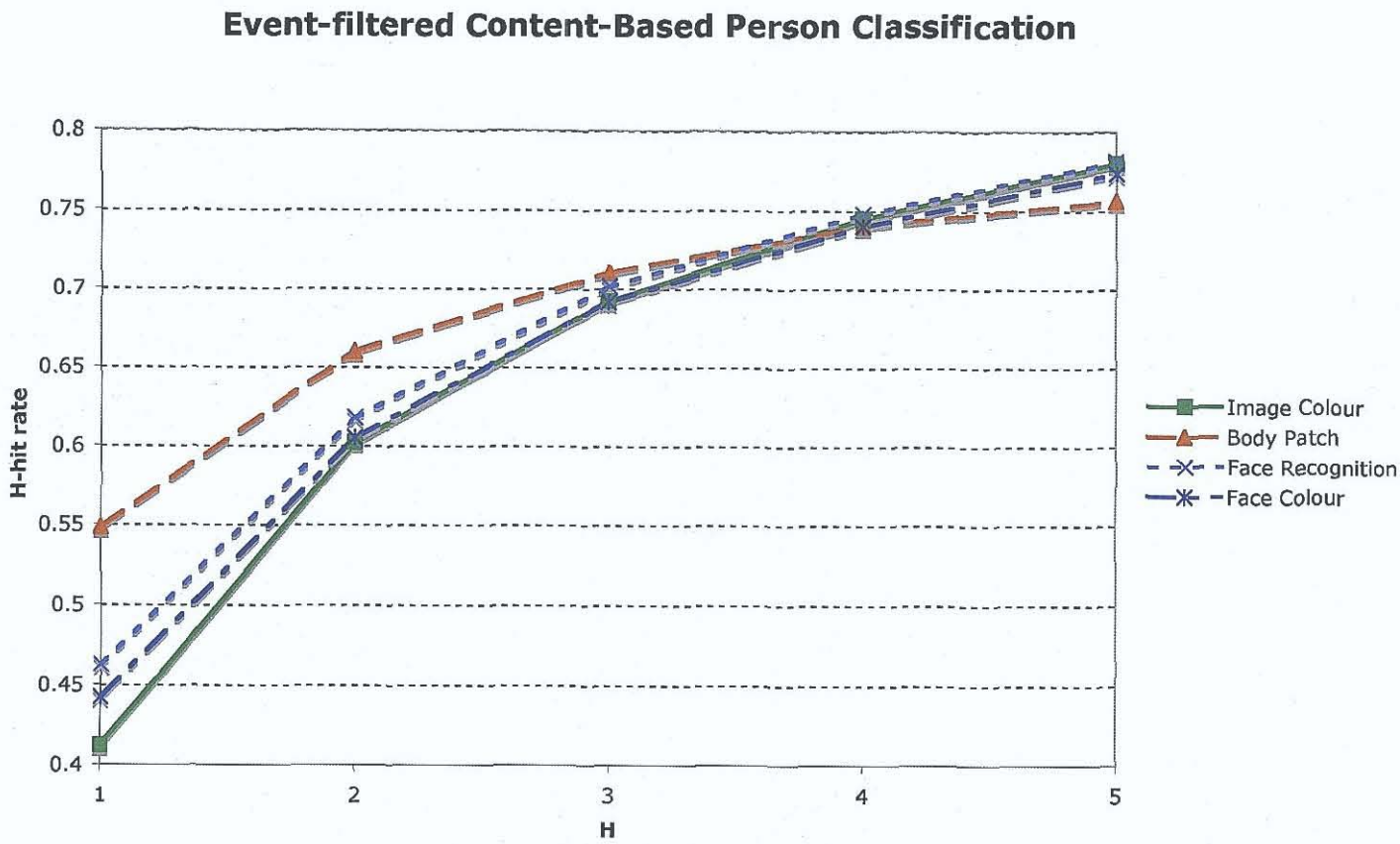


Figure 6.20: Content-based person classification results: H-hit rate for various values of H.

For user-filtering, the *image colour* feature significantly outperforms all of the region level features, with a *5-hit rate* of 0.758 compared with a best region-based approach of 0.718, for user-filtered *body patch*. This result is somewhat surprising since it suggests that, within personal photo collections, global image similarity is a better indicator of person re-occurrence than the similarity of region-based features such as *face recognition* and *body patch*, which are specifically designed to model people. We believe the reason for this is that user-filtered *global image colour* classification, which ranks candidate identities based on the similarity of photos containing known people to the photo containing the face region to be classified, will often rank photos from the same event highly because similar lighting conditions and similar locations mean that photos within the same event should be most similar to the query image. This means that this approach can perform well without event-filtering. For region-level approaches, which rank candidate identities based on the similarity of known regions to the region to be classified, it is more likely that a region from the query photo could be highly similar to a region from a photo outside of the current event, and by going outside the current event the possibility of errors in classification is increased.

Event-filtering always improves performance over user-filtering, a difference that is statistically significant for *face recognition* and *face colour*. Event-filtered *face recognition*, with a *5-hit rate* of 0.781, marginally outperforms event-filtered *image colour*, with a *5-hit rate* of 0.78, although this difference is not statistically significant. Also there is no significant difference for *5-hit rate* between event-filtered *face colour* and event-filtered *image colour*, or between event-filtered *body patch* and event-filtered *image colour*. For both user-filtering and event-filtering, all content-based approaches perform significantly better than context-based *MLE*, which we use here as a baseline. This shows that content-based approaches are capable of improving on the ranking of a filtered set of identities given by context-based *MLE* approaches.

The good performance of *face colour* is slightly surprising since it is based solely

on the colour of face regions in images Event-filtered *face colour* has a *5-hit rate* of 0 773 compared with 0 756 for event-filtered *body patch* and 0.781 for event-filtered *face recognition*, with neither approach significantly better than event-filtered *face colour*. *Face colour* significantly outperforms context-based *MLE* for both user-filtering and event-filtering. This suggests that, within the constrained environment of personal photo collections, and particularly when filtering by event, *face colour* can be a useful feature for identity classification, and that within this constrained environment it is able to model skin tone and other colour-based variations between identities (e g. colour of hair occluding face, sunglasses) We also note that, in the MediAssist collection, the majority of faces are of Irish people, and there is a general lack of racial variety in the faces present. We would expect that, in a more heterogeneous collection displaying more variety in the colour of faces present, then face colour might be even more useful.

Figure 6.20 shows the *H-hit rate* for the event-filtered content-based approaches for various values for h Event-filtered *body patch* is highly significantly better than all other event-filtered approaches for *1-hit rate*, with a score of 0.548, and significantly better than all other approaches for *2-hit rate*, with a score of 0.66 This shows that the simple, colour-based, *body patch* feature is more powerful than other region-based features for identity classification in highly constrained environments The fact that not all faces have a corresponding body patch region (the reasons for this are described in Chapter 5), in addition to the occasional occlusion of the torso, the occasional overlapping of the torso with the torsos of adjacent people in the same photo, and the fact that sometimes multiple people wear very similar clothes at the same event, are factors which inhibit the performance of *body patch* for larger values of h. Although *body patch* is the most discriminative content-based feature for identity classification, as can be seen from the *1-hit rate* and *2-hit rate* event-filtered results, for cases where this approaches fails to rank the correct identity in the top 2 or 3 identity suggestions, it is more likely to fail completely, which is the reason that the *5-hit rate* suffers. Other approaches do not have this problem, so

although they are less likely to rank the correct identity in the top 1 and 2 identity suggestions, they are more robust and will rank the correct identity in the top 5 identity suggestions more often.

The other region-based approaches, namely event-filtered *face recognition* and event-filtered *face colour*, also perform well for low values of h, with both features significantly outperforming event-filtered *image colour* for *1-hit rate*. For *2-hit rate*, event-filtered *face recognition* significantly outperforms both event-filtered *image colour* and event-filtered *face colour*. For values of h greater than 2 there tends to be no significant difference between the event-filtered content-based approaches, suggesting that event-filtering obfuscates the difference between approaches if evaluating using *H-hit rate* with larger values of h.

6.5.2 Content-based Person Retrieval

Figure 6.21 shows the content-based person retrieval results for the photo-level retrieval scenario, where the task is to retrieve a photo containing the person in the query. For user-filtering the best-performing content-based feature is *image colour*. This is consistent with the results from the classification scenario. For *MAP* and *P10*, user-filtered *image colour* performs highly significantly better than all other user-filtered features, and for *P20* and *P30* user-filtered *image colour* performs significantly better than all other user-filtered features

Event-filtering always improves performance over user-filtering, which is also consistent with the results from the classification scenario In fact the improvement is stronger for the retrieval scenario, with all content-based features gaining a highly significant improvement with the use of event-filtering. Event-filtering is again of most benefit for the region-level features, particularly *body patch*. Event-filtered *body patch* performs significantly better than all other event-filtered content-based features for all evaluation measures with a *MAP* of 0.493 and a *P10* of 0.613. This compares with a best *MAP* of 0.455 and a best *P10* of 0.511 for context-based approaches, an improvement that is statistically significant for all evaluation measures.

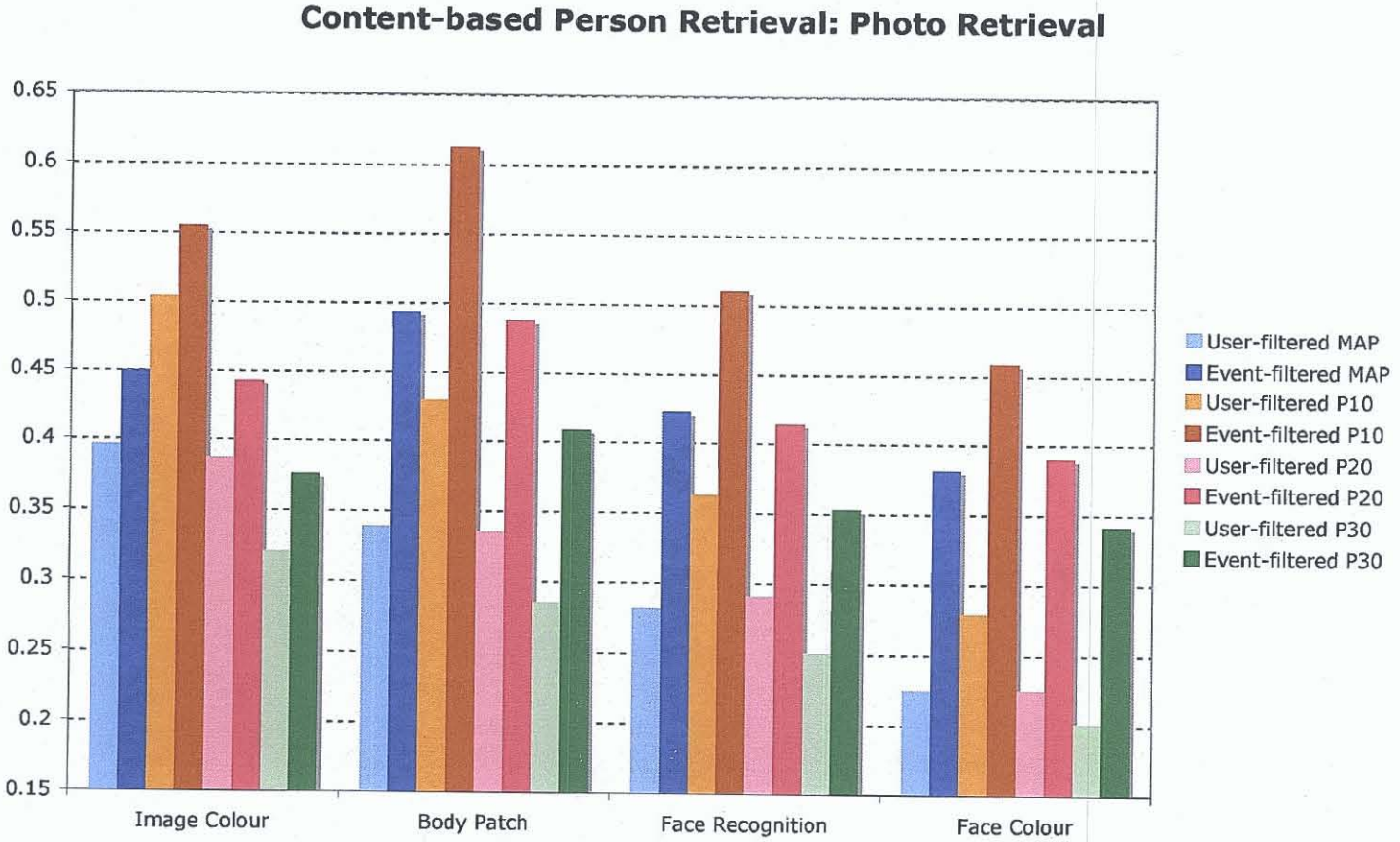


Figure 6.21: Content-based person retrieval results: photo-level retrieval.

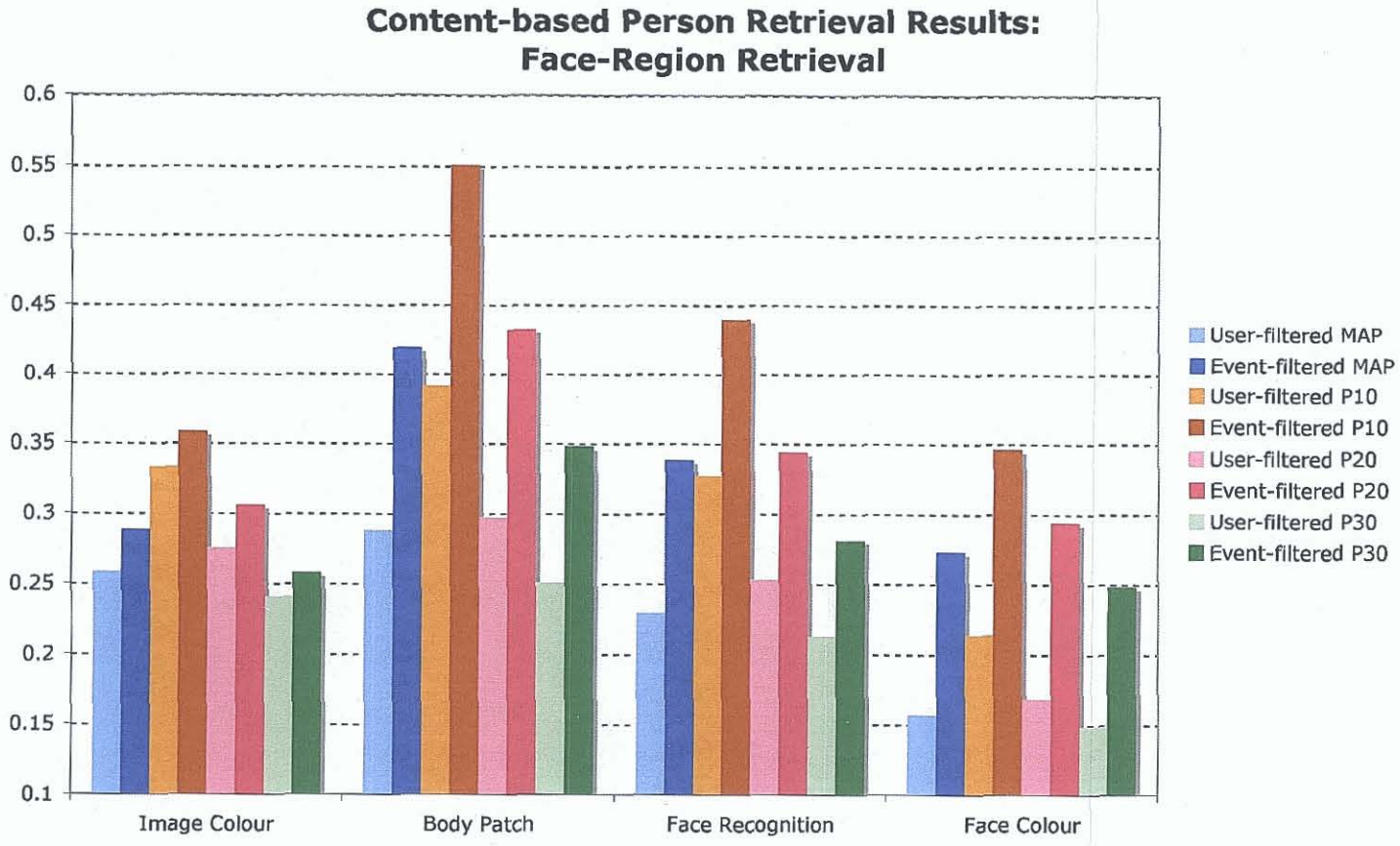


Figure 6.22: Content-based person retrieval results: face region-level retrieval.

The fact that missing body patch regions, occlusion and overlapping do not affect the performance of event-filtered *body patch* in the retrieval scenario in the same way as in the classification scenario shows that the high precision of the results for cases where the *body patch* works compensates for the cases where it fails, and the cases where it fails do not adversely affect its overall retrieval performance. The next best-performing feature is event-filtered *image colour*, which is significantly better than event-filtered *face colour* for all evaluation measures, and significantly better than event-filtered *face recognition* for *P10*, *P20* and *P30*.

The event-filtered *body patch* results for the photo-level retrieval scenario are very encouraging but for a real system to benefit from this feature the user must annotate identities at the face region level, and not just at the photo level, as described in Chapter 6. For this reason, we also evaluate the face region retrieval scenario, described in Section 6.2.2, where the task is to retrieve the face-region within the photo corresponding to the query identity. If the correct region is returned, then the user will be able to annotate candidate regions correctly retrieved by the system.

The results for the face region retrieval scenario are shown in Figure 6.22. The best-performing feature is *body patch*, with an event-filtered *MAP* of 0.419, and *P10* of 0.551. The result for event-filtered *body patch* is highly significantly better than all other content-based approaches for all evaluation measures. Event-filtered *image colour* suffers in the face region retrieval scenario, and is significantly outperformed by event-filtered *face recognition* and event-filtered *body patch* for all evaluation measures. This is not surprising since the *image colour* feature does not distinguish between regions, assigning the same similarity score to all regions within the same image. Again, event-filtered *face colour* is significantly outperformed by the other region-based features, showing that this is the least effective region-level feature for retrieval.

The results for the event-filtered *image colour* feature for the photo-level scenario, with *MAP* of 0.449 and *P10* of 0.555, are better than those for the best-performing approach in the face region retrieval scenario (event-filtered *body patch*), with *MAP*

of 0.419 and *P10* of 0.551. This difference is small for *P10*, however, and is only statistically significant for *MAP*. This means that, for the photo-level annotation scenario, the system will return more accurate photo-level retrieval results, using photo-level features, than the region-level results from the best-performing region-based approach for the face region annotation scenario. This suggests that the benefits of region-based approaches may not help user annotation, because the user will need to retrieve the correct region in order to annotate to the region-level, and face region-level retrieval using region-based features is not as accurate as photo-level retrieval using global image features. We will examine the differences between photo retrieval and face region-level retrieval again in Section 6.6.2, which evaluates combined approaches to person retrieval.

6.6 Evaluation of Combined Approaches to Person Classification and Retrieval

In this section we evaluate combined approaches to person classification and retrieval. Firstly we look at combined context-based approaches, followed by combined content-based approaches. Finally, we examine approaches which combine context-based approaches with content-based approaches. All combined approaches use the *CombSum* fusion approach, and all weighted combined approaches used the *Weighted CombSum*, both of which are described in Chapter 5.

6.6.1 Combined Person Classification

We restrict our evaluation of combined approaches to person classification to the ranking of the top 20 most popular people in each user's collection. We have shown in Section 6.4.1 that the results for top 20 ranking are consistent with the results when ranking all people. As with the learning of weights for context-based smoothing approaches to person classification, the weights for weighted combination approaches

are learned on the test set by optimising the evaluation measure *5-hit rate*.

Combined Context-based Person Classification

In Figure 6.23 we present the results for combined context-based approaches to person classification. For the *combined context hierarchical smoothing all* approach, which combines the best-performing *temporal proximity 4-layer hierarchical smoothing* approach, the best-performing *spatial proximity 4-layer hierarchical smoothing* approach and *co-occurrence 3-layer hierarchical smoothing*, we can see that unweighted *combined context hierarchical smoothing all* performs worse than *temporal proximity 4-layer hierarchical smoothing* This difference is highly statistically significant *Combined context hierarchical smoothing all (weighted)*, however, achieves a statistically significant improvement over *temporal proximity 4-layer hierarchical smoothing*, increasing the 5-hit rate from 0.872 to 0.879.

Since location-aware photo capture devices are not currently in common use, we also evaluate a *combined context hierarchical smoothing no location* approach, which does not use location information, but combines *temporal proximity* and *co-occurrence*. Unweighted *combined context hierarchical smoothing no location* again performs worse than *temporal proximtity 4-layer hierarchical smoothing*, a difference which is highly statistically significant. *Combined context hierarchical smoothing no location (weighted)* has a 5-hit rate of 0.876, which is a small but statistically significant improvement over *temporal proximity 4-layer hierarchical smoothing*.

Combined context nearest neighbour, with a 5-hit rate of 0.868, achieves a very slight improvement over *temporal proximity nearest neighbour*, which has a 5-hit rate of 0 866, although this difference is not statistically significant. Weighting improves the 5-hit rate to 0.874, which is only slightly below the performance of *combined context hierarchical smoothing all (weighted)*, although the difference is statistically significant.

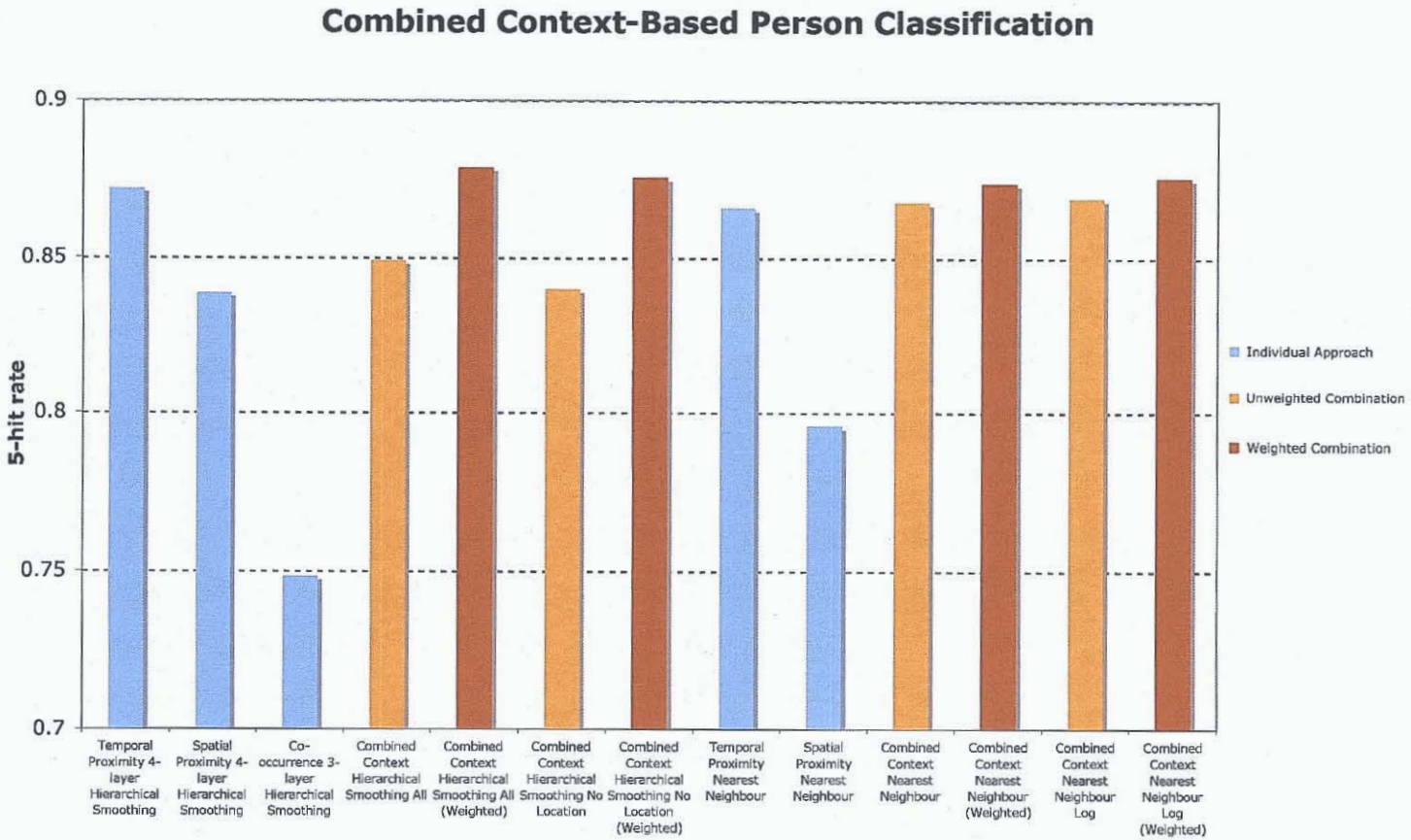


Figure 6.23: Combined context-based person classification results.

As described in Chapter 5, the distribution of values for *Nearest Neighbour* approaches is skewed, with, for example, some photos separated by seconds and others separated by months or years. To counteract the problems caused by this we can take the log of the *temporal proximity nearest neighbour* and *spatial proximity nearest neighbour* distances before normalising and combining them with other approaches. In Figure 6.23 we also give the results for *combined context nearest neighbour log*, which gives a slight improvement over *combined context nearest neighbour*, improving *5-hit rate* from 0.868 to 0.869 for unweighted combination, and from 0.874 to 0.876 for weighted combination. Although these improvements are small, they are statistically significant, and we will see in Section 6.6.2 that the improvement is larger for the retrieval scenario.

We can see from the results in Figure 6.23 that combining *temporal proximity* information with other context-based approaches gives only a small improvement in performance. Without weighting, in fact, combining *temporal proximity 4-layer hierarchical smoothing* with *spatial proximity 4-layer hierarchical smoothing* and *co-occurrence 3-layer hierarchical smoothing* approaches actually harms performance. This reaffirms the fact that *temporal proximity* is the best context-based feature for identity classification, and suggests that the other context-based features tend to be somewhat redundant in the presence of *temporal proximity* information when we have a richly annotated photo collection.

Combined Content-based Person Classification

The results for combining content-based approaches to person classification are shown in Figure 6.24. For user-filtering, shown in Figure 6.24(a), combining content-based features gives a highly significant improvement over all individual approaches for both weighted and unweighted combination. For *1-hit rate*, *2-hit rate* and *3-hit rate* there is no significant difference between *combined content user-filtered* and *combined content user-filtered (weighted)*. For *4-hit rate* and *5-hit rate*, *combined content user-filtered (weighted)* performs significantly better than *combined content*.

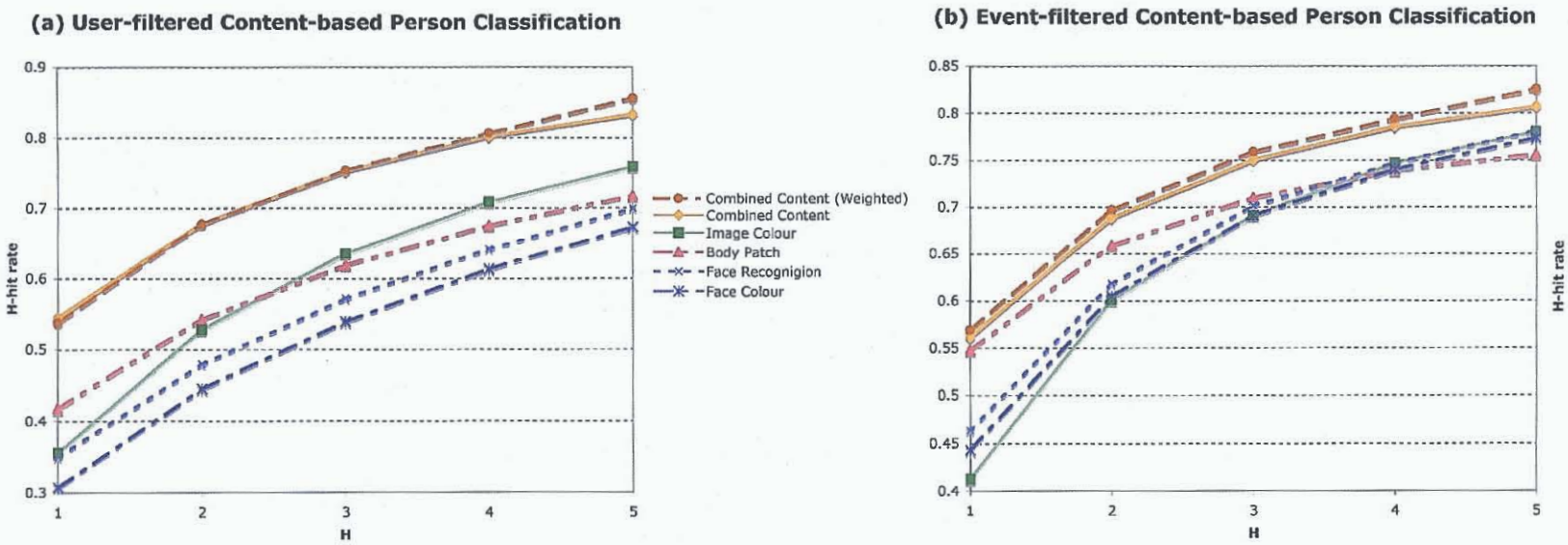


Figure 6.24: Combined content-based person classification results.

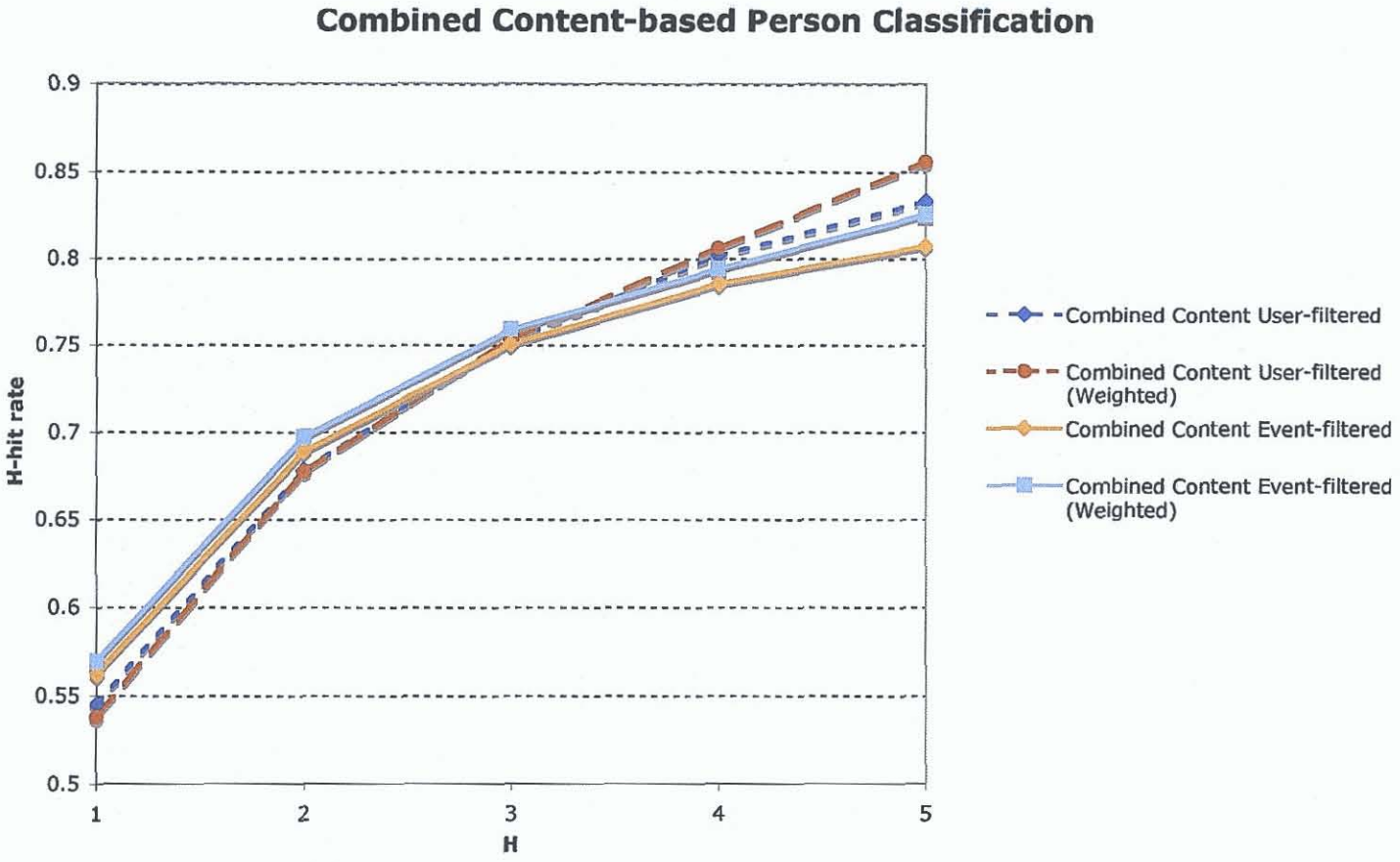


Figure 6.25: Combined content-based person classification: User-filtered and Event-filtered.

Combined approaches give a similar improvement for event-filtering, shown in Figure 6.24(b), with *combined content event-filtered* and *combined content event-filtered (weighted)* performing highly significantly better than all individual content-based approaches for *3-hit rate*, *4-hit rate* and *5-hit rate*. For *1-hit rate* and *2-hit rate*, both event-filtered combined approaches perform highly significantly better than *image colour event-filtered*, *face recognition event-filtered* and *face colour event-filtered*, although neither perform significantly better than *body patch event-filtered*. Again, there is no significant difference between *combined content event-filtered* and *combined content event-filtered (weighted)* for *1-hit rate*, *2-hit rate* and *3-hit rate*, although *combined content event-filtered (weighted)* performs significantly better for *4-hit rate* and *5-hit rate* Since the weights are optimised using *5-hit rate* it is obvious that weighting will help for this measure, and we would expect *4-hit rate* performance to be similar to that of *5-hit rate*. The fact that weighting does not help for other evaluation measures suggests that there is little benefit to be gained from weighting when combining content-based approaches to person classification

In Figure 6.25 we compare *combined content event-filtered* with *combined content user-filtered*. For *1-hit rate*, we can see that *event-filtered* approaches outperform *user-filtered* approaches, although the difference is not statistically significant For *2-hit rate*, *3-hit rate* and *4-hit rate* there is no significant difference between *user-filtering* and *event-filtering*. For *5-hit rate*, *user-filtering* outperforms *event-filtering*, although this difference is only significant between *combined content user-filtered (weighted)* and the unweighted *combined content event-filtered*; it is possible that this difference is caused by the weighting rather than the filtering used. These results suggest that the more accurate *event-filtering* performs better for *1-hit rate* but, for some users at least, *user-filtering* benefits from being able to rank identities from outside the current event when using the coarser evaluation measure *5-hit rate*

Combined Context-based and Content-based Person Classification

Figure 6.26 shows the *5-hit rate* person classification results for combining the best *combined context hierarchical smoothing* techniques with *combined content* approaches. Comparing context-based with content-based results, the context-based *combined context hierarchical smoothing* approach significantly outperforms *combined content event-filtered* for this evaluation measure It also outperforms *combined content user-filtered*, although this difference not statistically significant

Because the *hierarchical smoothing* approaches rely on weighted smoothing, and because *combined context hierarchical smoothing all* only improves performance over the best individual context-based hierarchical smoothing approach when using weights, we only evaluate weighted combinations of *combined context hierarchical smoothing* and content-based approaches. Weighted combination of content and context always gives a large, and highly significant improvement in performance over the combined content-based and combined context-based approaches alone. *Combined context hierarchical smoothing all / combined event-filtered content* achieves a *5-hit rate* of 0 931, a highly significant improvement over *combined context hierarchical smoothing*, which has a *5-hit rate* of 0.879. *Combined context hierarchical smoothing no location / combined event-filtered content* has a *5-hit rate* of 0.93, again showing that we do not lose much in terms of performance without location information.

Combined approaches which use *combined content event-filtered* always perform highly significantly better than those that use *combined content user-filtered.* This is in spite of the fact that, as shown in Section 6 6.1, there is no clear difference between *event-filtering* and *user-filtering* for combined content-based classification This suggests that the better *1-hit rate* of event-filtering, although not statistically significant, is indicative of superior performance, and that this superior performance is emphasised when we combine the content-based results with the context-based results

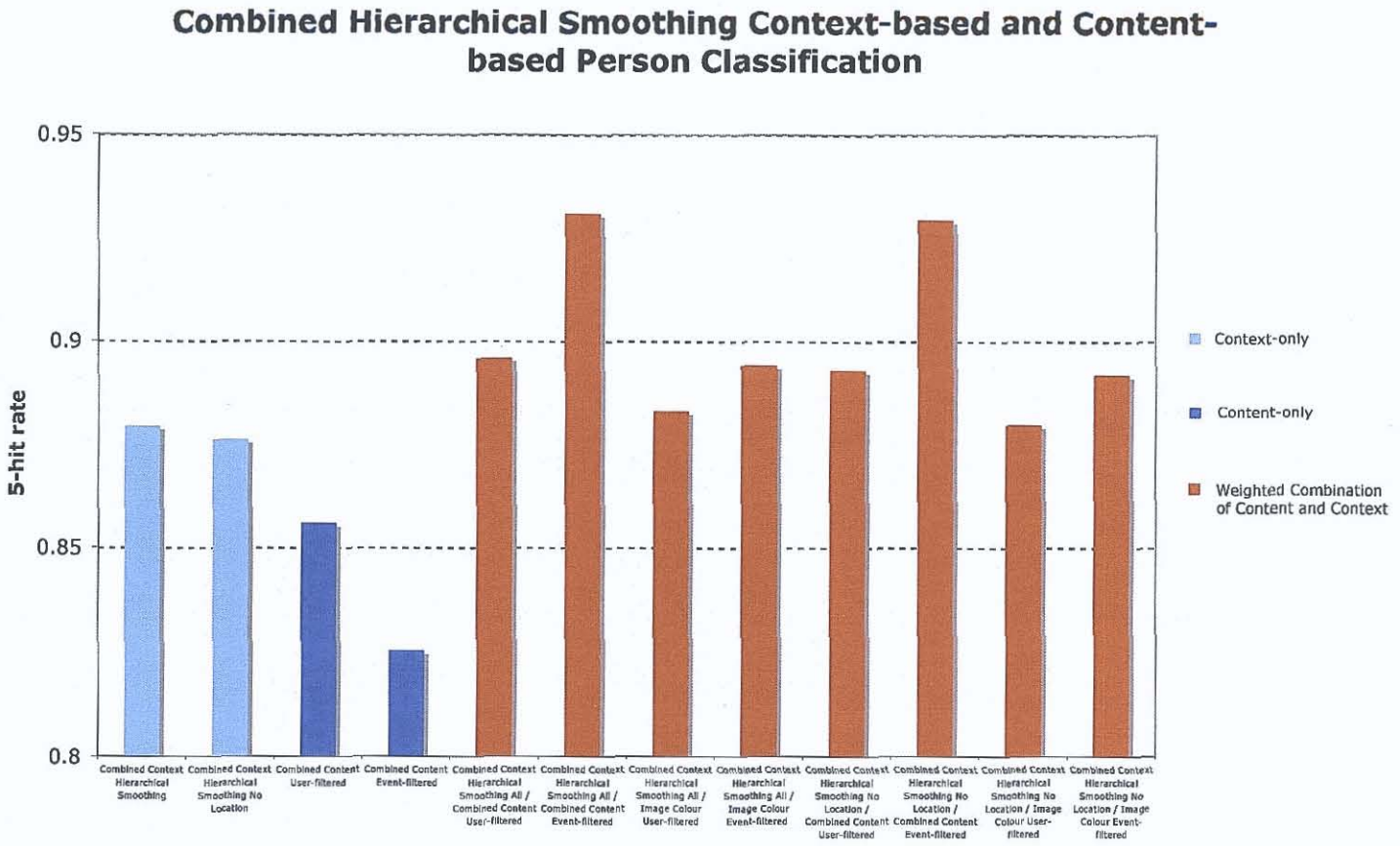


Figure 6.26: Combined context-based and content-based person classification: Weighted content and Hierarchical Smoothing context-based.

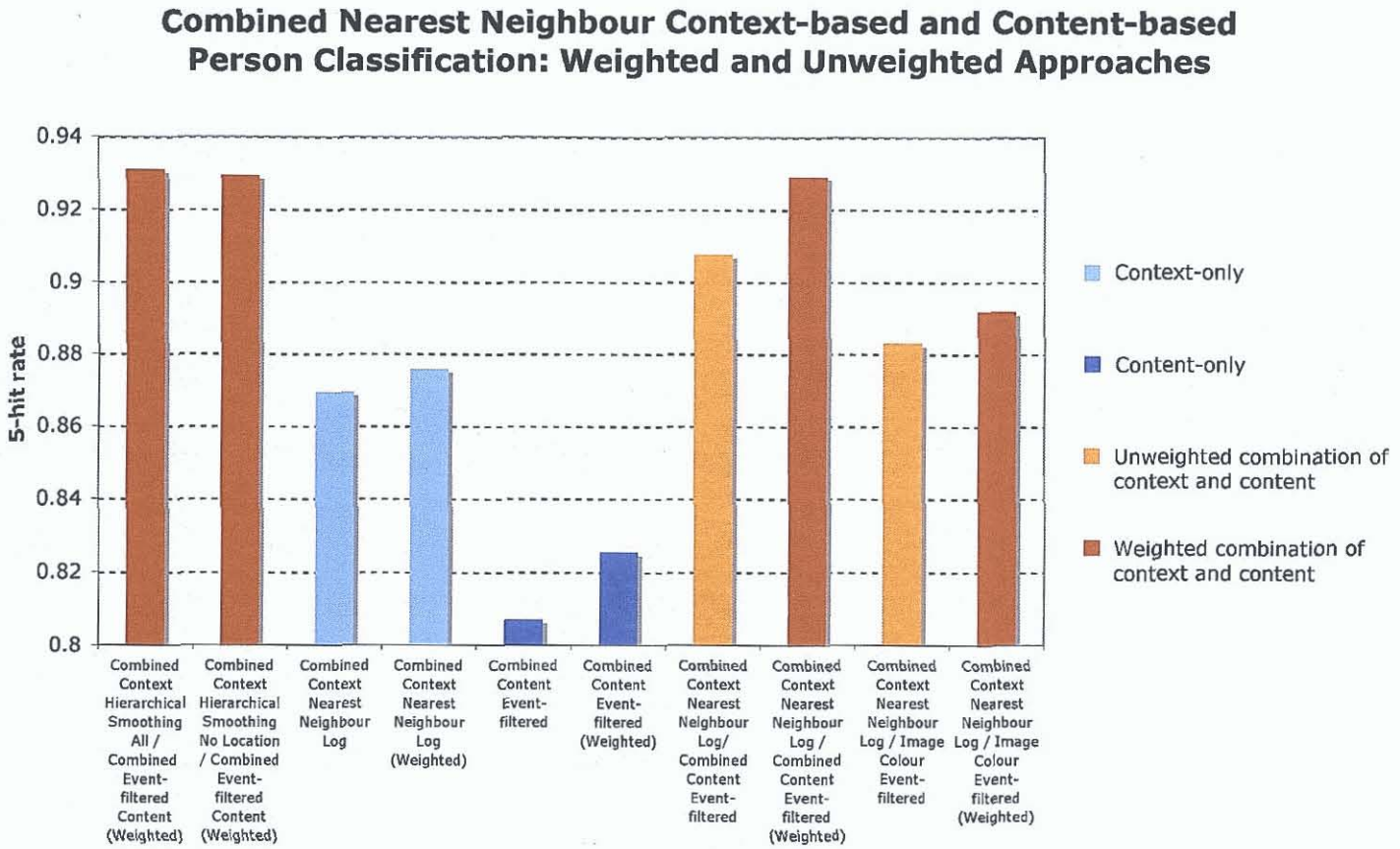


Figure 6.27: Combined context-based and content-based person classification: Nearest Neighbour context-based.

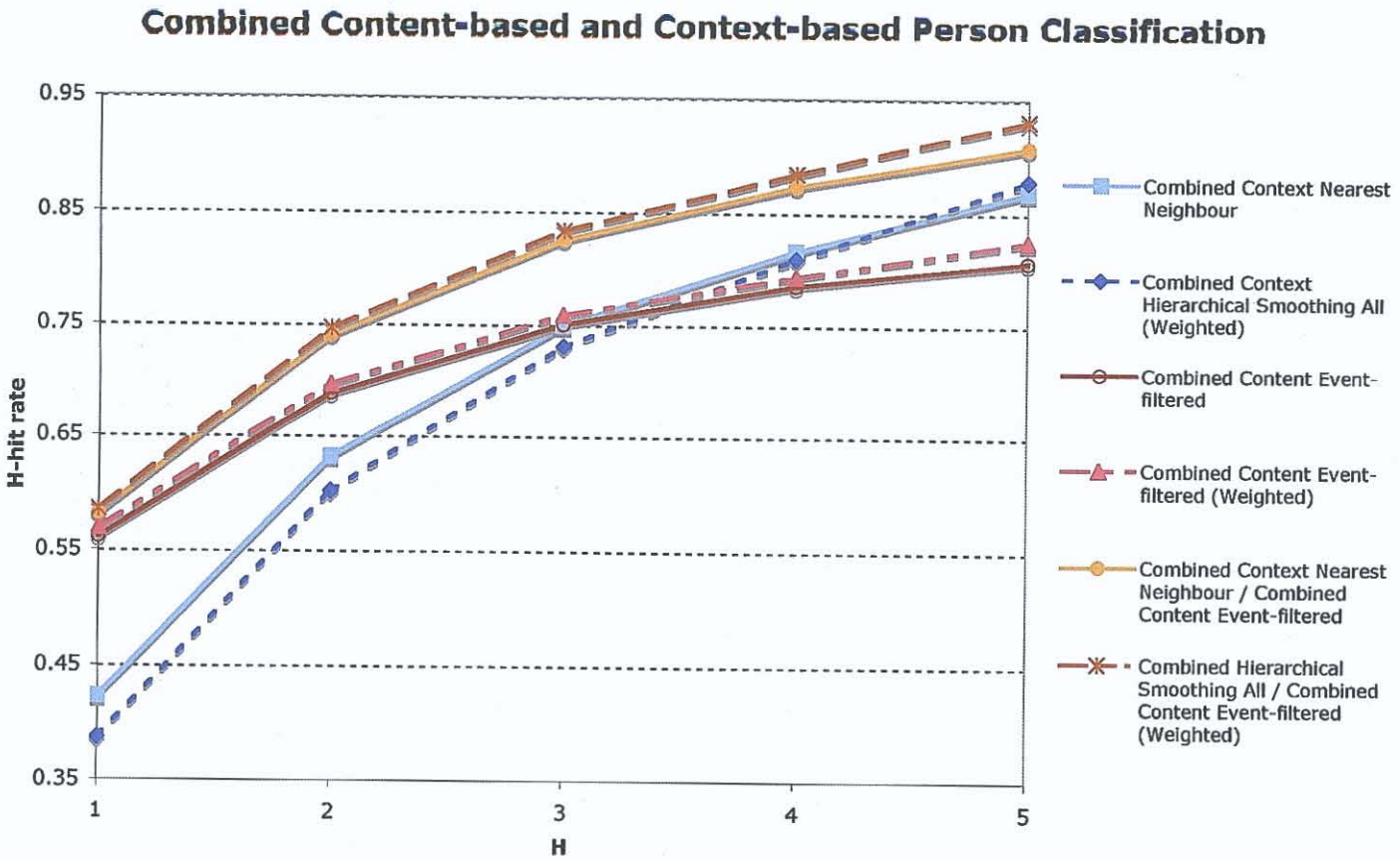


Figure 6.28: Combined content-based and context-based approaches to person classification: H-hit rate for various values of H.

We also evaluate combining context-based classification with content-based *image colour*. This allows us to compare the performance that we can achieve if we have photo-level annotations rather than face region-level annotations, since face region-level content-based approaches can only be used if we have region-level annotation. *Combined context hierarchical smoothing all / event-filtered image colour* achieves a *5-hit rate* of 0.894, which is a highly significant improvement over *combined context hierarchical smoothing all*, but still at long way off the performance of the region-based approaches, which have a best *5-hit rate* of 0.931.

Figure 6.27 shows the results from combining the best *combined context nearest neighbour* context-based techniques with *combined content* approaches. The performance of *combined context nearest neighbour log / combined content event-filtered (weighted)*, with a *5-hit rate* of 0.929, is only slightly outperformed by *combined context hierarchical smoothing all / combined event-filtered content (weighted)*, and the difference between these approaches is not statistically significant. *Combined context nearest neighbour log / combined content event-filtered (weighted)* highly significantly outperforms both *combined context nearest neighbour log (weighted)* and *combined content event-filtered (weighted)*. *Combined context nearest neighbour log / image colour event-filtered (weighted)* scores a *5-hit rate* of 0.892, which is highly significantly outperformed by *combined context nearest neighbour log / combined content event-filtered (weighted)*. This approach, however, does significantly outperform both *combined context nearest neighbour log (weighted)* and *image colour event-filtered*, and shows the level of performance we can achieve using only image-level features, without the region-level content-based features.

For unweighted combination approaches we restrict our evaluation to approaches which use *combined context nearest neighbour* as the combined context-based approach, because unweigthed combinations of *context hierarchical smoothing* actually diminish performance compared with *temporal proximity hierarchical smoothing* alone, and because these smoothing approaches already rely on smoothing weights. Although both of the unweighted approaches, *combined context nearest neighbour /*

combined content event-filtered and *combined context nearest neighbour / combined content image colour*, are significantly outperformed by their weighted equivalents, in each case the unweighted combination achieves a significant improvement over context alone and content alone The best-performing unweighted approach is *combined context nearest neighbour log / combined content event-filtered*, with a *5-hit rate* of 0.908.

Figure 6.28 shows combined content and context-based results, evaluated using *H-hit rate* for different values of *H*. Comparing content-based approaches with context-based approaches, we can see that content-based approaches outperform context-based approaches for *1-hit rate* and *2-hit rate*, with weighted and unweighted content-based approaches both outperforming context-based approaches for these evaluation measures. For *4-hit rate* and *5-hit rate*, however, context-based approaches perform better with the difference statistically significant for *5-hit rate*. The best content-only *1-hit rate* of 0.569 is far better that the best context-only *1-hit rate* of 0.422, showing that combined content is far more accurate for this more strict evaluation measure. The better *4-hit rate* and *5-hit rate* of the context-based approaches reflects the fact the event-filtered content-based approaches sometimes suffer when there are no annotations of the correct person name in the current event

We can see from Figure 6 28 that combined context-based and content-based approaches outperform content-only and context-only approaches for all values of *H*. *Combined context hierarchical smoothing all / combined content event-filtered (weighted)* highly significantly outperforms all content-only and context-only approaches for *2-hit rate*, *3-hit rate*, *4-hit rate* and *5-hit rate*. The unweighted *combined context nearest neighbour / combined content event-filtered* performs significantly better than unweighted content-only and context-only approaches for all values of *H*, and although it is outperformed by the weighted *combined context hierarchical smoothing all / combined content event-filtered (weighted)* approach, this difference quite small and is only significant for *4-hit rate* and *5-hit rate*.

6.6.2 Combined Person Retrieval

For the evaluation of context-based person retrieval, we evaluate combined context-based approaches for the *photo retrieval* scenario only. Because content-based approaches and combined context-based and content-based approaches rely on face region-level features in addition to phot- level features, we evaluate combined content-based approaches and combined context-based and content-based approaches for both the *photo retrieval* scenario and the *face region retrieval* scenario.

Similarly to how weights are learned for context-based smoothing approaches to person retrieval, the weights for weighted combination approaches to person retrieval are learned by optimising *MAP* on the test set.

Combined Context-based Person Retrieval

Figure 6.29 shows the results for combined context-based approaches to person retrieval. *Combined context hierarchical smoothing all (weighted)* performs highly significant better than *temporal proximity hierarchical smoothing*, the best-performing *hierarchical smoothing* approach, for all evaluation measures. The unweighted *combined context hierarchical smoothing all*, however, is outperformed by *temporal proximity hierarchical smoothing* for *MAP*, *P10* and *P20*, although the difference is not statistically significant for any of these measures As with the classification scenario, combined approaches perform well without the location information, with *MAP* of 0.470 and *P10* of 0 524 for *combined context hierarchical smoothing no location (weighted)*, compared with a *MAP* of 0.484 and *P10* of 0.535 for *combined context hierarchical smoothing all (weighted)*.

Combined context nearest neighbour approaches show a similar improvement. There is little difference between *combined context nearest neighbour (weighted)* and *combined context nearest neighbour log (weighted)*, and both slightly outperform *temporal proximity nearest neighbour*.

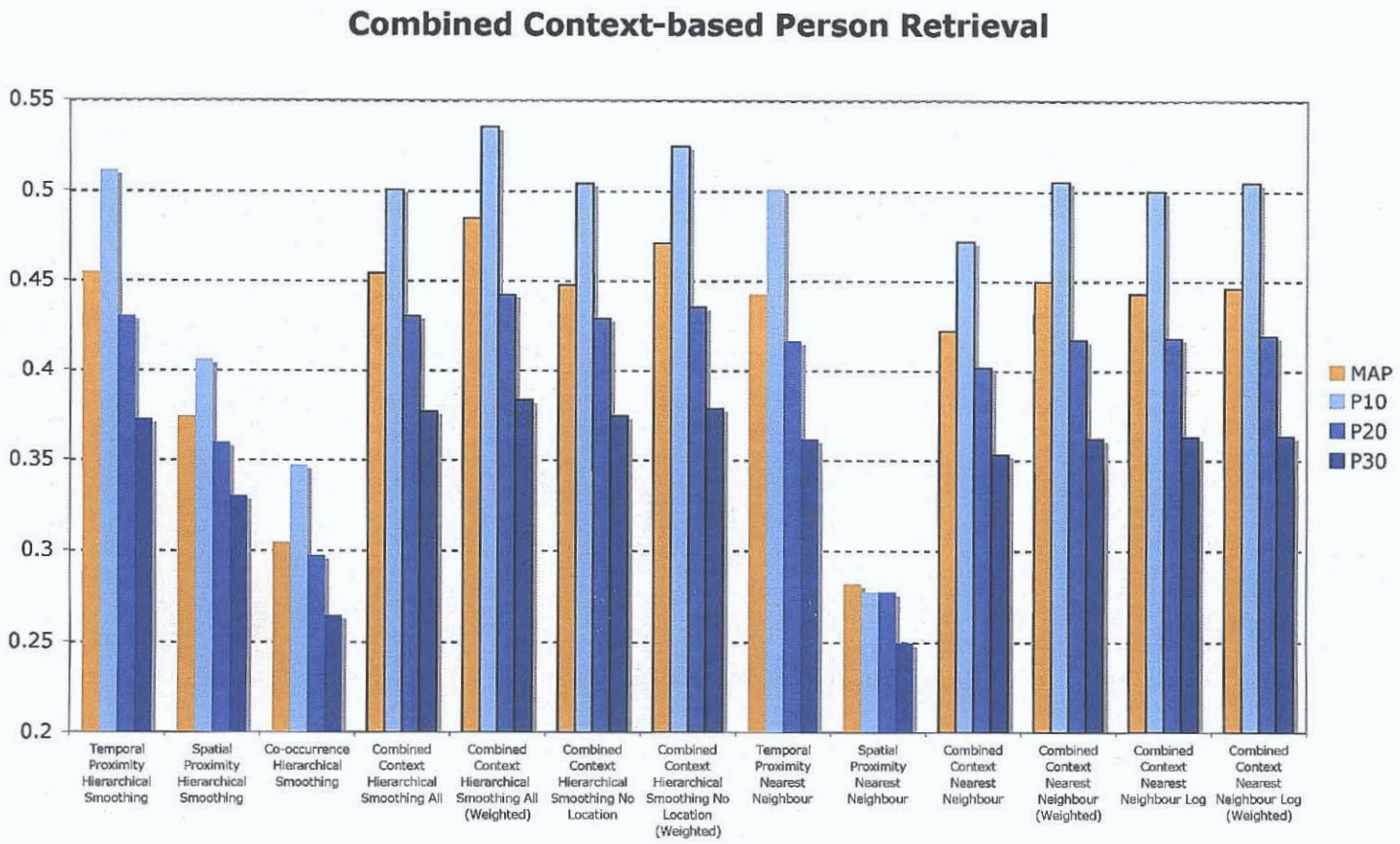


Figure 6.29: Combined context-based person retrieval results.

For unweighted combination of *temporal context nearest neighbour* approaches, *combined context nearest neighbour log* significantly outperforms *combined context nearest neighbour* for all evaluation measures *Combined context nearest neighbour*, in fact, performs worse than *temporal proximity nearest neighbour*, with a *MAP* of 0.422 and *P10* of 0.472 compared to a *MAP* of 0.442 and *P10* of for 0.501 *temporal proximity nearest neighbour* alone. We believe the reason that using log values is more important for the retrieval scenario than for the classification scenario is that in this scenario we are ranking all unknown faces, so it is much more likely that the maximum time difference will be very large, which can cause problems when normalising the scores from *temporal proximity nearest neighbour*. Unweighted *combined context nearest neighbour log* gives a slight improvement in *MAP*, *P20* and *P30* over *temporal proximity nearest neighbour*, although this is not statistically significant

Comparing *combined context hierarchical smoothing* approaches with *combined context nearest neighbour* approaches, we can see that *combined context hierarchical smoothing all (weighted)* outperforms *combined context nearest neighbour log (weighted)*, and this difference is significant for *MAP* and *P30*. Unweighted *combined context hierarchical smoothing all* outperforms *combined context nearest neighbour log* for all evaluation measures, although this is not statistically significant. *Combined context hierarchical smoothing all*, however, is not a truly unweighted approach because the individual *hierarchical smoothing* approaches use smoothing weights. For this reason we will not consider these approaches when evaluating unweighted combined context-based and content-based person retrieval approaches below.

Combined Content-based Person Retrieval

In Figure 6.30 we can see the results for combined content-based retrieval for the photo retrieval scenario. For user-filtering, unweighted and weighted *combined content* approaches both highly significantly outperform all individual approaches for all evaluation measures, with *combined content user-filtered (weighted)* improving *MAP*

from 0.396 to 0.493 and *P10* from 0.504 to 0 608 compared with the best individual user-filtered approach For event-filtering, *combined content event-filtered (weighted)* performs highly significantly better than all individual event-filtered approaches for all evaluation measures.

The unweighted *combined content event-filtered* significantly outperforms all individual event-filtered approaches for *MAP* and *P10*, and significantly outperforms all individual approaches except *body patch event-filtered* for *P20* and *P30*.

Comparing user-filtered results with event-filtered results, *combined content event-filtered* highly significantly outperforms *combined content user-filtered* for *MAP*, *P10* and *P20*, and significantly outperforms it for *P30*. Also, *combined content event-filtered (weighted)* highly significantly outperforms *combined content user-filtered (weighted)* for *MAP*, *P10* and *P20*

The combined content-based retrieval results for the face region retrieval scenario, shown in Figure 6.31, show a similar improvement with the use of combined approaches. For user-filtering, both unweighted and weighted approaches again highly significantly outperform all individual approaches for all evaluation measures. For event-filtering, *combined content event-filtered (weighted)* significantly outperforms all individual approaches for all measures.

Unlike for the photo retrieval scenario, the unweighted *combined content event-filtered* approach only improves on the best-performing individual approach, *body patch event-filtered*, for *MAP* and *P10*, and for these two measures the difference is not significant. For *P20* and *P30*, on the other hand, *body patch event-filtered* actually outperforms *combined content event-filtered*, a difference which is significant for *P30*. This emphasises the strong performance of the *body patch event-filtered* approach for the face region retrieval scenario, where the difference in performance between this approach and other approaches is large enough that, without using appropriate weights, combining it with other approaches will not give a clear improvement in performance.

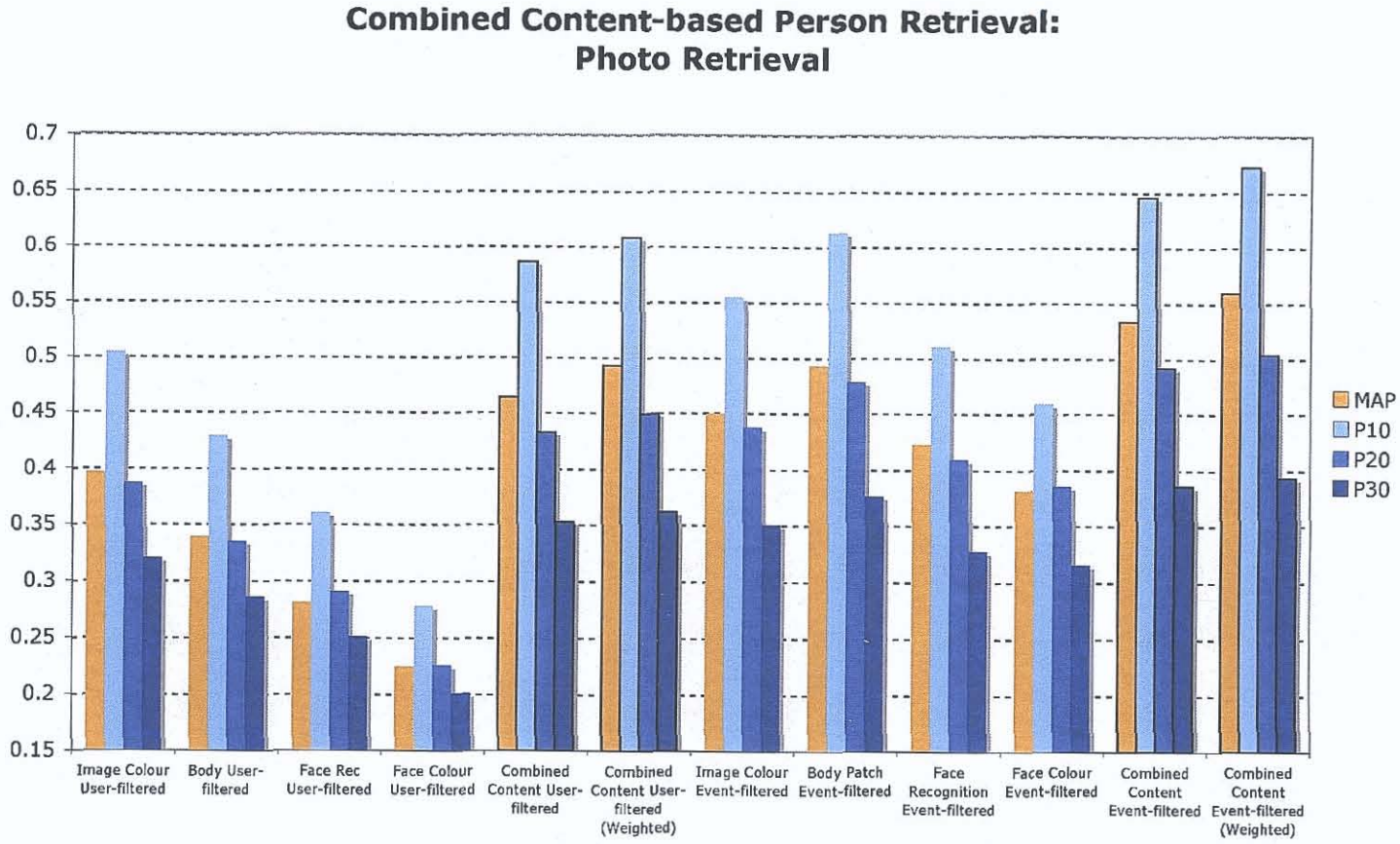


Figure 6.30: Combined content-based person retrieval results: photo retrieval.

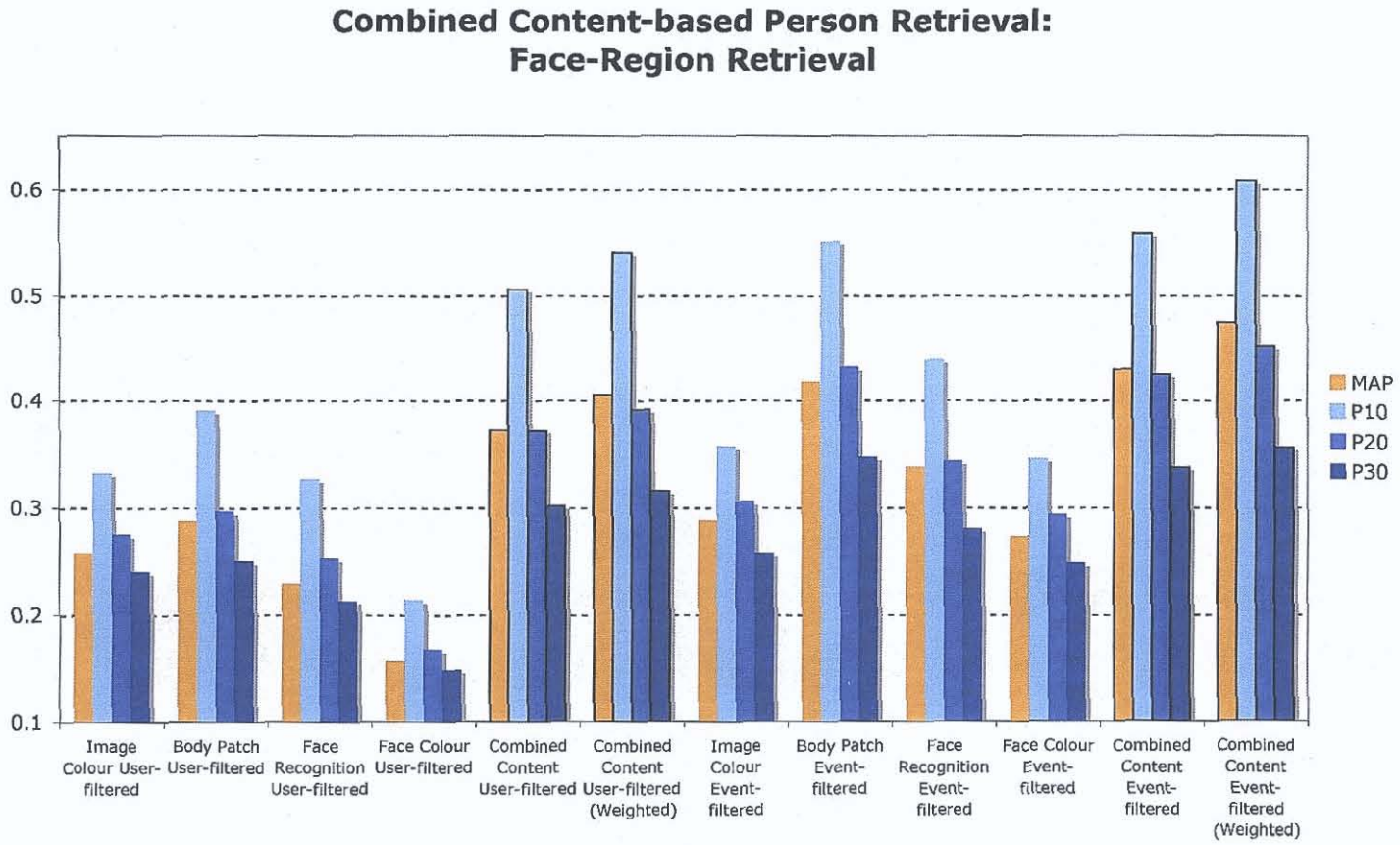


Figure 6.31: Combined content-based person retrieval results: face region retrieval.

We believe the reason for the particularly strong performance of *body patch* in the retrieval scenario, in comparison with the classification scenario, is that the results for the retrieval scenario are dominated by those cases for which the *body patch* feature is most confident, as these are the results that are found towards the top of the ranked list. In those cases where the *body patch* feature is most confident, its results are more reliable than the results from other features. This causes *body patch* to have a better *1-hit rate* performance for person classification, and for retrieval it means that the difference between *body patch* and other features is greater than for the classification task, where the overall performance is determined by average classification accuracy for all regions. This difference is further emphasised in the more difficult face region-level retrieval task

As with the photo-level retrieval scenario, event-filtering outperforms user-filtering in the face region-level retrieval scenario. For both weighted and unweighted combinations event-filtering highly significantly outperforms user-filtering for *MAP*, *P10* and *P20*, and significantly outperfoms user-filtering for *P30*

The photo-level retrieval results, with a best *MAP* of 0.559 and a best *P10* of 0.672, are always better than those from face region-level retrieval, which has a best *MAP* of 0.473 and a best *P10* of 0.608. This is expected, of course, because face region-level retrieval is a more difficult task and so will always be outperformed by photo-level retrieval. If we look at the photo-level retrieval results without using region-based features, however, we have *MAP* of 0.45 and *P10* of 0.555 for *image colour event-filtered*, which is outperformed by the face region-level retrieval result for *combined content event-filtered (weighted)* for all evaluation measures, although this is only significant for *P10* This suggests that, for combined content-based person retrieval approaches, region-based retrieval using region-based features can outperform photo-based retrieval using global image-based features. We will see below, however, that this improvement is lost when we combine content-based approaches with context-based approaches.

Combined Context-based and Content-based Person Retrieval

Figure 6.32 shows the combined context-based and content-based results for the photo-level retrieval scenario. If we compare context-based retrieval approaches with content-based approaches, we can see that the event-filtered content-based approaches outperform the context-based approaches, with *combined content event-filtered (weighted)* highly significantly outperforming *combined context hierarchical smoothing all (weighted)*, *combined context hierarchical smoothing no location (weighted)* and *combined context nearest neighbour log (weighted)* for *MAP*, *P10* and *P20*. Similarly, the unweighted *combined content event-filtered* highly significantly outperforms the unweighted *combined context nearest neighbour log* for *MAP*, *P10* and *P20*. This shows that, for person retrieval, using event-filtering with content-based approaches performs better than the best context-based approaches.

Weighted combinations of content-based and context-based approaches to photo-level retrieval always achieve a significant improvement in performance over the best-performing content-based approach, with *combined context hierarchical smoothing all / combined content event-filtered (weighted)* improving *MAP* from 0.559 to 0.605 and improving *P10* from 0.672 to 0.692. Removing the location information affects performance only slightly, giving a *MAP* of 0.601 and a *P10* of 0.687.

The difference between *combined context hierarchical smoothing all / combined content event-filtered (weighted)* and *combined context hierarchical smoothing no location / combined content event-filtered (weighted)* is only statistically significant for *MAP*. *MAP*, however, is the evaluation measure we optimise when learning the weights, so it is likely that there is no real improvement in performance from using location information with this dataset.

If we discard the region-level features and assume that our system only has photo-level annotations available to it, then we can see that *combined context hierarchical smoothing all / image colour event-filtered (weighted)* achieves a *MAP* of 0.524 and a *P10* of 0.596, which is a highly significant improvement over both *combined context hierarchical smoothing all* and *image colour event-filtered*.

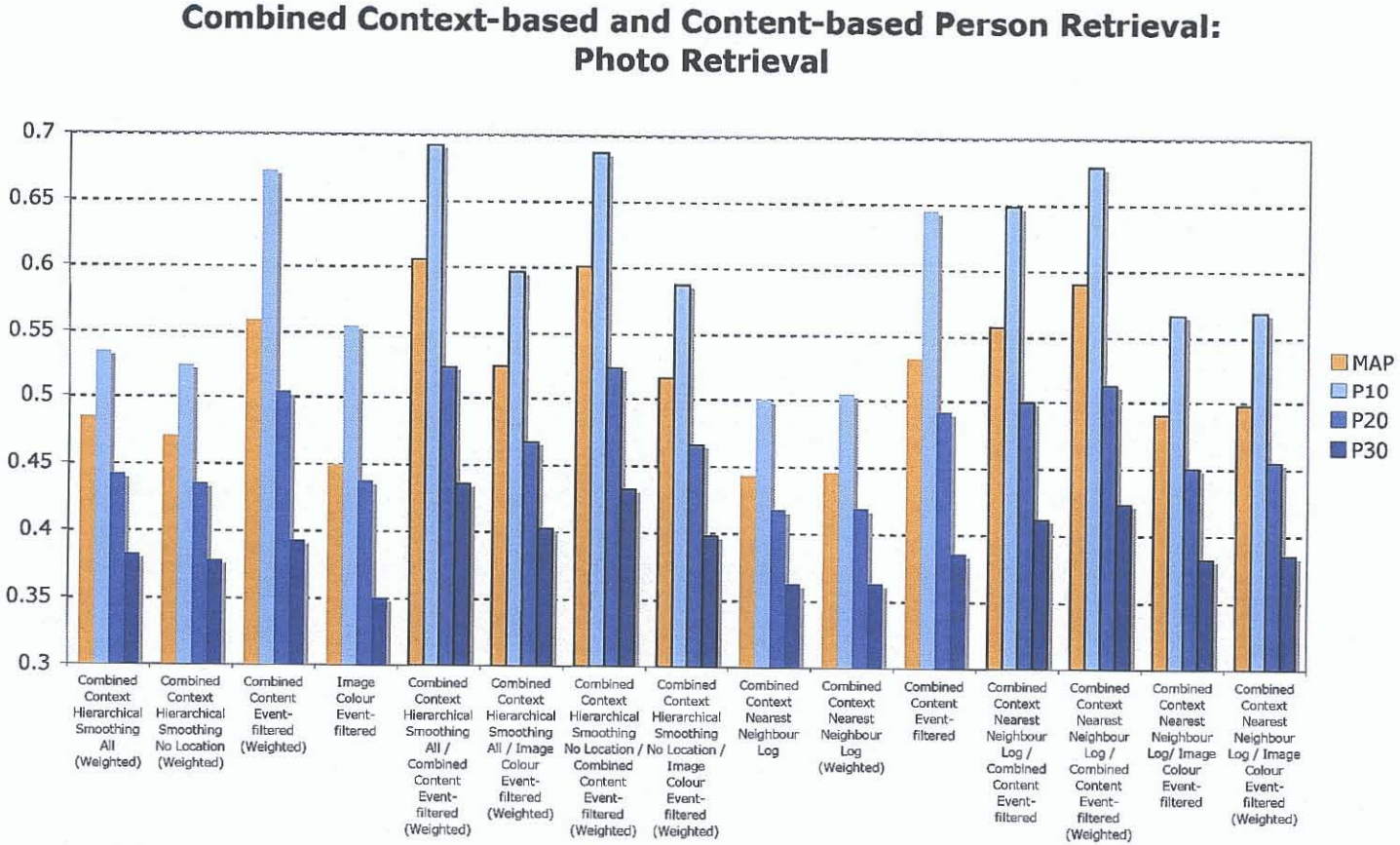


Figure 6.32: Combined context-based and content-based person retrieval results: photo retrieval.

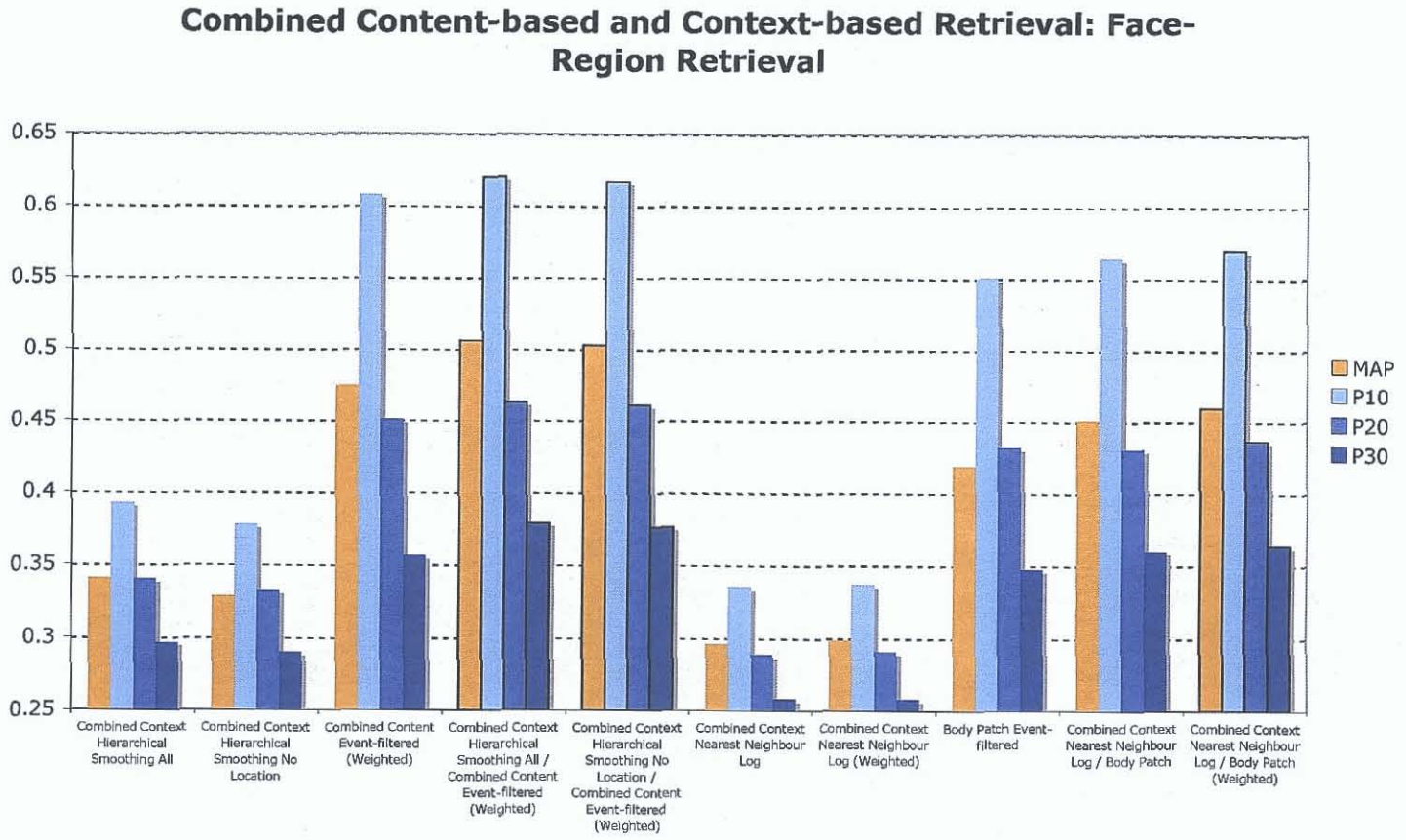


Figure 6.33: Combined context-based and content-based to person retrieval results: face region retrieval.

Looking at the unweighted *nearest neighbour* approaches, we can see that *combined context nearest neighbour log / combined content event-filtered* improves all evaluation measures compared with *combined content event-filtered* alone, improving *MAP* from 0.533 to 0.557, and *P10* from 0.645 to 0.648. *Combined context nearest neighbour log / combined content event-filtered* is significantly better than *combined content event-filtered* for *MAP* and *P30*.

If we only use photo-level content-based features, *combined context nearest neighbour log / image colour event-filtered* gives a highly significant improvement for all evaluation measures compared with *combined context nearest neighbour log* alone, and it gives a significant improvement over *image colour event-filtered* for *MAP*, *P10* and *P30*.

Figure 6.33 shows the combined content-based and context-based results for the face region-level retrieval scenario. *Combined context hierarchical smoothing all / combined content event-filtered (weighted)* significantly improves performance compared with *combined context hierarchical smoothing all* alone and *combined content event-filtered (weighted)* alone for all evaluation measures, with a *MAP* of 0.506 and a *P10* of 0.619.

Again, unweighted approaches offer an improvement over content alone and context alone. The best unweighted approach, *combined context nearest neighbour / body patch event-filtered*, significantly improves all evaluation measures compared with *combined context nearest neighbour*. Compared with *body patch event-filtered* alone and *combined content event-filtered* alone, *combined context nearest neighbour / body patch event-filtered* improves all evaluation measures, a difference which is significant for *MAP* and *P30*.

Photo-level retrieval results using photo-level features outperform face region-level results using region level features for *MAP*, *P20* and *P30*, although this difference is only significant for *P30*. For *P10*, face region-level retrieval, with a score of 0.619, outperforms photo-level retrieval using photo-level features, with a score of 0.596, although this is not statistically significant. This suggests that the accuracy of

combined context-based and content-based approaches to face region-level retrieval is little better than the accuracy of approaches to photo retrieval which use photo-level features only, and so an annotation system which uses region-level features may not in fact benefit from these region-level features, because such a system requires region-level annotation from the user. If this region-level annotation is facilitated by person suggestions provided by face region-level retrieval, then these results make it unclear whether it is more efficient to facilitate region-level annotation using person suggestions provided using face region-level retrieval, or to to facilitate photo-level annotation using person suggestions provided by photo-level retrieval. In practice we prefer to use face region-level annotation· since there is not a large difference in performance, we prefer to extract the more detailed face region-level annotations from the user, for a similar amount of user effort

6.7 Summary

We have evaluated context-based and content-based approaches to person classification and retrieval in partially annotated context-aware personal photograph collections. For person classification and retrieval, we have shown that our proposed context-based *smoothing* and *hierarchical smoothing* language model approaches, and our *nearest neighbour* approaches all outperform the *MLE* approach.

The best-performing context-based approach for both the person classification and person retrieval tasks is *temporal proximity*, which significantly outperforms *spatial proximity*, *cyclic temporal context* and *co-occurrence*. *Hierarchical smoothing* outperforms *nearest neighbour*, although this difference not statistically significant for the *temporal proximity* feature, showing that the simple *temporal proximity nearest neighbour* approach is very powerful for person classification and retrieval.

Combining context-based approaches improves performance for both classification and retrieval, with weighted approaches giving the best performance. Unweighted combination approaches give an improvement when combining *nearest*

neighbour approaches but not when combining *hierarchical smoothing* approaches.

For content-based person classification and retrieval we evaluate four content-based features: *face recognition*, *face colour*, *body patch* and *image colour*. For classification, each of these features improves on simple context-based *MLE* ranking, with all features benefiting from event-filtering For event-filtered content-based person classification, *face recognition* outperforms all other approaches in terms of *5-hit rate*, although if we use the finer-grained evaluation measure *1-hit rate* then *body patch* is the best-performing feature. For content-based person retrieval, event-filtering approaches again outperform user-filtering approaches, and for both the photo-level retrieval scenario and the face region-level retrieval scenario *body patch* is the best-performing approach.

Combining different content-based approaches to person classification is particularly effective and gives a highly significant improvement over the best-performing individual content-based approaches. There is little difference between weighted and unweighted combined content-based approaches to person classification For person retrieval, combined content-based approaches again improve performance over the best individual approach For the photo-level retrieval scenario, weighted and unweighted content-based approaches both give an improvement over the best individual approach. For the face region-level retrieval scenario, only weighted combination approaches give a significant improvement over the best individual approach

Content-based approaches to person classification outperform context-based approaches for the *1-hit rate* and *2-hit rate* evaluation measures. Context-based approaches to classification, however, perform better than content-based approaches for the coarser *4-hit rate* and *5-hit rate* evaluation. For person retrieval, event-filtered content-based approaches outperform context-based approaches, although these event-filtered content-based approaches are, of course, context-aware. Combining context-based and content-based approaches to person classification and retrieval improves the performance over either approach used alone, an improvement given by both weighted and unweighted approaches

The best face region retrieval results using region-level features are no better than the best photo-level retrieval results which use photo-level features. Since the purpose of the proposed retrieval approaches is to provide annotation suggestions to facilitate semi-automatic person annotation, this suggests that there may be no benefit in encouraging users to annotate their images to the face region level. We prefer, however, to extract the more detailed face region-level annotations from the user for a similar cost in terms of user effort

6.8 Conclusions

While the best-performing context-based approach to person classification and retrieval is the hierarchical smoothing approach, we believe in a real deployment it would be preferable to use the nearest neighbour context-based approach. This simple approach does not need to learn parameters like the language model approach, and can be expected to work well with a less heavily annotated training set, unlike the language model approach. Our results also show that using event-filtering for content-based approaches is always preferable to user-filtering. So, in conclusion, for both person classification and retrieval our preferred approach would be to use combined nearest neighbour context-based analysis and combined event-filtered content-based analysis, and to use a weighted combination of these two approaches to give an approach based on context and content. We believe it would be possible to learn a set of generic weights that would be appropriate for all users

Although our evaluation uses a 50% annotated training set, we believe that our approaches can give effective results with a smaller training set The approaches should work quite well in any events that contain a small amount of annotated faces, enabling batch annotation in these events and accelerating the annotation process.

Chapter 7

Evaluation of Semi-Automatic Person Annotation

In this chapter we evaluate proposed interfaces for semi-automatic person annotation in a personal photo management system. We compare the classification-based semi-automatic annotation approach from the MediAssist system with the enhanced, retrieval-based semi-automatic batch annotation interface proposed in Chapter 4. Seven users each completed a number of interactive annotation tasks using each of the two systems, and we evaluated the performance of each system using both quantitative measures of interactive annotation effectiveness, and qualitative measures of subjective user satisfaction with the system.

7.1 Test Collection and Users

For the evaluation of semi-automatic person annotation, we use a subset of the 9 users whose collections were used for the evaluation of person classification and retrieval in the Chapter 6. Of those 9 users, one is the author of this thesis, who could not participate as an unbiased user for interactive experiments. One further user was also unavailable for interactive experiments, leaving 7 users with large enough location-aware photo collections for interactive experiments. Each user has

contributed their own photos to the MediAssist system and so each user will be carrying out annotation experiments on their own personal photos that they contributed to the system, and so they will have intimate knowledge of these photos and the people they contain.

For the evaluation of person classification and retrieval in Chapter 6 we partitioned each user's person annotations into training sets and test sets, using cross-validation techniques to create 5 of such partitions. For these interactive experiments we take one such partition for each user and take the training set as the set of faces already annotated by that user Given a 50% annotated collection from each partition, the remaining 50% of unannotated faces are available for annotation during interactive experiments.

All users were either postgraduate students in computing or lecturing staff in the computing faculty in Dublin City University. As such, they can all be considered to be advanced users. Of the 7 users, 6 report that they use photo management system "a few times a month", with one user reporting the he "never" uses photo management systems, and 3 of the 7 users reported that they currently spend no time whatsoever annotating their personal photo collection.

Information about the personal photo collections of the 7 users for interactive experiments is given in Table 7 1. These users have an average of 2,255 photos in the MediAssist system, with an average of 544 unannotated faces, found in an average of 365 photos, in each collection. We also asked each user to estimate the size of their entire photo collection (i.e. those photos in the MediAssist system in addition to those photos in the user's personal photo collection not made available to the MediAssist system). On average the users estimated they had 7,857 photos, which is over 3 times the average collection size in the MediAssist system. This suggests that, if the users forget which photos they contributed to MediAssist and which they did not, then they could possibly spend time searching for photos which are not in the collection. We will outline the steps that we take to minimise this in Section 7 3.1.

User	Total Photos: MediAssist (Estimated Total)	Un-Annotated Faces	Photos With Un-Annotated Faces	Number of Annotation Tasks	Average Un-Annotated Occurrences per Task
A	5,231(20,000)	328	259	4	30
B	3,435 (5,500)	1209	789	8	58
C	2,672 (8,000)	184	142	6	18.2
E	1,974 (12,000)	305	202	6	26
G	1,208 (6,000)	835	605	10	66
J	753 (1,500)	601	349	8	25.9
L	513 (2,000)	344	211	8	17 9
Avg	2,255 (7,857)	544	365	7	35

Table 7.1: Details of the collections of the 7 users for interactive semi-automatic person annotation experiments. User labels are consistent with Table 3 1 in Chapter 3 and Table 6.1 in Chapter 6. The 'estimated total' is the user's own estimate of how many photos they have in their entire personal photo collection.

More information about these collections can be found in Table 3 1 in Chapter 3 and Table 6 1 in Chapter 6.

7.2 Experimental Systems

Each user's collection of personal photos was indexed using the best-performing person classification and retrieval approaches, as proposed in Chapter 5 and evaluated using the methods described in Chapter 6, to provide classification and retrieval engines for the interactive systems. For both person-classification and person-retrieval, the best-performing approach was *combined context hierarchical smoothing all / combined content event-filtered (weighted)*, so this approach is used for both the classification and retrieval engines in the interactive experiments reported in this Chapter. All user annotations made during the experiment are stored in the system, and faces which are annotated will not be suggested as candidates for future annotations Due to the execution times required by the prototype implementation of these indexing approaches, however, we do not re-index the collection after each

annotation, meaning that candidate faces are not compared to the new annotations when calculating classification and retrieval results The development of approaches to efficiently update the index after each annotation is clearly worthy of future investigation, although we consider it outside the scope of this work. We note, however, that even with this handicap the annotation systems performed well in this experimental scenario and, as we will see in what follows below, were well-received by all users

The two interactive systems which we used to evaluate the effectiveness of our proposed semi-automatic person annotation approaches are outlined in the next two subsections.

7.2.1 MediAssist: Semi-Automatic Person Annotation

The first, baseline, system is a slightly simplified version of the MediAssist photo management system described in Chapter 3. The system allows browsing and searching of personal photo collections based on context-based and content-based features. When the user encounters faces that have not been annotated the system presents a ranked list of suggested names for these faces, powered by the classification techniques proposed in Chapter 5.

There are two differences between the system used here and that described in Chapter 3. Firstly, to simplify the interface, the text search feature described in Section 3.6.2 is disabled. In addition, because we are interested in evaluating person identity annotation and not in annotating the presence of faces, we use a manually annotated ground truth of detected faces. The means that, as in Chapter 6, we assume that perfect face detection is available. Because of this, the semi-automatic annotation tools which allow the user to remove incorrect face detections and to add missed faces are now redundant in this system, and so we remove these tools from the interface.

7.2.2 MediAssist Enhanced: Semi-Automatic Batch Person-Annotation

The second system adds the semi-automatic batch person annotation enhancements to the basic MediAssist system, described in Chapter 6, to the *MediAssist: Semi-Automatic Person Annotation* system. These enhancements provide the user with suggested faces to match specific names, prompting the user to confirm the suggested faces as correct or to reject them as being incorrect, with the batch annotation suggestions powered by the person retrieval techniques proposed in Chapter 5. We call this system *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* because it allows the user to annotate a person in a *batch* fashion, suggesting a batch of faces to annotate, given a person name.

We expect that using this batch annotation feature will facilitate more efficient annotation, with the user annotating more faces in less time. We also hope that the system will lead to greater user satisfaction with the system, and encourage users to input annotations.

7.3 Experimental Methodology

7.3.1 Task

In order to evaluate the effectiveness of the proposed systems for semi-automatic person annotation, we presented users with the scenario that they have been using the MediAssist sytem for their day-to-day photo management, and that they have previously annotated a subset of the faces in their personal photo collection using this system. We then ask each user to complete a number of annotation tasks, where each task requires the user to annotate all, or as many as possible, of the occurrences of a specific target person in their collection within a specified time limit The users were introduced to *MediAssist: Semi-Automatic Person Annotation* as 'System A' and *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* as 'System

B' The use of each system for the annotation tasks was alternated, with one system used for the first annotation task, the other system for the next annotation task, and so on, with a time limit of 5 minutes set for each task. Also, the choice of which system was to be used for the first annotation task by each user was alternated, with half of the users starting with *MediAssist: Semi-Automatic Person Annotation* and the other half starting with *MediAssist Enhanced: Semi-Automatic Batch Person Annotation.*

As with the person-retrieval experiments described in Section 6.4 2, we only considered people/faces with an occurrence frequency of over 20 within each user's collection as suitable for evaluation, as otherwise there would not be enough unannotated faces for the user to annotate. This meant that some users had more people who were suitable for annotation tasks than other users. The user with the lowest number of suitable people was given 4 annotation tasks, 2 tasks for each system variant. Rather than limiting all users to this small number of tasks, we gave users with a greater number of suitable people/faces in their collection more annotation tasks. The number of annotation tasks for each user can be seen in Table 7 1

The day before running the experiment, each user was introduced to the *MediAssist: Semi-Automatic Person Annotation* system and given training on how to use the system. The user was then instructed to spend between 30 minutes and 1 hour familiarising themselves with the system. The reason for this is that we were attempting to replicate a scenario where the user uses this system for their day-to-day photo management, and where they have already annotated a subset of the faces in their collection using this system. Accordingly, by familiarising themselves with the system, the users should be comfortable with the search and browsing tools provided, and they should be familiar with the subset of faces in their collection that are currently annotated. In addition, as shown in Table 7.1, each user's complete, real world, personal photo collection is much larger than the subset made available to MediAssist for these experiments. Spending time with the system also served as a reminder to the users about which photos from their collections they made available

to the MediAssist system, and which photos they did not make available

7.3.2 Evaluation Measures

We evaluate the performance of each system using both quantitative and qualitative measures. For the quantitative evaluation we logged all user interaction with the system during the experiments. From these logs we can measure the number of occurrences of the target person successfully annotated at any time during the experiment. We then calculate a measure we call *annotation coverage*, the percentage of target person occurrences annotated, for each task, as follows:

$$\textit{Annotation Coverage} = \frac{\text{Num Target Person Occurrences Annotated by User}}{\text{Num Target Person Occurrences Available for Annotation}}$$

We measure user effort in terms of the number of interactions with the system, recorded as the number of mouse clicks. A mouse click was registered if the user used any of the system's query formulation tools, if the user made use of the browsing tools to change the current photo view, or if the user clicked any of the annotation buttons in the interface. For this experiment, scrolling was not recorded as a mouse click. We also calculate a measure we call *clicks per annotation*, which can be seen as a measure of the user effort required for each successful annotation:

$$\textit{Clicks Per Annotation} = \frac{\text{Number of Mouse Clicks during Annotation Task}}{\text{Num Target Person Occurrences Annotated by User}}$$

We calculate these quantative measures for each annotation task. We then calculate the average results for each user before averaging across users, to ensure that users with more annotation tasks do not dominate the results. This is consistent with the evaluation methodology in Chapter 6, where we average results for each user, then average across users, to avoid users with more data from biasing the results.

For qualitative evaluation we provided users with a post-experiment question-

naire that asked the users to report agreement or disagreement with various statements about the system using a 7-point Likert scale (Likert, 1932).

We use statistical tests to test the hypothesis that the differences between the *systems* is significant against the null hypothesis the difference is due to chance For the quantitative evaluation, as in Chapter 6, the statistical test used is randomization testing (Kempthorne and Doerfler, 1969) For the qualitative results, measured on a Likert scale, we use the Wilcoxon signed-rank test as this is more suitable for ordinal data (Wilcoxon, 1945). The input into the tests are the results for each user for the two systems, so we are effectively testing if the differences between runs are consistent across users, or if they are due to chance variations between users. As in the Chapter 6, whichever test is used, if a result is better than another result at a significance level of 0.05 we say it is *significant*, and if it is better at a significance level of 0.01 then we will say it is *highly significant*

7.4 Interactive Semi-Automatic Person Annotation Results

In this section we present the results of the user experiments, firstly presenting quantitative results measuring annotation effectiveness, followed by the qualitative results measuring user satisfaction.

7.4.1 Annotation Effectiveness: Quantitative Results

Figure 7.1 plots the average annotation coverage, measured as the percentage of target person occurrences annotated against time, taken at 10-second intervals for both the *MediAssist. Semi-Automatic Person Annotation* and *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* systems. We can see that growth is very slow for the first 30 seconds for both systems. After approximately 40 seconds, *MediAssist Enhanced· Semi-Automatic Batch Person Annotation* increases its

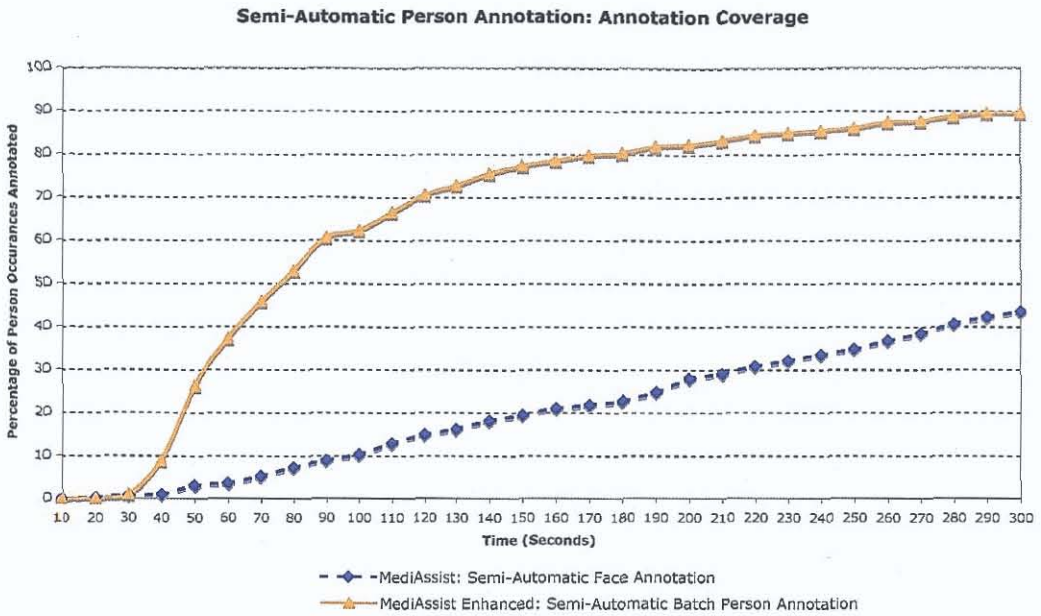


Figure 7.1: Annotation coverage plotted against time for interactive person annotation.

annotation coverage dramatically, roughly corresponding to the point at which the user encounters the batch annotation pop-up window and begins annotating faces in a batch manner. This dramatic growth continues until about 90-120 seconds have passed, corresponding to the lowering in accuracy of the suggested batch annotations. The annotation coverage continues to grow steadily, however, levelling out in the last 30 seconds or so and reaching an average of 89.5% of occurrences of the target person annotated after the 5-minute time limit has passed.

MediAssist: Semi-Automatic Person Annotation, on the other hand, achieves a slow and steady growth in annotation coverage as the user browses the collection and finds occurrences of the target person. After the full 5 minutes have passed, an average of 43.4% of occurrences of the target person have been annotated. *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* achieves a highly significantly better annotation coverage score than *MediAssist: Semi-Automatic Person Annotation* at all time points after 60 seconds have passed. In fact, after 60 seconds *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* has an annotation coverage of 37.3%, compared to 3.7% for the *MediAssist: Semi-Automatic Person*

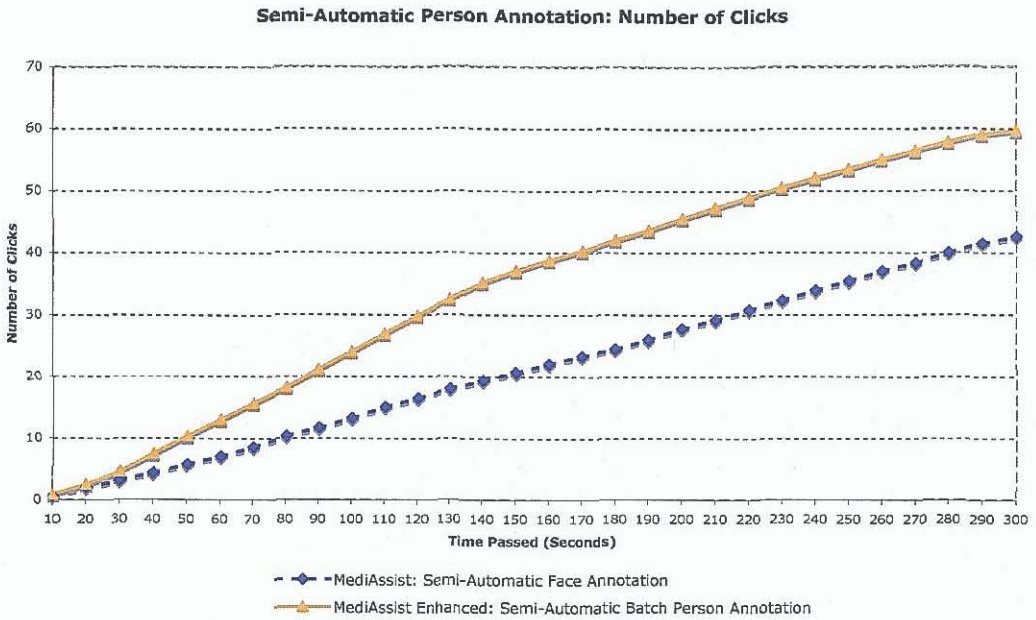


Figure 7.2: Number of user mouse clicks plotted against time for interactive person annotation.

Annotation system, which is just over 10 times more annotation coverage. By the end of the task, after 5 minutes have elapsed and the batch annotation system's annotation coverage growth has slowed, it has over twice the annotation coverage compared to *MediAssist: Semi-Automatic Person Annotation.*

Figure 7.2 looks at the average number of mouse clicks made by the user in interacting with the system. *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* has an average of 59.6 clicks after the 5 minutes have passed, compared with 42.5 clicks for the *MediAssist: Semi-Automatic Person Annotation*, reflecting the fact that with the enhanced system the user is busy confirming and rejecting annotation suggestions, and so has more interaction with the system.

Figure 7.3 shows the average number of mouse clicks per annotation. We begin to plot the results only after 100 seconds have passed because for a number of the tasks the user had still not made any annotations before that time, meaning that there is insufficient data to calculate average scores before this point. For *MediAssist Enhanced: Semi-Automatic Batch Person Annotation*, we can see that this ratio reaches its lowest value after 120 seconds at 2.1 clicks per annotation. After this

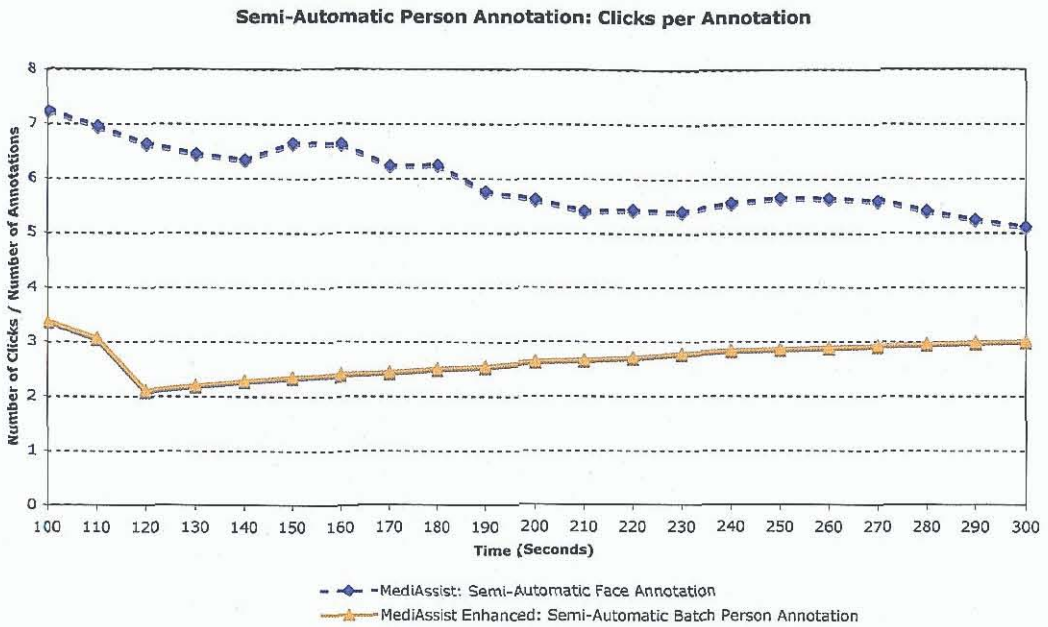


Figure 7.3: Number of user mouse clicks per annotation for interactive person annotation.

it begins to rise gradually, caused by the fact that there tends to be less accurate results coming from the batch annotation suggestions at this stage. For *MediAssist: Semi-Automatic Person Annotation*, the gradual trend is a downward one, reflecting the fact that the user tends to spend more mouse clicks formulating search queries and browsing towards the start of the task, and as time passes the user spends more mouse clicks annotating photos and less time browsing and searching. At 100 seconds and 110 seconds the difference between *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* and *MediAssist: Semi-Automatic Person Annotation* is statistically significant, and at all points after 120 seconds the difference is highly significant.

In addition to increased efficiency, *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* also has the advantage that, by encouraging users to reject face suggestions for a particular person and to label other suggestions as 'unknown', a great deal of supplementary annotation is input to the system. For the average annotation task using this system, 135 person suggestions were rejected by the users and 10.7 faces were labelled as unknown. These extra annotations can be used

by the system, when it re-indexes the users' collections, to further improve the suggestions provided in future annotation suggestion iterations. For the *MediAssist. Semi-Automatic Person Annotation* system an average of only 0.1 faces per task were labelled as unknown, while rejecting candidate faces for a given name is not possible with this system. These additional annotations are important because they justify the assumption we made in Section 6.2.1 that, given an appropriate user interface, users can be encouraged to annotate unknown faces in their collection. Also, for the batch annotation suggestion techniques to work effectively, the system needs to receive feedback about incorrect suggestions because otherwise the system will keep suggesting these for a given target person, particularly if they occur in a similar context (e.g. in the same event) to known occurrences of that person.

The results in this section, particularly for annotation coverage, clearly show that for this particular task *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* provides a much more efficient annotation interface. Although in a real world situation a user may not always concentrate on one identity at a time as they annotate the people in their collections, we believe the magnitude of the difference between each system emphasises the efficiency of this approach, and if users are presented with a list of accurate annotation suggestions they would take some to to confirm them. In fact, if we had removed the restriction on only annotating one identity per annotation task then we believe that the difference between the two approaches would be emphasised even further. This is because, at the point when the accuracy of the batch annotations begins to deteriorate, the user could switch 'tasks' and begin to annotate a different person. The size of the data-sets and number of users available meant that this additional experiment was beyond the scope of this work, but we believe that the magnitude of the differences between the two approaches suggests that the enhanced approach is superior for general person annotation scenarios in addition to being superior for the specific person annotation task evaluated here.

Also, as noted in Section 7.2, this prototype system does not fully update the

Statement	Median (a)	Median (b)
'The system is easy to use'	5	6
'The system's response time is quick enough'	5	5
'The system allows efficient annotation of people in personal photo collections'	3	6
'The annotation suggestions provided by the system were accurate/useful'	5	6
'The system is satisfying to use'	5	6
'I would use the annotation tools from this system in my day to day photo management if they were available'	4	6

Table 7.2: Median responses to a number of statements about each system on a 7-point Likert scale, with a score of 0 corresponding to 'Strongly Disagree' and a score of 6 corresponding to 'Strongly Agree'. (a) MediAssist: Semi-Automatic Person Annotation (b) MediAssist Enhanced: Semi-Automatic Batch Person Annotation.

index after each annotation, meaning that the accuracy of the results from the batch annotation system deteriorates more quickly than they would have if a full re-indexing was taking place on the fly. By developing techniques to allow efficient updating of the person retrieval index after each annotation, the batch annotation suggestions offered by this interactive system could be further improved, which in turn would further improve the efficiency of the interactive batch annotation system.

7.4.2 User Satisfaction: Qualitative Results

Table 7.2 shows median user responses to several statements about each system on a 7-point Likert scale, with 0 representing 'Strongly Disagree' and 6 representing 'Strongly Agree'. The full Likert scale was as follows: 'Strongly Disagree', 'Quite Disagree', 'Disagree', 'Neutral', 'Agree', 'Quite Agree' and 'Strongly Agree'. For both systems the user responses were generally positive, and for all statements except 'The system allows efficient annotation of people in personal photo collections' the median response for each system represents agreement. For this statement *MediAssist: Semi-Automatic Person Annotation* has a median response of 3, representing 'neutral'.

For all the statements shown, except 'the system's response time is quick enough', the responses for *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* are significantly better than those for *MediAssist: Semi-Automatic Person Annotation*. The users rated *MediAssist Enhanced: Semi-Automatic Batch Person Annotation* particularly highly with a median score of 6, the maximum score, for all other questions. The users had little difficultly learning how to use the annotation tools in the system, reflected in their agreement that the system is 'easy to use' and 'satisfying to use'. Perhaps most important is the high level of agreement with the statement 'I would use the annotation tools from this system in my day to day photo management if they were available'. 5 of the 7 users indicate that they 'strongly agree' with this statement, the other 2 users responding with 'quite agree', the next strongest level of agreement. This shows that, for the group of users in these experiments, the batch annotation tools were very useful and they would be very willing to use them to annotate their real-world personal photo collections.

The high scores for the *MediAssist: Semi-Automatic Person Annotation* system indicate that, although they were less satisfied with this system than the batch annotation system, users found the classification results useful for person annotation. This can be seen in the median score of 5, indicating 'quite agree', for the statement 'the annotation suggestions provided by the system were accurate/useful'.

There were occasions, however, when the batch annotation pop-up window proved irritating to the user, and we believe this aspect of the system could be improved. We believe that placing the suggestions on the bottom or to the side of the screen where they would be less intrusive would create a better user experience, as the user would be free to ignore the suggestions if they wished, rather than be forced to close a pop-up window.

7.5 Summary

We have evaluated an interactive semi-automatic batch person annotation system, comparing it to a baseline semi-automatic face annotation system in a controlled experiment, requiring 7 users to complete a number of annotation tasks. The new system significantly outperforms the baseline system in terms of both quantitative and qualitative measures. The average annotation coverage from the proposed system after 60 seconds of user interaction is over 10 times that of the baseline system. After the maximum allowed 5 minutes of user interaction, the proposed system achieves an annotation coverage of 89.5%, over twice that of the baseline system. The proposed system also outperforms the baseline system in terms of mouse clicks, requiring fewer clicks per annotation and therefore less user effort. The system also elicits extra annotation from the user in terms of annotation of unknown people and rejecting incorrect annotation suggestions for known people.

This strong performance of the system indicates that the users had no problems in learning how to use this interface, and indeed qualitative evaluation from users gave a median response of 'strongly agree' to the statements that the system is 'easy to use' and 'satisfying to use'. Indeed, the users rated the proposed system very highly for all qualitative measures, and indicated a strong willingness to use the annotation tools provided by this system in their real world photo management systems.

In summary, the proposed batch person annotation system is very efficient for annotation of people in personal photo collections, the interface was easily learned by users and the system would be welcomed by users if available in their real world photo management systems.

Chapter 8

Conclusions and Summary

In this thesis we have described the MediAssist system, a context-aware photo management system for location-stamped personal photos and we have developed and tested a set of hypotheses related to personal photo management, using this real system implementation as a testbed The system enhances available metadata using content-based and context-based analysis, and facilitates searching, browsing and semi-automatic annotation of personal photos. We proposed an approach to semi-automatic person-annotation in personal photo collections that facilitates the annotation of people in personal photo collections in a batch manner.

In developing the system functionality and scaling it to manage personal photo collections, we have proposed person classification and person retrieval techniques based on the use of both content and context, to facilitate this annotation by providing annotation suggestions to users. These approaches are based on analysis of the context of photo capture, principally the spatial and temporal context, along with analysis of the image content of the photo. The proposed context-based language model approach is an extension of the work of Naaman et al (2005). Another context-based approach which we test uses nearest neighbour classification and retrieval based on temporal and spatial proximity. Our proposed content-based approaches to person classification and retrieval use face recognition and body patch features, in addition to face colour and image colour, two novel features for the

task of person classification and retrieval We have implemented these proposed approaches, and the interactive tools for semi-automatic batch person-annotation, and integrated them into the MediAssist interactive system.

We have conducted a thorough evaluation of these proposed approaches, comparing the relative performance of approaches based on context, then based on content and finally based on a combination of the two, using the real photo collections of 9 users of the MediAssist sysem. We used a richly annotated training set for this task to enable us to fully explore the relative performance of context-based approaches, and the performance of content-based approaches that rely on a local temporal proximity (i e. within an event). Finally, we conducted a user study with 7 of these users, annotating faces in their personal photos using a semi-automatic batch person-annotation interface.

8.1 Conclusions

In Chapter 1 we stated two related hypotheses about person annotation which we set out to prove in this thesis. They are:

Hypothesis 1 (H1)

> *Semi-automatic person-annotation techniques for context-aware personal photo collections can be developed and can perform effectively*

Hypothesis 2 (H2)

> *Person classification and retrieval techniques, which use context-based analysis in addition to content-based analysis, can be developed and can perform effectively*

We believe that the evaluation reported in Chapter 6 proves *Hypothesis 2* by showing that context-based and content-based analysis can perform effectively and

that combined context-based and content-based approaches to person classification and retrieval work better than either in isolation.

We believe that the evaluation in Chapter 7 proves *Hypothesis 1* by showing, in a small pilot study, that real users, annotating their own personal photo collections, can easily and effectively use the proposed annotation tools, and these users show a strong willingness to use these tools in their day to day photo management.

Our approach to semi-automatic batch annotation has the advantage that it integrates the annotation process into the user's natural interaction with the system, providing annotation suggestions at appropriate moments as the user browses or searches their collection. By performing most of their analysis within the local context of known faces, our event-filtered person classification and retrieval approaches reduce the amount of analysis performed by the system, as candidate faces only need to be compared to known faces in the same event. The system allows the user to quickly annotate multiple occurrences of a known person within an event and this is confirmed through quantitative evaluation on users' photo collections. The approach has the drawback, however, that the annotation suggestions in events with little or no previous annotation are unlikely to be accurate.

In the sections below we will outline our individual conclusions, based on the empirical studies carried out in this thesis, about the performance of context-based and content-based, as well as combined context-based and content-based approaches to automatic person classification and retrieval in personal photo collections. We will then draw some conclusions based on a user study with the prototype implementation of the proposed system

8.1.1 Context-based approaches to Person Classification and Retrieval.

In evaluating the performance of different approaches to person classification and retrieval, we found that smoothing language models, and hierarchical language mod-

els in particular, significantly outperform maximum likelihood estimate (MLE) approaches For person retrieval, however, smoothing is not quite as effective because all faces to be ranked are smoothed with the same background probability for the query person, although hierarchical smoothing overcomes this problem.

Nearest neighbour approaches perform very well for temporal proximity, approaching the performance of hierarchical language models for this context-based feature. Nearest neighbour approaches perform poorly for spatial proximity, however, when compared with hierarchical language models. This means that, for location-based context, information about relative frequency of occurrence is needed to achieve the best performance, while for temporal proximity ranking by distance can also perform well. Nearest neighbour approaches also have the advantage that they can be expected to give reasonable performance even with very sparse annotations, for example if only one face in the entire collection event is annotated, other faces could be ranked based on temporal proximity to that face.

The best-performing context-based feature is temporal proximity, followed by spatial proximity. Combining temporal proximity and spatial proximity gives little improvement over temporal proximity alone. We believe, however, that in a more sparsely annotated collection, where local temporal information is not so often available, location information would become more useful and the combination of spatial and temporal proximity would show a larger improvement.

Combining context-based approaches gives significant improvements of each individual approach. For combining hierarchically smoothed language model approaches it is necessary to optimise mixing weights to achieve this improvement Combined nearest neighbour approaches show an improvement in performance without the use of weights, although the improvement is small due to the relatively poor performance of nearest neighbour spatial proximity.

8.1.2 Content-based approaches to Person Classification and Retrieval.

All of the content-based features which we evaluated in our suite of experiments proved useful for person classification and retrieval For classification, the image colour feature is surprisingly effective in terms of *5-hit rate*, although it is the worst performing content-based feature for the more strict evaluation measure *1-hit rate*. For *1-hit rate* body patch is the best-performing content-based feature. Face colour is the worst performing content-based feature. Combining these features for classification is particularly effective, significantly improving performance. This shows that, for person classification, if we have annotation information in the local context of candidate faces, then the use of colour-based image features as alternatives to face recognition can help classification accuracy. In more heterogeneous collections, in terms of the skin colour of the people present, it is possible that the performance of the face colour feature might improve further.

The best-performing feature for person retrieval is body patch. Since retrieval performance is most affected by candidate faces towards the top of the ranking, this shows that when the body patch feature is confident it is more reliable than the other features. Combining content-based features for retrieval gives a strong improvement. For the photo-level retrieval scenario this improvement is achieved by weighted and unweighted approaches, while for the face region-level scenario there is again a large improvement with the use of weights, but only a slight improvement from unweighted combination.

We found that event-based filtering improves performance for all individual content-based approaches. For body patch, face colour and image colour this confirms the assumption that these features are most useful within a local temporal proximity. For face recognition, which does not make any assumption about local context, this improvement suggests that the feature benefits from the smaller search space of an event. Combined content-based event filtering in the person classification

scenario outperforms user-filtering for *1-hit rate* and *2-hit rate*, but is outperformed by user-filtering for *4-hit rate* and *5-hit rate*. This emphasises the drawback of event-filtering, namely that it cannot suggest a name not already known to be in an event, causing it to suffer for this coarser evaluation measure. In the person retrieval scenario, event-filtering always outperforms user-filtering.

8.1.3 Combining Content and Context for Person Classification and Retrieval.

The results for both person classification and person retrieval show that performance can be improved if we combine context-based analysis with content-based analysis. For person classification and for photo-level retrieval this improvement is statistically significant for both weighted and unweighted approaches. For face region-level person retrieval, this improvement is significant for *MAP* and *P30* but not for *P10* and *P20* The use of weights for this task gives a larger improvement, giving encouragement that context-based features can offer real improvement for person retrieval.

We believe that these results show there is a place for context-based approaches to person classification and retrieval for semi-automatic person annotation. Content-based approaches will continue to improve but will always struggle in difficult cases, for example rotated faces, slightly occluded and non-frontal faces. The use of context-based approaches, which do not need to know anything about the content of an image, can step in and, as has been shown in this thesis, improve the performance in cases where standard approaches struggle. Context-based approaches, along with the image colour feature, also have the advantage that they can be used to suggest person names for photos even if it is not known that there is a face in the photo.

8.1.4 Face Region-level or Photo-level Annotation?

We also ask the question as to whether it is more efficient to provide annotations at the level of the photo and make use of photo-level features for retrieval, or is it more efficient to provide annotation at the level of the face where there can be more than one face in a photo, and make use of face-level features for retrieval. The performance of photo-level retrieval using photo-level features is similar to the performance of face region-level retrieval using region-level features. We believe, based on this, that it is more useful to annotate at the level of the face, because the quality of the annotation suggestions is equivalent, but the user is providing a more detailed annotation to the system.

8.1.5 Interactive User Evaluation.

Our user study confirmed that real users can work effectively with the proposed interface for semi-automatic batch person-annotation, and that the quality of the annotation suggestions, given a partially annotated collection, is good enough to facilitate effective interactive semi-automatic annotation. Results of a qualitative user survey showed a strong inclination from users to use these annotation tools for annotating their real world photo collections. We did not present users with a full implementation of the system which updates the person search engine after every annotation, however, and it would be interesting to evaluate how users would interact with the full system if they started using the system with no initial annotations.

8.2 Directions for Future Work

In this section we outline some possible directions for future work.

Automatic Annotation. It cases where all content-based and context-based features show high confidence, it is worth investigating if we can automatically annotate the face without prompting the user for confirmation. This can save the user anno-

tation effort and, if performed reliably, will add valuable training data to the person classification and retrieval models. We believe it is reasonable to assume that, if the temporal distance between the photo capture for two photos is low, and face recognition, body patch, face colour and image colour features are all highly similar, then automatic annotation could be possible. For near-duplicate photos, a face position feature would also be useful for this.

Cluster-based Batch Annotation. An alternative approach to batch person annotation is to automatically create clusters of similar faces, allowing face clusters to be annotated in a batch manner. This is orthogonal to our approach, and such cluster-based annotation could be an alternative annotation strategy in our interface. Using the content-based and context-based features proposed in this thesis could improve the quality of the clusters for such an approach.

Active Learning. The person retrieval approaches proposed in this thesis work best within the same event, and so there is the disadvantage that the user will tend to only annotate the people in one event at a time using this interface. We suggest that, in addition to using the approaches proposed in this thesis for person retrieval, the suggestion list could be supplemented by a small number of faces from events that are *not* known to contain the target person. These supplementary suggestions could use context and the face recognition feature, for example, so they would return faces similar to the target person known to occur in locations familiar to the target person and during timeframes when the target person was known to be present, but in new events. These suggestions would not be as accurate as the main suggestions. If they are relevant and the user annotates them, however, they will add another event to the person model and prompt accurate event-based annotation suggestions from these new events at the next annotation iteration. This should accelerate the annotation process. This concept of intelligent querying to retrieve the most useful samples to label for the purpose of training a model is known in the machine learning community as *active learning* (Cohn et al., 1992).

Leveraging Annotations from Friends. In large online photo-sharing communities, a user is likely to have friends who use the system and who will have photos taken at the same event. In addition to the functionality of being able to annotate each others photos, the system could use annotations submitted by one user to suggest annotations for the photos of another user, if they are known to be friends and if the photos were known to be captured at the same time and place. Accordingly, the content-based and context-based approaches proposed in this thesis, some of which assume annotation in the same event, could be utilised when none of the user's photos from that event have been annotated. Alternatively, a friend's user collection could be used as a background model for smoothing approaches, particularly when the friends collection is more heavily annotated.

Automatic Learning of Weights. All of the weighted approaches presented in the evaluation in this thesis use biased, or 'oracle', weights optimised on the test set. This gives an upper limit on the performance of given approaches to weighted data fusion, and can be used to benchmark strategies that learn weights in an unbiased manner. We believe that using cross-validation techniques (Devijver and Kittler, 1982) within the set of annotated faces will enable us to learn weights that improve on the performance of unweighted approaches, and approach the performance that can be achieved with the use of optimal weights.

Additional Sources of Context. Other sources of context can be explored for the tasks of person classification and retrieval. Davis et al. (2005b), for example, have previously proposed bluetooth co-presence information and photo-sharing information for this purpose. One can also imagine calendar entries with details of meetings and parties being used to suggest potential names in photos.

8.3 Summary

In this thesis we have described an approach to the semi-automatic annotation of people in personal photo collections based on semi-automatic batch person annotation. The approach is powered by context-based and content-based approaches to person classification and retrieval. The usefulness of this approach is in demonstrating that analysis of contextual information along with image content can be used to improve the accuracy of person classification and retrieval in personal photo collections.

We believe our results show that there is a place for context-based person classification and retrieval techniques alongside traditional techniques such as face recognition, even as these traditional techniques continue to improve. The presence of contextual information can reinforce content-based information, and can provide good annotation suggestions in situations where these content-based approaches fail.

We have also shown that the proposed semi-automatic batch person-annotation tools can be used effectively by real users, and that real users show a strong willingness to use these tools.

Bibliography

Anguelov, D., Lee, K., Gokturk, S. B., and Sumengen, B. (2007). Contextual Identity Recognition in Personal Photo Albums. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN.

Aris, A., Gemmell, J , and Lueder, R. (2004). Exploiting Location and Time for Photo Search and Storytelling in MyLifeBits. Technical report, Microsoft Research.

Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464.

Bederson, B. B. (2001). Photomesa: a Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. In *UIST '01: Proceedings of the 14th annual ACM symposium on User Interface Software and Technology*, pages 71–80, Orlando, FL.

Bissol, S., Mulhem, P., and Chiaramella, Y. (2003). Dynamic Learning of Indexing Concept for Home Image Retrieval. In *International Workshop on Content-Based Multimedia Indexing*, Rennes, France.

Boll, S., Sandhaus, P., Scherp, A , and Thieme, S. (2007). MetaXa - Context- and Content-driven Metadata Enhancement for Personal Photo Albums. In *International Multimedia Modeling Conference (MMM 2007)*, pages 332–343, Singapore.

Boutell, M. R. and Luo, J. (2005). Beyond Pixels: Exploiting Camera Metadata for Photo Classification. *Pattern Recognition*, 38(6):935–946.

Burges, C. J. C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chen, G. and Kotz, D. (2000) A Survey of Context-Aware Mobile Computing Research Technical report, Dartmouth College, Hanover, NH.

Chen, J., Tan, T., and Mulhem, P (2003a). Using Speech Annotation for Home Digital Image Indexing and Retrieval. In *International Workshop on Content-Based Multimedia Indexing*, pages 195–200, Rennes, France.

Chen, L., Hu, B.-G., Zhang, L., Li, M., and Zhang, H. (2003b) Face Annotation for Family Photo Album Management *International Journal of Image and Graphics*, 3(1):81–94.

Cohn, D., Atlas, L., and Ladner, R (1992). Improving Generalization with Active Learning. *Machine Learning*, 15(2):201–221.

Cooper, M. D., Foote, J., Girgensohn, A , Cooper, M. D., Foote, J., Cooper, M. D , Foote, J., Girgensohn, A., and Wilcox, L. (2003). Temporal Event Clustering for Digital Photo Collections. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 364–373, Berkeley, CA.

Cooray, S., O'Connor, N., Gurrin, C., Jones, G., O'Hare, N., and Smeaton., A. F. (2006). Identifying Person Re-occurrences for Personal Photo Management Applications. In *International Conference on Visual Information Engineering*, pages 144–149, Bangalore, India.

Cui, J., Wen, F., Xiao, R., Tian, Y., and Tang, X. (2007). EasyAlbum: an Interactive Photo Annotation Aystem Based on Face Clustering and Re-ranking. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 367–376, San Jose, CA.

Davis, M., House, N. V., Towle, J., King, S , Ahern, S., Burgener, C , Perkel, D , Finn, M., Viswanathan, V., and Rothenbery, M. (2005a). MMM2. Mobile Media

Metadata for Media Sharing. In *Conference on Human Factors in Computing Systems*, pages 1335–1338, Portland, OR.

Davis, M., King, S , Good, N., and Sarvas, R. (2004). From Context to Content: Leveraging Context to Infer Media Metadata. In *MULTIMEDIA '04· Proceedings of the 12th annual ACM international conference on Multimedia*, pages 188–195, New York, NY.

Davis, M., Smith, M., Canny, J., Good, N., King, S., and Janakiraman, R. (2005b). Towards Context-Aware Face Recognition. In *MULTIMEDIA '05· Proceedings of the 13th annual ACM international conference on Multimedia*, pages 483–486, Singapore.

Davis, M., Smith, M., Stentiford, F., Bambidele, A., Canny, J., Good, N., King, S., and Janakiraman, R. (2006). Using Context and Similarity for Face and Location Identification. In *Proceedings of the IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA

Devijver, P. A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach.* Prentice Hall, London.

Dey, A. K (2001). Understanding and Using Context. *Personal Ubiquitous Computing*, 5(1):4–7.

Dimitrova, N. (2004). Context and Memory in Multimedia Content Analysis. *IEEE MultiMedia*, 11(3):7–11

Doan, K., Plaisant, C., and Shneiderman, B. (1996). Query Previews for Networked Information Services. In *Advances in Digital Libraries Conference*, pages 120–129, Los Alamitos, CA.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (Second Edition).* Wiley-Interscience, New York, NY.

Everingham, M , Sivic, J., and Zisserman, A. (2006) Hello! My name is Buffy - Automatic Naming of Characters in TV Video. In *Proceedings of the British Machine Vision Conference*, pages 899 -908, Edinburgh, UK.

Fisher, R A. (1922). On the Mathematical Foundations of Theoretical Statistics *Philosophical Transactions of the Royal Society*, 222:309–368.

Fox, E. A. and Shaw, J. A (1994). Combination of Multiple Searches. In *Proceedings of the Third Text REtreival Conference (TREC-1994)*, pages 243–252, Gaithersburg, MD.

Gargi, U. (2003). Modeling and Clustering Photo Capture Streams. In *MIR '03: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 47–54, Berkeley, California

Gemmell, J., Bell, G., and Lueder, R. (2006). MyLifeBits: A Personal Database for Everything. *Communications of the ACM*, 49(1):89–95.

Gemmell, J., Bell, G., Lueder, R., Drucker, S., and Wong, C. (2002). MyLifeBits. Fulfilling the Memex Vision. In *MULTIMEDIA '02. Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238, Juan-les-Pins, France.

Gemmell, J., Lueder, R., and Bell, G (2003a). Living with a Lifetime Store. In *ATR Workshop on Ubiquitous Experience Media*, pages pp.69–76, Kyoto, Japan

Gemmell, J., Lueder, R., and Bell, G. (2003b). The MyLifeBits Lifetime Store. In *ETP '03· Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*, pages 80–83, Berkeley, CA.

Girgensohn, A , Adcock, J., and Wilcox, L. (2004a). Leveraging Face Recognition Technology to Find and Organize Photos. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 99–106, New York, NY.

Girgensohn, A., Adcock, J., and Wilcox, L (2004b). Organizing Photos of People. In *Seventeenth Annual ACM Symposium on User Interface Software and Technology*, pages 37–38, Santa Fe, NM.

Graham, A., Garcia-Molina, H., Paepcke, A., and Winograd, T. (2002). Time as Essence for Photo Browsing Through Personal Digital Libraries. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 326–335, Portland, OR.

Gurrin, C., Jones, G. J. F., Lee, H , O'Hare, N., Smeaton, A. F., and Murphy, N. (2005). Mobile Access to Personal Digital Photograph Archives In *MobileHCI '05: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & services*, pages 311–314, Salzburg, Austria.

Jaffré, G. and Joly, P (2004). Costume. a New Feature for Automatic Video Content Indexing. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pages 314–325, Avignon, France.

JEITA (2002). *Exchangeable Image File Format for Digital Still Cameras. EXIF Version 2.2.* Japan Electronics and Information Technology Industries Association, http://www.exif.org/.

Kempthorne, O. and Doerfler, T. E. (1969). The Behaviour of Some Significance Tests Under Experimental Randomization. *Biometrika*, 56(2).231–248.

Kuchinsky, A , Pering, C., Creech, M. L., Freeze, D., Serra, B., and Gwizdka, J. (2001). FotoFile: A Consumer Multimedia Organization and Retrieval System. In *CHI '99: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 496–503, Pittsburgh, PA.

Lafferty, J and Zhai, C. (2001). Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 111–119, New Orleans, LA.

Larsen, S. F., Thompson, C. P., and Hansen, T. (1995). Time in Autobiographical Memory. In Rubin, D. C., editor, *Remembering Our Past: Studies in Autobiographical Memory*, pages 129–156, Cambridge, UK. Cambridge University Press.

Lee, J. H. (1997) Analyses of Multiple Evidence Combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pages 267–276, Philadelphia, PA.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, (140):1–55.

Lim, J.-H., Tian, Q , and Mulhem, P. (2003). Home Photo Content Modeling for Personalized Event-Based Retrieval. *IEEE Multimedia, Special Issue on Multimedia Content Modeling and Personalization*, 10(4):28–37.

Liu, C. (2003). A Bayesian Discriminating Features Method for Face Detection. *Pattern Analysis and Machine Intelligence*, 25(6):725–740.

Longley, P. A., Goodchild, M. F , Maguire, D. J., and Rhind, D. W. (2001). *Geographic Information Systems and Science*. John Wiley & Sons, Chichester, UK

Loui, A. C. and Savakis, A. E (2000). Automatic Image Event Segmentation and Quality Screening for Albuming Applications. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1125–1128, New York, NY.

Malobabic, J., LeBorgne, H., Murphy, N., and O'Connor, N. E. (2005). Detecting the Presence of Large Buildings in Natural Images. In *Proceedings of CBMI 2005 - 4th International Workshop on Content-Based Multimedia Indexing*, Riga, Latvia.

Manjunath, B S., Ohm, J R., Vasudevan, V. V., and Yamada, A. (2001). Color and Texture Tescriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715.

Manning, C D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Mc Donald, K (2005). *Discrete Language Models for Video Retrieval.* PhD thesis, Dublin City University, Dublin, Ireland.

Mc Donald, K. and Smeaton, A. F. (2005). A Comparison of Score, Rank and Probability-based Fusion Methods for Video Shot Retrieval. In *CIVR 2005 - International Conference on Image and Video Retrieval*, pages 61–70, Singapore.

Mills, T. J., Pye, D., Sinclair, D., and Wood, K. (2000). Shoebox: A Digital Photo Management System Technical report, AT&T Laboratories, Cambridge, UK

Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., and Paepcke, A. (2004a). Context Data in Geo-Referenced Digital Photo Collections. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 196–203, New York, USA.

Naaman, M., Harada, S., Wang, Q., and Paepcke, A. (2004b). Adventures in Space and Time: Browsing Personal Collections of Geo-Referenced Digital Photographs. Technical report, Stanford University

Naaman, M., Paepcke, A., and Garcia-Molina, H. (2003). From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates. In *10th International Conference on Cooperative Information Systems (COOPIS 2003)*, Sicily, Italy.

Naaman, M., Song, Y. J., and Paepcke, A. (2004c). Automatic Organization for Digital Photographs with Geographic Coordinates. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 53–62, Tuscon, AZ.

Naaman, M., Yeh, R. B., Garcia-Molina, H., and Paepcke, A. (2005) Leveraging Context to Resolve Identity in Photo Albums. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 178–187, Denver, CO.

O'Hare, N., Gurrin, C , Jones, G. J F., Lee, H., O'Connor, N. E., and Smeaton, A. F. (2007). Using Text Search for Personal Photo Collections with the MediAssist System. In *SAC 2007 - 22nd Annual ACM Symposium on Applied Computing*, Seoul, Korea.

O'Hare, N., Gurrin, C., Jones, G. J. F., and Smeaton, A. F. (2005a). Combination of Content Analysis and Context Features for Digital Photograph Retrieval. In *2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pages 323–328, London, UK.

O'Hare, N., Gurrin, C., Lee, H , Murphy, N , Smeaton, A. F., and Jones., G. J F (2005b). Digital Photos: Where and When? In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 261–262, Singapore.

O'Hare, N., Lee, H., Cooray, S., Gurrin, C , Jones, G., Malobabic, J., O'Connor, N., Smeaton, A. F., and Uscilowski, B. (2006) MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. In *CIVR2006 - 5th International Conference on Image and Video Retrieval*, pages 529–532, Tempe, AZ

Pigeau, A. and Gelgon, M. (2003) Spatio-temporal Organization of one's Personal Image Collection with Model-based ICL-Clustering. In *3rd International Workshop on Content-Based Multimedia Indexing (CBMI'2003)*, pages 111–118, Rennes, France.

Pigeau, A. and Gelgon, M. (2004). Incremental Statistical Geo-Temporal Structuring of a Personal Camera Phone Image Collection. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, pages 878–881, Washington, DC.

Pigeau, A. and Gelgon, M. (2005). Building and Tracking Hierarchical Geographical & Temporal Partitions for Image Collection Management on Mobile Devices In

MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, pages 141–150, Singapore.

Platt, J. C. (2000). AutoAlbum. Clustering Digital Photographs using Probabilistic Model Merging. In *CBAIVL '00: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00)*, pages 96–100, Washington, DC.

Ponte, J. M. and Croft, W B. (1998). A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 275–281, Melbourne, Australia.

Ray, S. F. (2000). *Camera Exposure Determination*, pages 310–335 Focal Press, Oxford, UK.

Robertson, S E., Walker, S., , Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, MD.

Rodden, K. and Wood, K. R. (2003). How do People Manage their Digital Photographs? In *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 409–416, Fort Lauderdale, FL.

Seidelmann, P. K., editor (1992). *Explanatory Supplement to the Astronomical Almanac*. University Science Books, Mill Valley, California.

Shen, C., Lesh, N., and Vernier, F. (2003). Personal Digital Historian: Story Sharing Around the Table. *ACM Interactions*, 102(22):15–22.

Shen, C., Lesh, N. B., Vernier, F., Forlines, C , and Frost, J. (2002). Sharing and Building Digital Group Histories. In *CSCW '02· Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 324–333, New Orleans, LA.

Shen, C., Moghaddam, B., Lesh, N., Beardsley, P., and Bardsley, R. S. (2001). Personal Digital Historian: User Interface Design. In *CHI '01: extended abstracts on Human Factors in Computing Systems*, Seattle, WA

Shneiderman, B. and Kang, H. (2000). Direct Annotation. A Drag-and-Drop Strategy for Labeling Photos. In *IV '00: Proceedings of the International Conference on Information Visualisation*, page 88, Washington, DC.

Sivic, J., Zitnick, C., and Szeliski, R. (2006). Finding People in Repeated Shots of the Same Scene. In *Proceedings of the British Machine Vision Conference*, pages 909–918, Edinburgh, UK

Smeulders, A., Worring, M , Santini, S., Gupta, A., and Jain, A. (2000). Content-based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Song, Y., Song, Y., and Leung, T. (2006). Context-Aided Human Recognition - Clustering. In *9th European Conference on Computer Vision (ECCV2006)*, pages 382–395, Graz, Austria.

Spinellis, D. D. (2003). Position-Annotated Photographs: A Geotemporal Web. *IEEE Pervasive Computing*, 2(2):72–79.

Srihari, R. K., Zhang, Z., Srikanth, M., Han, B., Rao, A , and Wu, X. (1999) Multimedia Indexing and Retrieval of Voice-Annotated Consumer Photos. In *Proceedings of Multimedia Indexing and Retrieval Workshop, SIGIR'99*, Berkeley, CA.

Suh, B. and Bederson, B. B. (2007). Semi-automatic Photo Annotation Strategies using Event Based Clustering and Clothing Based Person Recognition *Interacting with Computers*, 19(4):524–544.

Sun, Y., Zhang, H., Zhang, L., and Li, M. (2002). MyPhotos. a System for Home

Photo Management and Processing. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 81–82 ACM Press.

Talmi, K. and Liu, J. (1999). Eye and Gaze Tracking for Visually Controlled Interactive Stereoscopic Displays *Image Communication*, 14(10):99–810.

Toyama, K., Logan, R., and Roseway, A (2003). Geographic Location Tags on Digital Images. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 156–166, New York, NY.

Tuffield, M. M , Harris, S , David, Chakravarthy, A., Brewster, C , Gibbins, N., Hara, K. O , Ciravegna, F., Sleeman, D., Shadbolt, N R., and Wilks, Y. (2006). Image Annotation with Photocopain. In *Proceedings of the 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM'06)*, pages 1–14, Edinburgh, UK.

Wagenaar, W. A. (1986). My Memory: A Study of Autobiographic Memory over Six Years. *Cognitive Psychology*, 18:225–252.

Wenyin, L , Sun, Y., and Zhang, H. (2000). MiAlbum - a System for Home Photo Management using the Semi-Automatic Image Annotation Approach. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 479–480, Marina del Rey, CA.

Westerveld, T., de Vries, A., van Ballegooij, A., de Jong, F., and Hiemstra, D. (2003a). A Probabilistic Multimedia Retrieval Model and its Evaluation. *EURASIP Journal on Applied Signal Processing*, 2:186–198.

Westerveld, T., de Vries, A. P., Westerveld, A. T., de Vries, A P., and van Ballegooij, A. R. (2003b). CWI at the TREC-2002 Video Track. In *The Eleventh Text REtrieval Conference (TREC-2002)*, pages 207–216, Gaithersburg, MD.

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics*, 1(6):80–83.

Wilhelm, A., Takhteyev, Y., Sarvas, R., House, N. V., and Davis, M. (2004). Photo Annotation on a Camera Phone In *CHI '04: extended abstracts on Human factors in computing systems*, pages 1403–1406, Vienna, Austria.

Wolf, L. and Bileschi, S. (2006). A Critical View of Context. *International Journal of Computer Vision*, 69(2):251–261.

Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

Zhai, C. and Lafferty, J. (2001) A Study of Smoothing Methods for Language Models Applied to Information Retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 334–342, New York, NY.

Zhang, L., Chen, L., Li, M., and Zhang, H. (2003). Automated Annotation of Human Faces in Family Albums. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 355–358, Berkeley, CA.

Zhang, L., Hu, Y , Li, M , Ma, W , and Zhang, H (2004). Efficient Propagation for Face Annotation in Family Albums. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 716–723, New York, NY.

Zhao, M., Teo, Y., Zhao, M., Zhao, M., Teo, Y., Liu, S., Chua, T.-S., and Jain, R. (2006). Automatic Person Annotation of Family Photo Album. In *5th International Conference on Image and Video Retrieval (CIVR 2006)*, pages 163–172, Tempe, AZ.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face Recognition: A Literature Survey. *ACM Computing Surveys*, 35(4):399–458.