RECENT TRENDS IN VIDEO ANALYSIS: A TAXONOMY OF VIDEO CLASSIFICATION PROBLEMS

Matt Roach, John Mason

Department of Electrical
& Electronic Engineering
University of Wales Swansea
SA2 8PP, UK
eeroachm@swansea.ac.uk
http://galilee.swan.ac.uk/

Li-Qun Xu, Fred Stentiford

Content and Coding Lab
BTexact Technologies - Research
Adastral Park
Martlesham Heath
Ipswich IP5 3RE, UK

ABSTRACT

This paper reviews and analyses the problems facing video classification. It investigates how the semantic gap can be bridged. It presents a new taxonomy for video calssifaction based on a lituture survey. It conclueds that narrowing the domain is the current approach to bridgeing the semantic gap.

1. INTRODUCTION

In this paper video classification is investigated. Video classification is a broad topic and some approaches are also termed video interpretation or video understanding. The term video is used here to refer to multimedia sequences comprised of both sound and a series of images. If a single mode of the video is to be referred to individually, the terms acoustic and visual are adopted.

One important influential component of video classification is the input domain i.e. the type of video being classified. Generally the input domain is said to vary between narrow and wide domains. *"A narrow domain has a limited variability in all relevant aspects of its appearance"*, Smeulders *et al* [1]. When this limited variability in the input domain exists the semantic content is often well defined and this property is utilised to design more accurate systems. Medical and industrial inspection video is very often specialist and recorded under deliberately controlled conditions, i.e. a narrow domain. Medical video is used to help diagnosis. Industrial inspection systems are used primarily in quality control capacities. Satellite and surveillance are perhaps recorded under less controlled conditions but have quite a lot of similarity within their respective classes therefore narrowing the input domain. Satellite video applications include research on the growth of different environmental habitats and military operations. Two common types of surveillance video are security generally used for property or public areas and motoring which is used to monitor traffic flow. Conversely, *"A wide domain has an unlimited and unpredictable variability in its appearance even for the same semantic meaning"*, Smeulders *et al* [1]. Broadcast video has a wide domain. The term *broadcast video* is used here to describe the type of video considered within the scope of this work, namely wide domain, mainstream entertainment video. This includes video presented with different media e.g. VHS, DVD, TV etc. Broadcast video has probably the greatest intra-class diversity i.e. the widest domain. It is in this area of video classification that this paper is concerned.

Perhaps the origins of broadcast video classification can be said to lie in image analysis applications. The retrieval of images containing specific printed texts is one of the earliest tasks to be tackled. Later came the introduction of computer vision techniques and searching for more general objects. A paper of this work is provided by Tamura and Yokoya [2]. In more recent times advances in the field have been both marked and matched by the increase in computer power. Recent papers of content-based image retrieval include those by Smeulders *et al* [1] and Rui *et al* [3], the latter having made some interesting forecasts for future work. These developments can be seen as origins of multimedia analysis, now in its relative embryonic infancy. What might be seen as a progression of the previous image work is today's developments in automatic classification and annotation of multimedia material. Reviews are presented by Aigrain *et al* [4] and more recently Brunelli *et al* [5]. Also Wang *et al* [6] presented an account with an acoustic bias.

There are many different aspects and approaches to content-based multimedia analysis, the end-application being an important determining factor. Motivation for content-based multimedia analysis arises from applications associated with content based queries. Approaches are aimed at providing efficient browsing, searching and retrieval of multimedia material. Application domains include large distributed digital libraries, broadcasting or production archives and video databases. The largest multimedia database is the world wide web (WWW) and specific approaches to this domain have been proposed [7]. Applications include: video editing, video education and training, video database navigation. Future applications could include browsing video email, searching video conferencing records and video on demand. These tasks are deemed necessary to organise and make optimum use of the ever-growing mass of multimedia material.

The focus of this paper is to analyse the problem of video classification. Specifically the analysis is to understand the problem facing automatic content-based semantic classification approaches. The major goal of approaches in this area is to reduce the *semantic gap*. *"The semantic gap is the lack of coincidence*

between the information that one can extract from the" multimedia material *"and the interpretation that the same data have for a user in a given situation"* Smeulders *et al* [1]. Redefined here as *"the lack of coincidence between the formative and cognitive information"*. Formative information is the shape, form or pattern held within the sequence of multidimensional matrices that make up the video. Cognitive information is the information pertaining to 'knowing'. This involves the interpretation of the formative information by a human viewer.

In Section 2 further explanations of applications that would make use of automatic content-based semantic classifications are given. In section 3 the analysis of the structure of the video classification problem is presented followed by a discussion, Section 4.

2. APPROACHES TO VIDEO CLASSIFICATION

This section presents the three main application drives for video content analysis. There are those techniques that attempt to summarise video, briefly explained in Section 2.2. There are those that are example-based queries, described by retrieval by example and are reviewed in Section 2.1. Finally there are approaches that attempt to identify content of the video, placing it into one of a set of previously defined categories; here this is termed video labeling and presented in Section 2.3.

2.1. Retrieval by example

In the example-based case, often termed query by example or retrieval by example, the system is presented with one or more examples of the type of video sequence required. The system then searches for videos with similar attributes. This leads to the need for definitions of similarity. Here the terms formative and cognitive are adopted to generalise into two types of information contained within a video. Our interpretations of the meaning of formative and cognitive are given in Section 1

Approaches to query by example can vary in complexity of the similarity measure between the example(s) and the retrieved video sequences and functionality of the user interface. Early examples were applied to limited domains such as satellite imagery using formative similarity of roads and river networks by detecting edges. Bouthemy *et al* [8, 9] present approaches that use basic formative motion measures to retrieve clips with similar motion attributes, claiming that the results often lead to clips with the same nature i.e. genre or events being retrieved, sport videos were especially good. Later, approaches make deliberate attempts to measure the cognitive similarity; for example Dori [10] modeled cognitive similarity using a visual object-process diagram scoring the similarity using object attributes: colour, location and lighting. A commercial system that employs many measures of similarity is presented by Flickner *et al* [11] called the QBIC, query by image and video content, system. Chang *et al* [12] present an approach to this task that supports spatio-temporal queries and a survey of content-based video retrieval is presented by Yongsheng and Ming [13], whilst mainly focusing on processing in MPEG compressed video domain.

2.2. Video summarising

Video summarising attempts to capture semantic content of a video and present a general highlight of the video in a shorter period of time. The aim is to compress the video in a cognitive sense rather than in a formative sense. The abstraction of video is related to video retrieval as a full system would allow the browsing of a retrieved video.

Video (content) summarisation and video skimming, although often used interchangeably in literature, there exists certain difference between them. While the result of the former is often represented by certain well-designed Graphics Interface depicted in key frames, in a mosaic form, or a kind of importance measures etc, the latter is almost certain to be in the form of a shorter semantics-preserved video file. See for example [14]. In this sense, one of the most important desired applications is real-time "adaptive video streaming". Aigrain *et al* [4] suggest in their paper that video skimming is more successful in limited domains such as education or news videos; namely those with very explicit speech or text (closed caption) contents.

2.3. Video indexing or labelling

This differs from retrieval by example in that these approaches attempt to generalise and make models for the classifications. However there may be similarity measures that are comparable between the two different approaches. This type of classification, where a given video is assigned to one of a pre-defined set of classes, has received more attention in the literature. Arguably the topic is now sufficiently mature to make comparisons of recognition accuracies across different modes, features and classifiers. Important practical issues that influence accuracies, in common with most such classification tasks, can be summarised in terms of the classes themselves, and their intra- and inter-class variations. In other words the number of classes together with the specific classes (video genre) under test are likely to greatly influence classification accuracy. Another practical factor in assessment is the type of testing, whether it is closed- or open-set identification (giving a 1 in N or 1 in N+1 decision) or whether it is the verification of a claimed identity label (a yes/no decision). Closed-set testing, though often not explicitly stated, seems to be the most common in the current literature. Where closed-set testing means that the test video sequences are by design a member of one of the defined class.

3. ANALYSIS OF THE VIDEO CLASSIFICATION PROBLEM

Figure 1 shows that the content of broadcast video can be conceptually divided into two parts. First, the semantic content, i.e. the story line told by the video. This is split into genre, events and objects. These elements are present in all media, for example a book can be a member of a genre and the story line contains events and objects. The ultimate classification of the video will be based on these story line semantics, which are further described in the following section. Second, there are the inherent properties of the digital media video. These are often termed editing effects. The equivalent to editing effects for a book would be chapters, pages

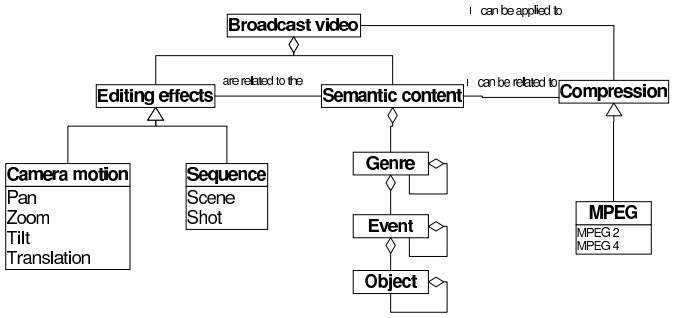

Fig. 1. General classification levels applied to broadcast video

and paragraphs although these are part of the book and are related to the story in a loose sense, they are not part of the story per se. This is the same for video: editing affects which are present in the video but are not part of the story.

3.1. Semantic content

In this section the different levels of semantic content classification are analysed. These classifications exist in the problem space; that is to say they have both acoustic and visual modes present. A video is split into two modes only in the solution space and the combination of these modes in the classification task is a desirable but formidable task. Section 3.1.1 presents the highest level of classification i.e. genre classification. Followed by event classification in Section 3.1.3 leading onto the lowest conceptual levels of classification, object classification in Section 3.1.4. Finally a description of the inter-relationships between the classification levels is summarised in Section 3.1.5.

3.1.1. Genre

Fig. 2. Here are some exmples of video genre

Broadcast video can be regarded as being made up of genre. The genre of a video is the broad class to which it may belong e.g. sport, news, cartoon etc, see Figure 2. Genre can themselves in turn be made up of genre. For example a sports program that reviews the results of all the soccer matches of the week can be a member of the genre soccer which in turn is a member of the genre sport. Genre classification at the same levels are mutually exclusive. That is to say a soccer program can not be a member of the genre hockey, as opposed to soccer, or a member of news as opposed to sport. The genre of a given video can be, and is often contested by reviewers or journalists. The determination of a genre is made by viewing the video content and often comes down to subjective views and semantic subtleties. However, when it comes to automate the process, researchers have deliberately chosen genre that are relatively well defined and commonly recognised. The brief history of the topic of genre classification shows that the genre tackled include: cartoon(5), news(8), commercial(6), music(2) and sport(8), where the number in brackets shows the number of occurrences of each genre in the literature [15, 16, 17, 18, 6, 19, 20, 21].

Some example approaches to genre classification begins with Fischer *et al* [15] whom presented one of the first attempts at genre classification in 1995, classifying: news, commercial, cartoon, tennis and car racing. Note that the sports are classified as separate genre this is because the authors suggest that a sport genre would be far too diverse to achieve reasonable success when attempting to classify it. They present a three step approach; First collecting basic acoustic and visual statistics, then as a second step they attempt to derive style attributes (SA) which include scene length, camera motion intensity, colour histograms for caption detection and the combination of the low-level statistics they also divide the acoustic data into speech, music and noise. Finally the distributions of these style attributes (SA) are used in the discrimination of genre. Liu *et al* [16, 17] present two approaches applied to TV genre classification. They investigate a range of statistical time and frequency features extracted from the acoustic signal. They use short term spectral estimates (23ms) and also include what are termed clip-based features which are measured over 1.5 seconds. These clip-based features are claimed by the authors to be more semantically meaningful. First they present results using a neural net classifier [16]. Subsequently, for the same experimental setup, they report an 11.9% improvement when using hidden Markov models (HMMs) [17]. The latter approach achieves an accuracy of 93.4% on 3 classes: sport (basketball and football), reports (news and weather) and commercials. here it is believed the improvement originates from the temporal modeling properties of the ergodic HMM.

From the previous examples it can be concluded that approaches to genre classification begin with low-level holistic features. This is because the types of features are robust to the diversity of material in a genre classification task. Secondly an attempt is made to include further semantic discrimination properties in their features. This is done in one or both of two ways: to combine othoginal low-level features [15] or to increase the period of time that the features are captured over [16].

In Figure 3a behavioral diagram illustrates scenarios for the genre classification previously defined. In this type of diagram some temporal information about the nature of genre changes is given. It can be seen that the genre changes to and from sport, a program named 'soccer roundup' broken up by commercials. Then the program changes to a program called 'music charts' which is a music video. This program too has commercials in-

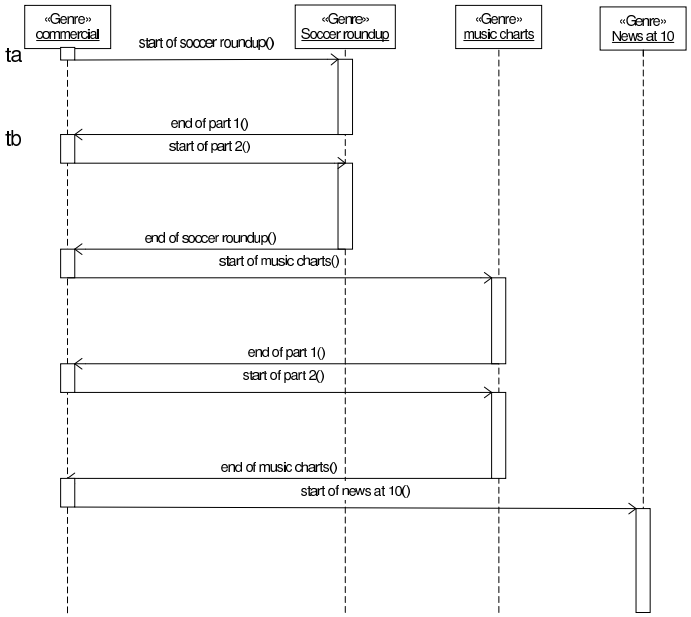

Fig. 3. An example of change of genre in a video sequence

terrupting it. Finally the genre changes to news when a program called 'News at 10' begins. These changes in genre are over relatively long periods of time. Typically the commercial breaks will last between 3 - 5 minutes and a program part between 10 - 20 minutes. Also seen in this diagram is the mutually exclusive nature of the genre of a video.

3.1.2. Related genre-level classifications

There are those approaches that can be said to be analogous to video genre classification. They are high-level classification tasks that originated in a single mode: visual or acoustic. Obliviously parts of these approaches can be utilised in the video case. Examples for the acoustic case include: speech music discrimination is a popular task. Carey *et al* [22] report on classification of acoustic data into speech and music limiting the data to just these two classes. They investigate the performance of a range of features including cepstral coefficients, amplitude, pitch and zero-crossings. They report that in each case including the derivatives of the feature improves accuracy and conclude that using cepstra and derivative coefficients gives the best performance. El-Maleh *et al* [23] discriminate speech and music and present their results for different music genre such as: Classical, Instrumental, Opera, Rock, Dance, Rap and Pop. Spina *et al* [24] present classification of general acoustic data into 7 classes. They investigate how many different general acoustic data classes can be successfully distinguished and how well the system can group the acoustic into homogeneous regions. They use fourteenth order cepstra (MFCC-14) and classify using models of Gaussian distributions. These approaches clearly have relevance to the task of video genre classification.

3.1.3. Events

Events are made up of objects and are defined by the objects interactions and interrelations over a finite period of time. Every video sequence contains events and in the same way as genre, events can be made up of other events. Furthermore there is an event constantly happening within a video, although some events are deemed more influential or important than others to the overall semantic content of the video. Therefore in the main approaches that attempt these types of events have been pursued, for example differentiating between classes of news shots [25], where the classes are: anchor shot, sound bite and voice over. In another event context a combination of visual static and dynamic features used in a limited environment is presented by Haering *et al* [26, 27] where event detection is applied to hunts in wildlife videos. Also Yow *et al* [28] analyse football videos for highlights. this is done by detecting the presence of the upright goal posts and tracking the motion of the ball, if the ball passes the posts this is deemed to be a shot on goal. Chang and Lee [29] classify events within a volley ball match. Colombo *et al* [30] use colour and motion features to generate semantic features entitled: action, happiness, relax, suspense and use these semantic based ratings to detect commercials that match a cognitive search criterion. These classifications usually operate under a narrow input domain as this is likely to improve success rates; they rely on the video being pre-classified into news, wildlife, sport and commercials respectively. In the sport case further refinement into soccer and volleyball is assumed.

If the input environment is limited then it enables more specific classification tasks to become tractable. Zelnik-Manor and Irani [31] present an event detection metric that is said to be more general. It is not aimed at specific events but is trained to detect similar event sequences within a video. There are also approaches that attempt to classify events that are less well defined for example moods such as, violence, happiness etc. Pfeiffer [32] reports on using the acoustic data mode for detecting the presence of violent scenes, which in this case are considered independent of the cultural environment of a user, typified by shots, explosions and cries. This approach is applied within the framework of the MoCA project (Movie Content Analysis). Event detection approaches in general add complexity to the feature extraction process to determine the more specific nature of events when compared with genre classification.

As with genre previously, a behavioral diagram of the nature of the change of events within a video sequence is given in Figurefig:eventchange. Note that the scale of the transitions has decreased this can be seen by looking a the two left columns showing the genre change and is referenced on both diagrams by t_a and t_b. In Figure 4 the first part of the program 'soccer roundup' is considered in more detail and some examples of event changes are illustrated. It is shown in the diagram that events change within a genre i.e. over a shorter period of time. Events are in the main also mutually exclusive usually defined this way for ease of classification. However the slow motion replay of a shot on goal can be considered as two events occurring at the same time. This can be seen in Figure 4 where 'slow-motion-replay' and 'shot-on-goal' are shown to exist in parallel.

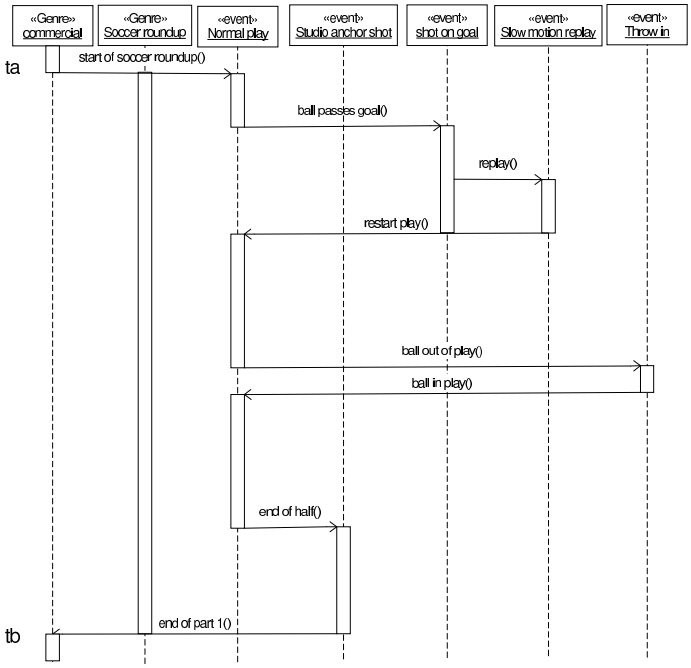

Fig. 4. An example of change of events in a video sequence

3.1.4. Objects

Finally, each event is shown to be made up of a number of objects. Objects are conceptually the lowest level of classification that can effect the semantic meaning of the video content. One of the most common objects to be attempted to be detected or classified is the human face [33, 34]. Object detection often requires very well structured feature extraction. Objects themselves are usually very well defined in structure and form therefore this knowledge of the formative appearance of an object is injected into the pattern recognition task. Generally this is done in two ways: to inject human understanding and knowledge of the problem into the feature extraction using rule-based algorithms for example mosaic faces i.e. there are two eyes which are above a nose which in turn is above a mouth that can move. Secondly pattern learning approaches are adopted when many examples of the object are given and a generalisation of the objects form is made. These types of approaches work well in limited domains for example a typical constraint put on face detection is that it must be facing front.

3.1.5. Inter-relationships between semantic classification levels

A taxonomy to represent the inter-relationships between genre classification and other content-based multimedia approaches is presented in Figure 5. This figure is an instantiation of the abstract view given in Figure 1. An analogous taxonomy is given by Smith and Chang [7] where they apply an image and video search engine for the world-wide web. In their system a semi-automatic process to build a key-term dictionary is proposed. These keywords are arranged into a proposed image and video subject taxonomy. Here the taxonomy arises from a literature review. High-level classification tasks are at the top of the diagram and low-level tasks at the bottom where the classifications become more

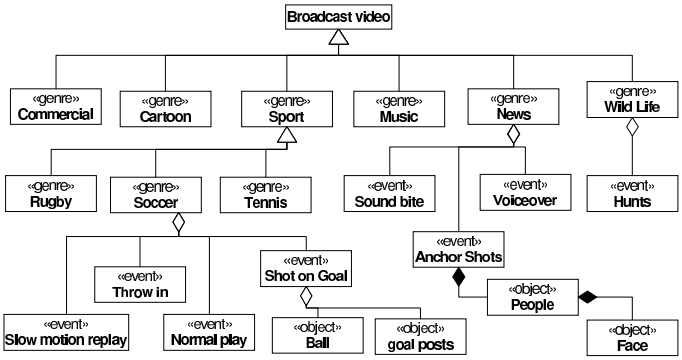

Fig. 5. A proposed taxomy of video classification

granular. The diagram illustrates the many to one relationship between the classification levels, by virtue of its expanding tree like structure, and it does so perhaps better than the abstract view in Figure 1. This means that for every step down in the level of classification level there are more classification tasks that could be attempted.

4. DISCUSSION

In this section a discussion of the capabilities of different classification approaches applied to different classification tasks is presented. From the literature it can be seen that there is an inverse relationship between the approach and the task. Is is seen that as the classification level decreases, i.e. the classifications become more granular, then the complexity or sophistication of feature level tends to increase. This is a general trend rather than an absolute rule. Event detection more often uses regional features to determine the occurrence of an event. For example Object detection approaches tend to use features that have more domain knowledge built into them. Linked to the increase in complexity of the feature extraction is the narrowing of the domain. As the classification level is decreased the prior classification level is often assumed, explicitly or implicitly. This makes the input domain narrower enabling an increase in feature or feature extraction complexity the material is inherently less diverse therefore allowing more knowledge to be injected into the features. This increase in feature complexity is aimed at reducing the semantic gap. Obviously it is the objective of many approaches to reduce the semantic gap [3] so that the features are more representative of what is considered important by humans. This will lead to automatic systems which rely on formative information being able to satisfy a human user who will make requests based on cognitive information.

Humans use high level semantics to interpret a video sequence using objects and their interrelations. High level classification tasks such as genre classification could be made by detecting objects and their interactions. For example if the word 'ball' was thought to be consistently used in sport or the visual object 'a ball' could be detected in a wide domain then these features could be used to assist in classifying the genre sport. However in practice the number of low level classification tasks for this bottom-up approach to work is beyond current technology. So approaches to high level tasks use low level features that are robust to the

inherent large-scale variation. This could lead to the conclusion that one good approach would be to first sub-divide the multimedia material into narrower domains, then apply more complex feature extractions in these narrow domains in attempts to extract more semantically meaningful results.

5. ACKNOWLEDGEMENTS

M. Roach, wishes to express his gratitude to EPSRC and BT for their financial support. The authors also thank, BT and UWS research group members for their inputs.

6. REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. on Pattern Analysis and Machine Itelligence*, vol. 22, no. 12, 2000.

[2] Hideyuki Tamura and Naokazu Yokoya, "Image database systems: A survey," *Pattern Recognition*, vol. 17, no. 1, 1984.

[3] Yong Rui, Thomas S. Huang, and Shih-Fu Chang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 1–23, 1999.

[4] P. Aigrain, H. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: A state of the art review," *Multimedia Tools and Applications*, vol. 3, pp. 179–202, 1996.

[5] R Brunelli, O. Mich, and C.M. Modena, "A survey on the automatic indexing of video data," *Visual Comm. and Image Representation*, vol. 10, pp. 78–112, 1999.

[6] J. Wang and T. Tan, "A new face detection method based on shape information," *PRL*, vol. 21, pp. 463–471, 00.

[7] J. R. Smith and S. F. Chang, "An image and video search engine for the World-Wide Web," *Storage and Retrieval for Image and Video Databases*, pp. 84–95, 1997.

[8] R. Fablet and P. Bouthemy, "Motion-based feature extraction and ascendant hierarchical classification for video indexing and retrieval," *3rd Int. Conf. on visual Information Systems, VISual'99, Amsterdam*, 1999.

[9] R. Fablet and P. Bouthemy, "Spatio-temporal segmentation and general motion characterization for videos indexing and retrieval," in *10th DELOS Workshop on Audio-Visual Libraries, Santorini*, 1999.

[10] D. Dori, "Cognitive image retrival," *Int conf. Pattern Recognition*, vol. 1, pp. 42–45, 2000.

[11] M. Flickner, "Query by image and video content," *IEEE Computer*, pp. 23–32, 1995.

[12] S-F. Chang, W Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A fully automated content based video search engine supporting spatio-temporal queries," *IEEE Trans. CSVT*, vol. 8, no. 5, pp. 602–615, 1998.

[13] Y. Yongsheng and L. Ming, "A Survey on content based video retrival," *http://www.cs.ust.hk/faculty/dimitris/COMP530/video_survey.pdf*.

[14] M.A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *Proc. IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 61–70.

[15] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *The 3rd ACM Int. Multimedia Conference and Exhibition*, 1995.

[16] Zhu Liu, Jincheng Huang, Yao Wang, and Tsuhan Chen, "Audio feature extraction and analysis for scene classification," in *IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, 1997.

[17] Zhu Liu, Jincheng Huang, and Yao Wang, "Classification of TV programs based on audio information using hidden Markov Model," in *IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, 1998.

[18] B-T. Truong, S. Venkatesh, and C. Dorai, "Automatic genre identification for content-based video categorization," *Int. Conf. Pattern Recgnition*, vol. 4, pp. 230–233, 2000.

[19] M.J. Roach, P. Martin-Granel, and J.S.D. Mason, "Camera motion extraction using correlation for motion-based video classification," in *Proc. IWVF4 Lecture Notes in Computer Science*, 2001, vol. 2059, pp. 552–562.

[20] M.J. Roach, J.S.D. Mason, and M. Pawlewski, "Video genre classification using dynamics," *Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.

[21] M.J. Roach and J.S.D. Mason, "Classification of video genre using audio," *Eurospeech*, 2001.

[22] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *ICASSP*, 1999.

[23] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/ music dicrimination for multimedia applications," *Int. Conf. on Acoustics, Speech and Signal Processing*, 2000.

[24] M. Spina and V. Zue, "Automatic transcription of general audio data: preliminary analysis," in *Proc. ICSLP*, 1996, pp. 594–597.

[25] K. Shearer, C. Dorai, and S. Venkatesh, "Incorporating domain knowlege with video and voice data analysis in news broadcasts," in *IEEE KDD-2000, Multimedia Data Mining workshop*, 2000.

[26] N.C. Haering, R.J. Qian, and M.I. Sezan, "Detecting hunts in wildllife videos," *IEEE Int. Conf. Multimedia Computing and Systems*, vol. 1, 1999.

[27] N.C. Haering, R.J. Qian, and M.I. Sezan, "A semantic event detection approach and its application to detecting hunts in wildlife video," *IEEE Trans. on Circuits and Systems for Video Tecnology*, 1999.

[28] D. Yow, B-L Yeo, M. Yeung, and B. Liu, "Analysis and presentation of soccer highlights from digital video," in *Proc. Asian Conf. on Computer Vision*, 1995.

[29] C-W Chang and S-Y Lee, "A video information system for sport motion analysis," *Journal of Visual Languages and Computing*, vol. 8, pp. 265–287, 1998.

[30] C. Colombo, A. Del Bimbo, and P. Pala, "Retrieval of commercials by video semantics," *CVPR*, pp. 572–577, 1998.

[31] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," 2001.

[32] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in *Proc. ACM Multimedia*, 1996, pp. 21–30.

[33] J.D. Brand, J.S.D. Mason, and M. Pawlweski, "Face detection in colour images," *Int. conf. Image Processing*, 2001.

[34] H. Wang and S-F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 7, no. 4, 1997.