THE ACETOOLBOX: LOW-LEVEL AUDIOVISUAL FEATURE EXTRACTION FOR RETRIEVAL AND CLASSIFICATION

Noel E. O'Connor, Edward Cooke, Herve Le Borgne, Michael Blighe, Tomasz Adamek

Centre for Digital Video Processing, Dublin City University, Ireland
{oconnorn,ejcooke}@eeng.dcu.ie

ABSTRACT

In this paper we present an overview of a software platform that has been developed within the aceMedia project, termed the *aceToolbox*, that provides global and local low-level feature extraction from audio-visual content. The toolbox is based on the MPEG-7 eXperimental Model (XM), with extensions to provide descriptor extraction from arbitrarily shaped image segments, thereby supporting local descriptors reflecting real image content. We describe the architecture of the toolbox as well as providing an overview of the descriptors supported to date. We also briefly describe the segmentation algorithm provided. We then demonstrate the usefulness of the toolbox in the context of two different content processing scenarios: similarity-based retrieval in large collections and scene-level classification of still images.

1. INTRODUCTION

Image and video indexing and retrieval continues to be an extremely active area within the broader multimedia research community. Interest is motivated by the very real requirement for efficient techniques for indexing large archives of audiovisual content in ways that facilitate subsequent user-centric browsing, searching and retrieval. This requirement is a by-product of the decreasing cost of storage and the now ubiquitous nature of capture devices, the result of which is that content repositories, either in the commercial domain (e.g. broadcasters or content providers' repositories) or the personal archives many of us now maintain, are growing in number and size at virtually exponential rates.

It is generally acknowledged that providing truly efficient user-centric access to large content archives requires indexing of the content in terms of the real world semantics of what it represents. Furthermore, it is acknowledged that real progress in addressing this challenging task requires key advances in many complementary research areas, including scalable coding of both audiovisual content and its metadata, database technology, user interface design, user context modeling, knowledge representation and modeling, automatic and semi-automatic annotation tools, and indexing and retrieval algorithms. The EU FP6 aceMedia integrated project is tackling many of these issues [1]. A key effort within the project is to link audio-visual analysis with ontological reasoning in order to extract semantic information [2]. In this context, low-level pre-processing is necessary in order to extract descriptors that can be subsequently linked to the ontology and used in the reasoning process. Since, in addition to ontological-based reasoning, the project has other research activities that require low-level feature extraction (e.g. scalable coding of metadata [3], image retrieval including relevance feedback [4]) it was decided to develop a common platform for descriptor extraction that could be used throughout the project.

The requirements for the platform for low-level feature extraction were that it should:

- provide extraction of a subset of MPEG-7 [5] features;
- facilitate the integration of new descriptors;
- facilitate extraction of global (i.e. corresponding to the entire image) and local (i.e. corresponding to sections of the entire image only) descriptors;
- be platform independent and suitable for integration into larger scale demonstration systems.

In this paper we describe this platform, which we have termed the *aceToolbox*. The remainder of the paper is organised as follows: a general overview of the aceToolbox is provided in Section 2 including a description of the architecture, the descriptors supported and the image segmentation algorithm. In Section 3, we provide two examples of the potential usefulness of the toolbox in indexing image content. Finally, in Section 4 we describe our future plans for both the extension of the toolbox and its use in different scenarios.

This work was supported by the European Commission under contract FP6-001765 aceMedia (URL: http://www.acemedia.org).

2. ACETOOLBOX OVERVIEW

In this section, we present an overview of the structure of the toolbox and briefly describe the audio and visual feature extraction techniques currently supported. We also present a brief overview of the segmentation algorithm provided to support local region-based visual descriptor extraction. The aceToolbox currently supports extraction of 13 low-level audio-visual descriptors. The design is based on the architecture of the MPEG-7 eXperimentation Model (XM) [6], the official reference software of the ISO/IEC MPEG-7 standard [5]. In addition to a more "light weight" and modular design, a key advantage of the aceToolbox over the XM is the ability to define and process regions in the case of image input. Such image regions can be created either via a grid layout that partitions the input image into user defined square regions or a segmentation tool that partitions the image into arbitrary shaped image regions that reflect the structure of objects present in the scene.

2.1. Architecture

Figure 1 illustrates a schematic diagram of the main functionalities of the aceToolbox. The aceToolbox processes audiovisual input i.e. images, video sequences and audio. In Figure 1 the input is illustrated as an image, which might be a keyframe from a video sequence or a single photo. For images, local descriptors are supported by segmenting the image into either a user defined rectangular grid or arbitrarily-shaped regions. Following this, feature extraction (with user selection of the features to be used) is performed on either the whole image, each image block or each image region as appropriate. The current version of the aceToolbox supports the descriptors listed in Table 1 and described in Section 2.3. Visual descriptors are classified into four types: colour, texture, shape and motion (for video sequences). Currently, there is only a single video-based and audio descriptor supported, but this will be extended in the future. The output of the aceToolbox is an XML file for each specified descriptor, which for image input relates to either the entire image or separate areas within the image. An example of typical output is shown in Figure 2.The toolbox adopts a modular approach, whereby APIs are provided to ensure that the addition of new descriptors is relatively straightforward. The system has been successfully compiled and executed on both Windows-based and Linux-based platforms.

2.2. Image segmentation

As discussed in [7], while imperfect, segmentation is an essential step for object recognition which often leads to improved scene understanding. For this reason, the aceToolbox contains a robust and efficient method for segmentation of images into large regions that reflect the real world

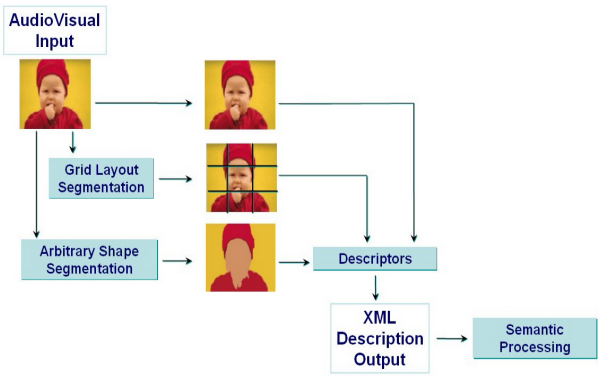

Fig. 1. Overview of the aceToolbox for image input

```
<?xml version='1.0' encoding='ISO-8859-1' ?>
<Mpeg7 xmlns = "http://www.mpeg7.org/2001/MPEG-7_Schema"
   xmlns:xsi = "http://www.w3.org/2000/10/XMLSchema-instance">
<DescriptionUnit xsi:type = "DescriptorCollectionType">
<Descriptor xsi:type = "HomogeneousTextureType">
<Region><RegionNumber>0</RegionNumber>
   <Average>139</Average>
   <StandardDeviation>157</StandardDeviation>
<Energy>176  191  ... </Energy>
<EnergyDeviation>175  193  ... </EnergyDeviation>
</Region>
    ...
</Descriptor>
...
</DescriptionUnit>
</Mpeg7>
```

Fig. 2. Example output of the aceToolbox

objects present in the scene. The segmentation algorithm, originally presented in [8], is based on an extension to the well known Recursive Shortest Spanning Tree (RSST) algorithm [9] that makes use of a new color model and the concept of *syntactic features* as originally proposed by Ferran and Casas in [10].

Many existing approaches to segmentation aim to create large regions using simple homogeneity criteria typically based only on color, texture or motion. The introduction to the segmentation process of syntactic features that represent geometric properties of regions and their spatial configurations provides a region merging criteria which helps prevent formation of regions spanning more than one semantic object. Such features provide an important source of information that could subsequently help to close the gap between low level features and a semantic interpretation of a scene. Whilst they do not provide a complete solution to reliably grouping regions into complex semantic objects, they can significantly improve the performance of automatic bottom-up segmentation techniques bringing them closer to the semantic level by allowing creation of large meaningful regions. Syntactic features can even be useful even in the case of supervised segmentation scenarios where they can facilitate more intuitive user interactions.

In our approach, syntactic features are integrated into the Recursive Shortest Spanning Tree (RSST) segmenta-

Colour	Texture	Shape	Motion	Audio
Dominant Colour Scalable Colour Colour Structure Colour Layout GoP/GoF	Homogeneous Texture Texture Browsing DCT Edge Histogram	Region Shape Contour Shape	Motion Activity	Fundamental Frequency

Table 1. Descriptors supported in the current version of the aceToolbox.

tion framework described in [11]. Syntactic features are extracted by structure analysis and are based on the shapes and spatial configuration of image regions. Region features are extended by a new color model and region boundary criterion. Subsequently, the merging order and merging criteria are re-defined based on the extracted syntactic features and the merging process continues until there are no remaining region pairs fulfilling the merging criteria. Features used include homogeneity, regularity (low complexity), compactness and inclusion and these control the merging order as well as the segmentation stopping criteria. We divide the merging process into stages so that the merging criteria can adapt as the segmentation progresses. In a given stage, links corresponding to pairs of neighboring regions, are allowed to compete for merging only if they fulfil the merging criteria defined for this stage. The merging order is controlled by a cost calculated as a weighted sum of costs related to color homogeneity and changes in shape complexity and adjacency. If a link does not fulfill the merging criteria specific to the current stage its cost is set to infinity. Each stage continues until there are no more links with finite cost associated with it. When a new stage starts, the merging criteria are redefined and costs for all links are recalculated.

Using a fixed set of parameter values, the algorithm works well on images from different collections, such as standard test images, professional databases, digital photographs and video keyframes from broadcast TV. Typically, a CIF image is segmented in less than 3 seconds on a Pentium III 600 MHz. Some illustrative segmentation results are presented in Figure 3. Further details and results, including a comparison with another popular segmentation technique for indexing and retrieval, are presented in [8].

2.3. Descriptors supported

All descriptors currently supported are MPEG-7 and are briefly described in the following. Further details can be found either in the MPEG-7 standard itself or in [5].

Dominant Colour (DC) takes an image or region as input and clusters the colours into a small number of representative colours. The number of dominant colors can vary from image to image, with a maximum of eight dominant colors being sufficient to represent an image. *Scalable Colour (SC)* generates a colour histogram in the hue-saturation-value (HSV) colour space that is encoded using a Haar transform thereby providing a scalable representation. *Colour Layout (CL)* is designed to capture the spatial distribution of colour in an image or region by clustering the image into 64 blocks and deriving the average colour of each block. These values are then transformed into a series of coefficients by performing an 8×8 DCT. *Colour Structure (CS)* represents an image or image region by both the colour distribution and the local spatial structure of the colour. It scans the image using an 8×8 pixel block and computes the number of blocks containing each colour and generates a Hue-Max-Min-Diff colour histogram. The *Group-of-Picture and Group-of-Frame (GoP/GoF)* colour descriptor is used for the joint representation of colour-based features for multiple images or multiple frames in a video segment. It extends the scalable colour descriptor and generates a colour histogram for a video segment or a group of pictures.

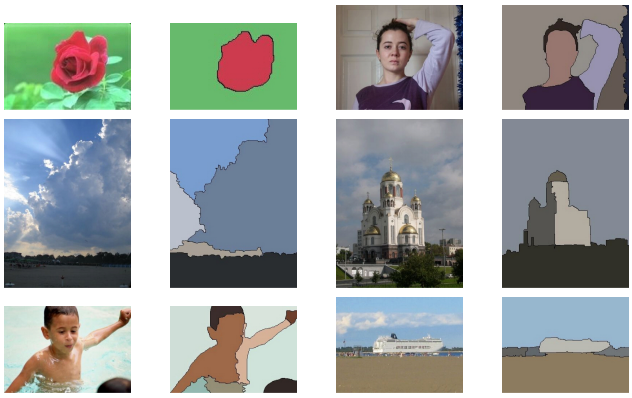

Fig. 3. Segmentation results for images from various sources. An image from the Corel data set and from our own test set are shown in row 1. The remaining images are from the aceMedia Content Repository

The *Edge Histogram (EH)* captures the spatial distribution of edges, which are identified using the Canny algorithm, by dividing the image into 16 non-overlapping blocks and then calculating 5 edge directions in each block. *Homogeneous Texture (HT)* describes directionality, coarseness

and regularity of patterns in images by partitioning the image's frequency domain into 30 channels and computing the energy and energy deviation of each channel and outputting the mean and standard deviation of the frequency coefficients. *Texture Browsing (TB)* specifies the perceptual characterization of texture similar to human characterization and is based on Gabor filtering. A multi-resolution decomposition of the image is used to determine the dominant orientations in the image using a Radon transform along the dominant orientations that determines the regularity of scale and direction of the texture. The *Discrete Cosine Transform (DCT)* is applied to 8×8 pixel blocks created using the Grid-Layout segmentation tool provided by the aceToolbox and the AC coefficients provide a simple measure of texture.

Region Shape (RS) is a descriptor that expresses the pixel distribution within a 2D object region using a complex 2D angular radial transformation (ART). The output is an array of normalised and quantized magnitudes of 35 AT coefficients. *Contour Shape (CSD)* is based on the MPEG-7 curvature scale-space (CSS) contour representation [5]. The output consists of the number of CSS peaks, the global eccentricity and circularity of the contour, the height of the most prominent CSS peak and the set of remaining peaks.

Motion Activity (MA) is a motion descriptor that captures the intensity of action of a video sequence. It is based on computing the standard deviation of motion vector magnitudes. The output consists of values representing the dominant direction, intensity, spatial distribution and a histogram representing the relative frequency of different levels of temporal activity.

Fundamental Frequency (FF) is the only audio descriptor currently supported. It provides information regarding musical pitch and the periodic content of speech signals and can also be used to give indications of melody. It is derived from the correlation between the signal and a lagged representation of the signal. It is computed using the local normalized auto-correlation function of the signal, taking its first maximum in order to estimate the local fundamental period [12].

3. APPLICATIONS OF THE ACETOOLBOX

3.1. Image retrieval in TRECVid 2004

The aceToolbox was used as the feature extraction mechanism for our participation in the TRECVid annual benchmarking activity [13]. A content-based information retrieval system, based on the Físchlár Digital Video System [14], was developed for participation in the interactive search task. Two versions of the system were developed: one supporting text- and image-based searching; the other supporting image-based searching only.

In order to support image-based searching the following features from the aceToolbox were automatically extracted from all TRECVid keyframes: Colour Layout, Scalable Colour, Edge Histogram and Homogenous Texture. All features were extracted for the entire image (region-based functionality was not supported at the time). The similarity between images was estimated by the L2 Minkowsky (Euclidean) distance for each of the features. At query time, the user could select which (or all) of the features were important for the specified topic via the system interface (see Figure 4) and each of these features were combined to produce a final feature ranked list. Relevance feedback was used to iteratively refine query results and converge on the user designated saved shots that were submitted for evaluation. User experiments with the two system variants were conducted with 16 volunteers who each processed 12 of the 24 TRECVid topics. Table 2 presents a comparison of the system performances across different types of queries for 4 users with 6 topics. The system using text combined with images provides a better average MAP and recall rate than the system using only images, however this is to be expected given the maturity of text retrieval research. Full system and experimental details can be found in [15].

Fig. 4. Screenshot of the system built for TRECVid 2004 that uses aceToolbox feature extraction. Features can be turned on/off by the user in the query panel in the upper left corner of the system interface.

3.2. Image classification

The aceToolbox was also used for image classification within multimedia databases. This classification of images is achieved by selecting an appropriate set of visual descriptors that capture the particular properties of a specific domain and the distinctive characteristics of each image class. One of the main challenges here is the identification of an appropriate combination or fusion method of the individual descriptors such that the overall classification result improves.

	Text-Image System		Image only System	
	MAP	Recall	MAP	Recall
User A	0.203	0.222	0.066	0.082
User B	0.190	0.181	0.094	0.097
User C	0.191	0.180	0.088	0.083
User D	0.133	0.139	0.074	0.077
Avg	0.179	0.181	0.081	0.085

Table 2. *Mean Average Precision* (MAP) and *Recall* results of interactive runs in TRECVid 2004.

For instance, local color descriptors and global color histograms are used in indoor/outdoor classification [16] to detect e.g. vegetation (green) or sea (blue). Edge direction histograms are employed for city/landscape classification [17] since city images typically contain horizontal and vertical edges. Motion descriptors are used for sports video shot classification [18].

In [19] the aceToolbox Colour Layout, Scalable Colour and Edge Histogram descriptors were used to extract features for a *beach*/*urban* scene classification probem. The aim of the work was to fuse several descriptors in order to improve the performance of several machine-learning classifiers. Fusion is necessary as descriptors would be otherwise incompatible and it would be inappropriate to directly include them in a distance metric. Three approaches are described in the paper: "merging" fusion combined with an SVM classifier, back-propagation fusion combined with a K-Nearest Neighbor classifier and a Fuzzy-ART neurofuzzy network. The image database used for the experiments is part of the aceMedia Content Repository[1], see Figure 5 for some examples of the images used. More specifically, it is part of the Personal Content Services database and consists of 767 high quality images divided in two classes: *beach* and *urban*. 60 images (40 from *beach* and 20 from *urban*) were used to train the neural network and the other 707 images (406 from *beach* and 301 from *urban*) were used to test the efficiency of the different classification approaches. 40 images from the *beach* category and 20 from *urban* were selected as the representative examples of the given classes for the considered database, then used as training dataset for the SVM classifier. The remaining 707 images of the database were used for testing. All possible combinations of the three visual descriptors were considered in all three approaches. Full details and experimental details can be found in [19].

4. CONCLUSION AND FUTURE WORK

Regarding the syntactic segmentation tool, in the future we plan to provide more fine-grained control of the process so that the result can be tailored to particular scene types. This will involve providing functionality to switch on/off substages of the segmentation process and modify the rules to be used in merging regions. In addition to more fine-grained control of the existing algorithm, we will also work on improving and extending the underlying algorithmic framework itself. To this end, work will target two directions: application of the segmentation process to video (e.g. the algorithm could be modified to avail of prior segmentation results so that temporal coherency of regions from one video frame to the next is ensured – see [20] for initial preliminary investigations) and the introduction of an alternative merging criterion motivated by models of the human visual system (e.g. saliency-based criteria).

Regarding additional descriptors, we plan to add the following MPEG-7 video-based features: Parametric Motion, Camera Motion and Motion Trajectory. Furthermore, we also plan to incorporate useful non-MPEG-7 descriptions e.g. Ridgelet-like transforms [21], Gabor filters, co-occurence matrices, region locator, mass and centre of mass. Furthermore some content structuring tools for video will also be investigated including shot-boundary detection and keyframe extraction and scene/sequence level analysis. Content structuring and motion-based descriptors will ensure a more comprehensive extension of the toolbox towards video content.

In terms of application of the toolbox, our preliminary experiments into image classification have shown that there is scope for significant research in the efficient fusion of low-level descriptors. For indexing and retrieval, we plan to use the toolbox in our participation in TRECVid 2005, using the segmentation functionality and associated region-based descriptors.

[1] http://driveacemedia.alinari.it/

5. REFERENCES

[1] I. Kompatsiaris, Y. Avrithis, P. Hobson, and M. Strintzis, "Integrating knowledge, semantics and content for user-centred intelligent media services: the acemedia project," in *WIAMIS 2004, Lisboa, Portugal*, April 2004.

[2] N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis, S. Staab, and S. Kollias, "An ontology infrastructure for multimedia reasoning," in *IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands*, July 2005.

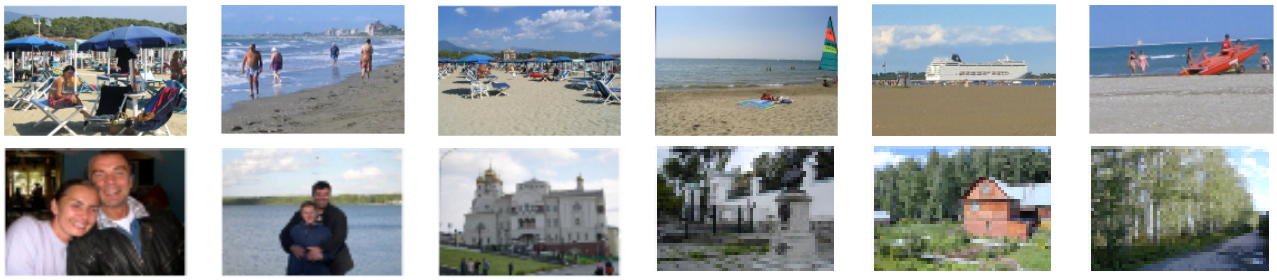

Fig. 5. Representative Images - First Row:Beach Images, Second Row: Urban Images

[3] D. Djordjevic, A. Dorado, E. Izquierdo, and W. Pedrycz, "Concept-oriented sample images selection." in *6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland*, April 2005.

[4] N. Sprljan, M. Mrak, G. Abhayaratne, and E. Izquierdo, "A scalable coding framework for efficient video adaptation." in *6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland*, April 2005.

[5] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MEPG-7: Multimedia Content Description Standard*. New York: Wiley, 2001.

[6] MPEG-7, "Visual experimentation model (xm) version 10.0," ISO/IEC/JTC1/SC29/WG11, N4062, 2001.

[7] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Color- and texture-based image segmentation using em and its application to image querying and classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1037, August 2002.

[8] T. Adamek, N. O'Connor, and N. Murphy, "Region-based segmentation of images using syntactic visual features," in *6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005), Montreux, Switzerland*, April 2005.

[9] S. H. Kwok, A. G. Constantinides, and W.-C. Siu, "An efficient recursive shortest spanning tree algorithm using linking properties," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 852–863, June 2004.

[10] C. F. Bennstrom and J. R. Casas, "Binary-partition-tree creation using a quasi-inclusion criterion," in *IEEE Computer Society Press, in the proceedings of the Eighth International Conference on Information Visualization (IV). London, UK*, 2004.

[11] A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image sequence analysis for emerging interactive multimedia services-the european cost 211 framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp. 802–813, Nov 1998.

[12] D. Gerhard, "Pitch extraction and fundamental frequency: History and current techniques," University of Regina, Tech. Rep. TR-CS 2003-6, November 2003.

[13] A. F. Smeaton, "Large scale evaluations of multimedia information retrieval: The trecvid experience," in *International Conference on Image and Video Retrieval (CIV 05), Singapore*, July 2005.

[14] A. Smeaton, C. Gurrin, H. Lee, K. M. Donald, N. Murphy, N. O'Connor, D. O'Sullivan, B. Smyth, and D. Wilson, "The físchlár-news-stories system: Personalised access to an archive of tv news," in *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval (RIAO 2004), Avignon, France*, April 2004.

[15] E. Cooke, P. Ferguson, G. Gaughan, C. Gurrin, G. Jones, H. Le Borgne, H. Lee, S. Marlow, K. McDonald, M. McHugh, N. Murphy, N. O'Connor, N. O'Hare, S. Rothwell, A. Smeaton, and P. Wilkins, "Trecvid 2004 experiments in dublin city university," in *TRECVID 2004 - Text REtrieval Conference TRECVID Workshop, Gaithersburg, Maryland, USA*, November 2004.

[16] M. Szummer and R. Picard, "Indoor-outdoor image classification," *IEEE international workship on content-based access of image and video databases*, 1998.

[17] A. Vailaya, A. Jain, and H. Zhang, "On image classification: City images vs landscapes," *Pattern Recognition*, vol. 31, pp. 1921–1936, 1998.

[18] D. Wang, Q. Tian, S. Gao, and W. Sung, "News sports video shot classification with sports play field and motion features," in *ICIP04*, 2004, pp. 2247–2250.

[19] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N.O'Connor, "Fusing mpeg-7 visual descriptors for image classification," in *International Conference on Artifical Neural Networks*, September 2005.

[20] A. Smeaton, H. Le Borgne, N. O'Connor, T. Adamek, O. Smyth, and S. D. Burca, "Coherent segmentation of video into syntactic regions," in *MVIP 2005 - 9th Irish Machine Vision and Image Processing Conference, Belfast, Northern Ireland*, August 2005.

[21] H. Le Borgne and N. O'Connor, "Natural scene classification and retrieval using ridgelet-based image signatures," in *Acivs 2005 - Advanced Concepts for Intelligent Vision Systems, Antwerp, Belgium*, September 2005.