GENERATION OF SEMANTIC CUES FOR SPORTS VIDEO ANNOTATION

J Kittler, K Messer, W J Christmas, B Levienaise-Obadia and D Koubaroulis

Centre for Vision, Speech and Signal Processing
School of Electronics, Computing and Mathematics
University of Surrey, Guildford GU2 7XH, UK

ABSTRACT

The use of video and audio features for automated annotation of audio-visual data is becoming widespread. A major limitation of many of the current methods is that the stored indexing features are too low-level - they relate directly to properties of the data. In this work we apply a further stage of processing that associates the feature measurements with real-world objects or events. The outputs, which we call "cues", denote the probability of the object being present in the scene. An additional advantage of this approach is that the cues from different types of features are presented in a homogeneous way.

1. INTRODUCTION

The ever increasing popularity of sport means that there is a vast amount of sports footage being recorded every day. Ideally, all this sports video should be annotated, and the meta-data generated on it should be stored in a database along with the video data. Such a system would allow an operator to retrieve any shot or important event within a shot at a later date. Also, the level of annotation provided by the system should be adequate to facilitate simple text-based queries. Such a system has many uses, such as in the production of television sport programmes, documentaries and helps ensure our culture preservation.

Due to the large amount of material being generated, manual annotation is both impractical and very expensive. However, automatic annotation is a very demanding and an extremely challenging computer vision task as it involves high-level scene interpretation.

Perhaps the most well known automatic video annotation system reported in the literature is Virage [1]. Virage has an open framework which allows for the integration of many audio and video analysis tools in real time and places the data into an industry-standard database such as Informix or Oracle. However, the number of analysis tools available is limited, although always expanding. The main problem with Virage is that no effort has been made to bridge the gap between the information provided by the low-level analysis tools and the high-level interpretation of the video, which is required for our application.

Other work, specific to some form of sports annotation include [7] in which camera motion is used to help in the automatic annotation of basketball. Mo et al. utilize state transition models, which include both top-down and bottom-up processes, to recognise different objects in sports scenes [6]. In [8] work has been undertaken to distinguish between sports and non-sports MPEG compressed video. Finally, MIT have been working on the analysis of American football video [2].

The ASSAVID project is concerned with the development of a novel system which will provide a semantic annotation of sports video. This annotation segments the sports video into semantic categories (e.g. type of sport) and permits the user to formulate queries to retrieve events that are significant to that particular sport.

ASSAVID will provide the complete system. The engine will comprise of a set of software tools to aid an operator in the generation of the high-level annotation for incoming sports video. These tools are based on a set of lower-level audio and video analysis tools, which we term cue detectors. A contextual reasoning engine will then be used to analyse the output of these cue detectors and attach semantic information to the video being analysed. The generated meta-data and video data will then be stored in a database which is based on a mixture of IBM's Media360 and Informix. ASSAVID will also provide a Java graphical user interface to the database which will allow the user to browse the video, view sequences and generate story boards, formulate queries and analyse and modify the generated indices.

In this paper we outline some of the general details about the design of ASSAVID, give some examples of the cues that we are working on along with some preliminary results.

2. USER REQUIREMENTS

An investigation into the user requirements for a sports annotation system was carried out within the sports department at the British Broadcasting Corporation. It was found that two separate logging (meta-data creation) processes are currently used when analysing the sports video and this functionality will need to be provided by ASSAVID. These are termed production logging and posterity logging.

Production logging is when there is a need to perform some of the annotation in real time, as the event is actually happening. Presently, this annotation is made manually and is mainly focused on marking shots for inclusion in a subsequent program. Posterity logging is typically performed off-line. There are fewer time constraints on how quickly

0-7803-6725-1/01/$10.00 ©2001 IEEE

the annotation needs to be built and the major aim is to get a very detailed description of the sports video being archived. Again, at present these logs are done by hand by skilled librarians. It is not uncommon for a one hour sequence to take over ten hours to index fully.

Many different identification tasks for the logging were also identified for ASSAVID. Some of the more important ones include shot change detection; shot description (e.g. camera movement, lens effects and framing terms); identification and classification of sport (e.g. football, tennis, interview); event identification within sport (e.g. goals, headers, race start. red cards etc.); sports personality detection (e.g. Alan Shearer, Tim Henman and audio aescriptors (e.g. crowd cheering/booing, gunshot). Most of these tasks require a high-level of understanding of the video being analysed.

3. CUE DETECTION

The objective in the automatic annotation of video material is to provide indexing material that describes as usefully as possible the material itself. In much of the previous work in this area, the annotation consisted of the output of various feature detectors. By itself, this information bears no semantic connection to the actual scene content — it is simply the output of some image processing algorithms. In this project we are taking the process one stage further. By means of a set of training processes, we aim to generate an association between the feature detector outputs and the occurrence of actual scene features. Thus for example we might train the system to associate the output of a texture feature detector with crowds of people in the scene. We can then use this mechanism to generate confidence values for the presence of a crowd in a scene, based on the scene texture. We denote the output of this process as a "cue". These cues can then be combined in the contextual reasoning engine to generate higher-level information, e.g. the type of sport being played. A simple scheme for the generation of for example some visual cues is shown in Fig. 1.

In addition to visual cues, we are generating cues based on speech, non-speech sounds and text from on-screen captions.

We can define the process of cue generation a little more formally as follows. The confidence measure we would like to determine is the probability of the occurrence of a cue C given a feature measurement m. This might for example be computed using Bayes Rule:

$$P(C|m) = \frac{p(m|C)P(C)}{p(m|C)P(C) + p(m|\overline{C})P(\overline{C})} \qquad (1)$$

where $p(m|C)$ (and $p(m|\overline{C})$) is a value from the p.d.f. of the feature measurement given the presence (or absence) of the cue. The forms of the p.d.f.s are estimated via the training process, the exact form of which depends on the type of measurement and the amount of data available containing the particular cue. The prior probabilities $P(C)$ also have to be determined. Initially we simply establish which cues are mutually exclusive, and assign equal probabilities accordingly.

The contextual reasoning engine may also require information about the spatial and temporal location of the measurements.

4. VISUAL CUES

For this system many different cue detection methods are being developed. In this section we briefly discuss three visual cue generation methods. Each method can be used to form a number of different cue-detectors provided that suitable training data is available. These methods are:

4.1. Neural network

Each cue-detector is a neural network trained on colour and texture descriptors computed at a pixel level on examples of image regions of the cue of interest (see [5]) and on image regions which are known not to contain the cue. The resulting trained network is then able to distinguish between the features of cue and non-cue pixels. A high output represents the case when the feature vector of the pixel belongs to same distribution as the cue and vice-versa.

To check for the presence of a cue in a test image the same colour and texture features are computed for each test image pixel and the feature vector is passed to the neural network. If many high outputs are observed then this gives an indication of how likely it is that the given cue is present in the image. Cues suitable for this method include sky, grass, tennis court and athletics track.

4.2. Multimodal Neighbourhood Signature

In the Multimodal Neighbourhood Signature approach (see [4]), object colour structure is represented by a set of invariant features computed from image neighbourhoods with a multimodal colour density function. The method is image-based – the representation is computed from a set of examples of the object of interest, a cue in this context.

In the implemented method, MNS are sets of invariant colour pairs corresponding to pairs of coordinates of the located density function modes from each neighbourhood. The MNS signatures of all the example images are then merged into a composite one by superposing the features (colour pairs). Considering each colour pair as an independent object descriptor (a *detector*), its discriminative ability is measured on a pre-selected *training set* consisting of positive and negative example images. A simple measure, the absolute difference of true and false positive percentages is computed. Finally, the n most discriminative detectors are selected to represent the object of interest. For the reported experiments n was set to 3.

For matching, we view each detector as a point in the detector space. A hypersphere with radius h is defined around each point. Given another image of the object, measurements are likely to lie inside the detector hyperspheres. A

binary n-tuple is computed for each test image, each binary digit assigned 1 if at least one test measurement was within the corresponding detector sphere, 0 otherwise. One of 2^n possible n-tuples are the measurements output from the matching stage. The relative frequency of each possible n-tuple over the positive and negative cue examples of the training set define an estimate of the probability of each measurement given the cue and not given the cue respectively. These 2 numbers are output to the decision making module.

4.3. Texture codes

The texture-based cue detector consists of two components: a training phase, in which a model for the cue is created using ground truth, and the cue extractor (see [3]). In the training stage, template regions from the keyframes are selected for each cue. Several templates are needed for each cue to account for appearance variations. Textural descriptors are extracted form the templates using a texture analysis module based on Gabor filters. These descriptors, with the number of occurrences, form the model for the cue.

In the cue extractor, the whole image is presented to the texture analysis module. Then, by comparing the result with the model, a coarse detection component selects the three templates which are most likely to be visually similar to an area of the image being annotated. The similarity is evaluated using the histogram intersection. We increase the computational efficiency by hashing the meta-data; this also enables us to compute the similarity measure only for templates which share descriptors with the input image. A localisation component finally identifies the areas of the image which the selected templates match most closely, and the image location which yields the best match confidence is retained. The highest confidence, with its location, are the output for the cue.

4.4. Experimental Results

Using each outlined method four different cue detectors were built, to distinguish between boxing, swimming, tennis and athletic track events.

These cue detectors were then applied to test images grabbed from a digital video tape of sports material. Examples of these images can be seen in the first column of figure 2. The confidence measures of the cue being either present or not-present, using each cue and on each image, were then calculated. These confidences obtained using the neural network based method are shown in the bar charts of column 2. The confidences found using the MNS method are shown in column three. Finally, the confidences using the texture-code based approach are shown in column 4.

As one can see from these results, in all cases the highest confidence obtained for each image is the correct cue for that sport for all three methods. As the methods used to generate these cues are complementary, the output confidences can be combined in the contextual reasoning engine. This will make the annotation much more robust and reliable.

5. CONCLUSIONS

In this paper we described a developing system for the annotation of sports video, named ASSAVID. This system is based on the concept of cues, which allow us to extract high-level information from sets of low-level features computed on the incoming sports video data. Three of the visual cue methods were then briefly outlined. These were demonstrated to work well on a set of images containing the sports of athletics, boxing, swimming and tennis.

Presently, we are continuing to develop a variety of cue methods, including methods based on features computed on the audio track and the motion vectors. We have also started to work on the contextual reasoning engine which will combine the outputs of the cues and make decisions as to the exact scene content of the video.

Acknowledgements This work has been performed within the framework of the ASSAVID project granted by the European IST Programme.

6. REFERENCES

[1] http://www.virage.com.

[2] S.S. Intille and A.F. Bobick. A framework for representing multi-agent action from visual evidence. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, July 1999.

[3] B Levienaise-Obadia, J Kittler, and W Christmas. Defining quantisation strategies and a perceptual similarity measure for texture-based annotation and retrieval. In IEEE, editor, *ICPR'2000*, volume III, Sep 2000.

[4] J. Matas, D. Koubaroulis, and J. Kittler. Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature. In D. Vernon, editor, *ECCV*, LNCS vol. 1842, pages 48–64, Berlin, Germany, June 2000. Springer.

[5] K Messer and J Kittler. A region-based image database system using colour and texture. *Pattern Recognition Letters*, pages 1323–1330, November 1999.

[6] H. Mo, S. Satoh, and M. Sakauchi. A study of image recognition using similarity retrieval. In *First International Conference on Visual Information Systems (Visual'96)*, pages 136–141, 1996.

[7] D.D. Saur, Y.-P. Tan, S.R. Kulkarni, and P.j. Ramadge. Automated analysis and annotation of basketball video. In *SPIE Storage and Retrieval for Still Image and Video Databases V, Vol.3022*, pages 176–187, 1997.

[8] V.Kobla, D.DeMenthon, and D.Doermann. Identifying sporst video using replay, text and camera motion features. In *SPIE Storage and retrieval for Media Database 2000*, pages 332–342, 2000.

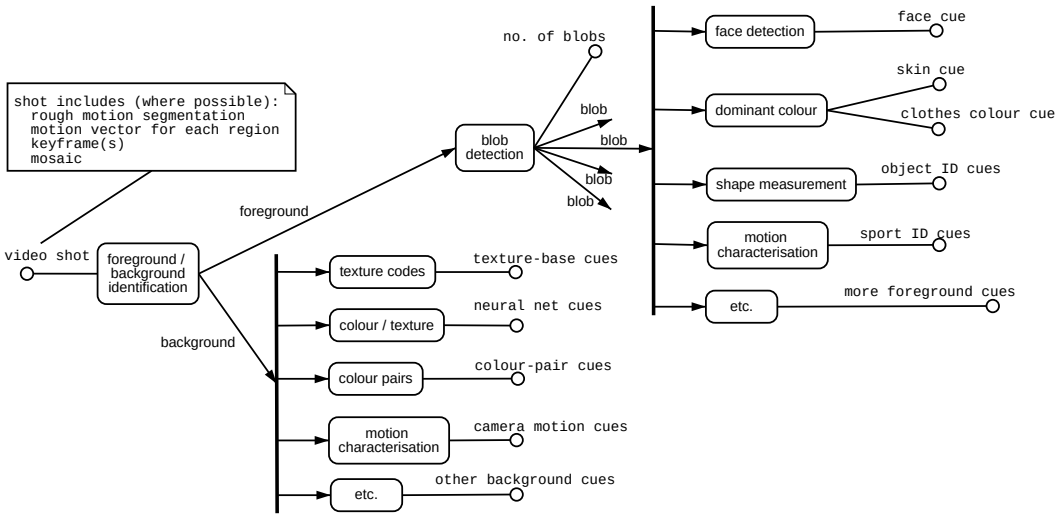

Fig. 1. Refinement of architecture assuming reliable foreground / background separation

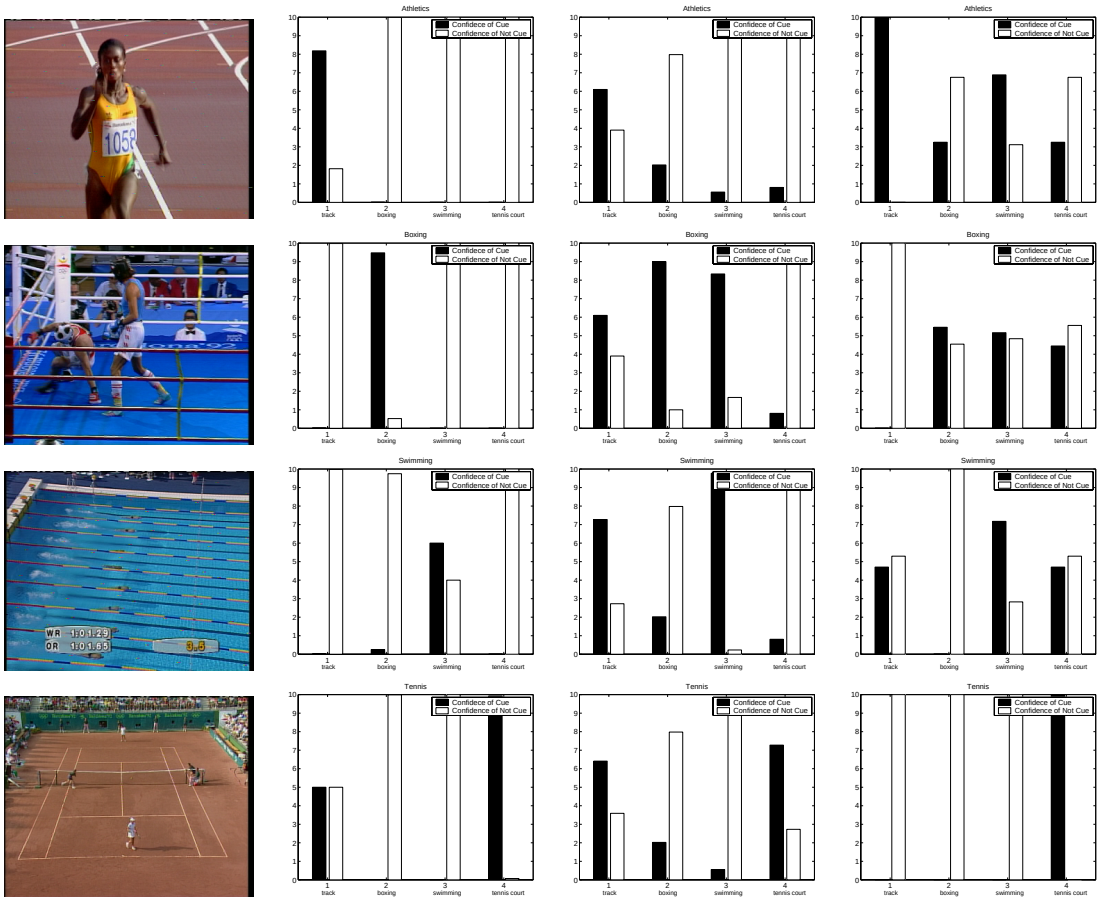

Fig. 2. `Column 1` Four test images taken from the DV tapes. `Column 2` Confidences for the neural network based cues. `Column 3` Confidences for the MNS based cues. `Column 4` Confidences obtained using texture codes. NB. all the confidences have been normalised to fit in the range zero to ten.