Colour-Based Object Recognition for Video Annotation

Dimitrios Koubaroulis[1], Jiří Matas[1,2], Josef Kittler[1]

[1] Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

[2] Center for Machine Perception, Czech Technical University, Prague, 120 35, CZ

Abstract

We propose a colour-based object recognition method for video annotation. The semantic gap between image measurements and symbolic labelling is bridged by assuming the existence of objects whose appearance can be associated with some desired image categories (labels). A colour-based method, the Multimodal Neighbourhood Signature (MNS) is used. We propose an automatic method for learning the object representation from multiple images. A new MNS matching strategy is also introduced, making use of a K-class classifier based on a binary feature vector computed from the object's MNS signature.

In the experimental section, the proposed method is evaluated for annotating sport video keyframes using raw broadcast video material provided by the BBC. Despite the poor quality of some of the images and a wide range of appearance variations (occlusion, illumination and viewpoint change, camera noise and cluttered background to name a few), correct (average 85%) object recognition and sport classification was achieved for a set of four selected objects/sports.

1 Introduction

Many organisations (e.g news agencies and broadcasting companies) keep large collections of images and video sequences. Working with such data sets requires a time consuming and costly effort to archive and retrieve items of interest from the collection. Automation of this process is highly desirable. The assignment of concise descriptions to image and video sequences (a task called *annotation*) has been the subject of content-based image and video retrieval research [2]. Several image/video sequence properties can be exploited to represent visual data such as colour, texture, detected text, motion, shot duration etc. Here we develop a colour-based annotation system.

Mapping the computed (here colour-based) measurements to symbolic labels which correspond to the objects present in an image as perceived by a human, is not trivial and is often called the *semantic gap* [10]. In this paper, we present an object-based approach for automatic annotation of video sequences. We bridge the semantic gap by assuming that an image label is computed as a function of the presence of specific physical objects in the image. Object recognition is a well-studied problem and a number of successful applications has been reported (e.g. [7, 8]). Our approach is only limited by the existence of characteristic objects whose presence adequately indicates an image category (class). Labelling each image with one of a set of possible labels is viewed as a classification problem. Image colour measurements are classified to one of a number of object 'classes' which are mapped (one to one) to a category label. This object-based approach is quite different from other methods where annotation is achieved e.g. via pixel-based classification (e.g. [9]).

In this work, we apply a colour-based object modelling and recognition method, called the Multimodal Neighbourhood Signature (MNS) [8], for sport video annotation. Assuming a set of example images/regions for learning object appearance, a feature selection algorithm and a novel MNS matching algorithm are introduced. In contrast with other object-based recognition algorithms, MNS does not make use of automatic spatio-temporal segmentation (as in [3]), neither does it focus on a specific application domain (e.g.annotation of basketball sequences). In [12] an augmented model of appearance was described, using a combination of visual features. In our experiments, good results were obtained using colour alone. MNS has been tested for image retrieval [8, 6], however image labelling (classification) using MNS has not been addressed.

The proposed method is tested on sports video data provided by the BBC for the ASSAVID project [1]. Our approach is particularly useful for this type of image data since there exist objects whose appearance is characteristic of a sport discipline. Such objects, for instance, are the boxing ring, the taek-won-do tatami and the athletics track to name a few.

2 The MNS object model

The MNS method, introduced by Matas et al. in [8], is image-based; only a set of images (or regions) are re-

1051-4651/02 $17.00 © 2002 IEEE

quired to describe object appearance. Local colour structure is represented by stable features computed from image neighbourhoods with a multimodal colour density function. The positions of the modes used for the computation of the invariants are robustly filtered, stable values, efficiently established in the RGB[1] colour space with the mean shift algorithm [4]. The features used in that paper, are functions of coordinates of pairs of the located density function modes from each neighbourhood. Each MNS signature consists of a number of selected invariants and representative locations. Features are selected using a suppression algorithm to eliminate almost identical measurements. In [8], MNS matching was implemented as a model-oriented stable matching problem [5] and successful application to image retrieval and object recognition was reported.

In published experiments using MNS, a single example image was used to describe object appearance. In this paper, a set of images of each sought object are assumed available to *learn* object appearance. An object representation is obtained by manually selecting a small number of image regions that show each sought object in a subset of the example images. The MNS signatures of all the example regions are merged into a composite MNS by superposing the features (colour pairs) and suppressing identical features.

2.1 Learning the object representation

The set of example images for all objects is used as a training set (excluding those used for computing the object MNS signature). From the object MNS, a small set of discriminative features is selected. In feature selection, the features (colour pairs in the signature) are considered independent. We view each feature as a point in the measurement space. A hypersphere with radius h is defined around each point. Each feature in the object MNS is matched against every feature of every image in the training set. For the comparison, the L_2 metric is used in the colour pair (RGB^2) space (see formula in [8]). The decision to whether a measurement is present in a test image is positive if at least one test measurement is within the corresponding object feature hypersphere, 0 otherwise. Consequently, the percentage of the sought object and other examples which has produced a particular measurement is calculated. The features are then sorted by the absolute difference of true (object) and false (other) positive percentages. This difference is taken as a measure of the discrimination ability of the feature. Finally, the n most discriminative features are selected to represent the object of interest.

[1]Other colour spaces (e.g. HSV) could be used without changing the algorithm. In experiments, MNS was insensitive to the space used.

2.2 Object recognition

In the original MNS paper, features were matched independently [8]. Here, cooccurence of features is exploited. After feature selection for each object, a set of n selected features defines a so-called *detector* for the particular object. Given measurements from another image of the object, they are likely to lie inside the object feature hyperspheres (designed exactly as above). Outputting 1 for each object feature found in the test image, and 0 for the others, a binary vector measurement $D = \{0,1\}^n$ is formed by grouping the n outputs.

Making a decision about the appearance of the object in the image is posed as a K-class classification problem, where K is the number of categories. We design a K-dimensional binary feature classifier, using the following structure of the likelihoods $P(x|C_i)$, where x is the observation vector and $C_i, i = 1..K$ is the class represented by object i. First, let us assume that for each class C_i , there is one object detector D_i. For each test image, the observation vector we consider consists of a concatenation of all detector outputs $D_i, i = 1..K$, resulting in a binary vector $m = d_i^j, i = 1..K, j = 1..n$ of size $K \times n$, where n is the length of the detector's output (assumed equal for all detectors here). No constraints are placed on the statistical model of binary features produced by a detector D_i; the probability distributions $P(D_i|C_i)$ are estimated in full from the training set. Note, that the d_i^js are not independent, however the class-conditional probabilities $P(D_i|C_i)$ of the D_is forming the observation m are assumed independent. The class-conditional probability of m is computed as

$$P(m|C_t) = \prod_i P(D_i|C_t) \quad (1)$$

where $t = 1..K$.

In the classification stage, a Bayesian approach with estimates $P(m|C_i)$ replacing the true probabilities is used. Assuming equal ('flat') prior probabilities, the maximum $P(m|C_i)$ is the output of a maximum aposteriori probability (MAP) classifier. In annotation applications, the data is typically available a priori and the prior probabilities are given or they can be estimated using e.g. some empirical Bayesian method. In our experiments, equal priors were assumed for each class, since the number of images representing each sport in the test set was controlled. A test image is rejected (and labelled "unknown") when the following (ad hoc) criterion is true for class C_i with maximum $P(m|C_i)$:

$$P(D_i|C_i) < P(D_i|\overline{C_i}) \quad (2)$$

The class-conditional probabilities $P(D_i|C_i)$ for each detector D_i and the probability $P(D_i|\overline{C_i})$ are computed as relative frequencies from the training set. To avoid the so called *zero-frequency problem* in probability estimation

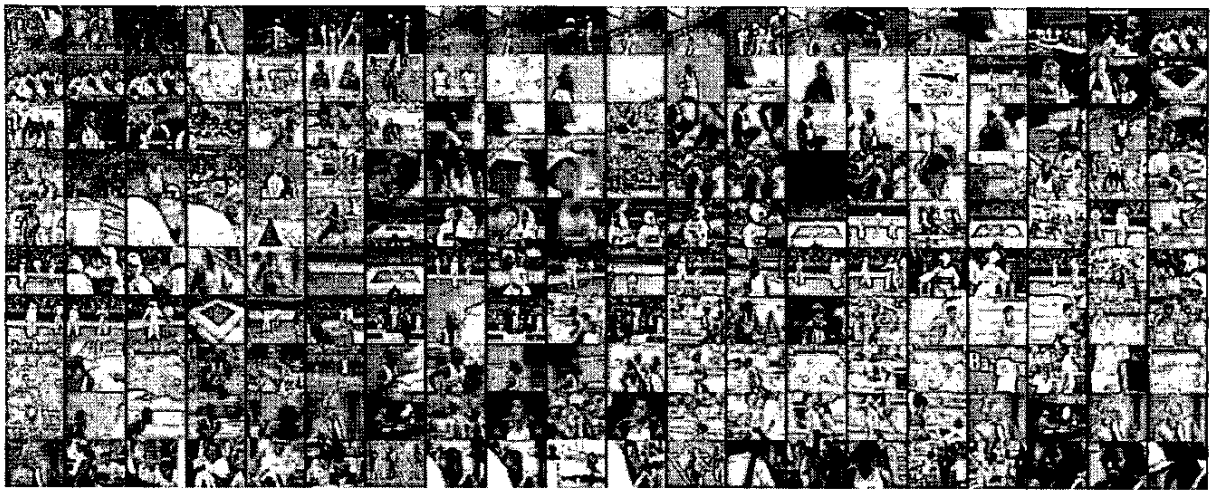

Figure 1. Sample frames from the BBC video sequences used in the experiment

due to the small number of examples in the training set, a smoothed estimate [11]

$$P(D_i|C_k) = \frac{f_v + 1}{T_k + K} \quad (3)$$

was used instead of the maximum likelihood estimate; where f_v is the frequency of observation $D_i = v$ and $T_k, k = 1..K$ is the number of images of class C_k in the training set.

The approach presented so far is image-based i.e. the spatio-temporal characteristics of the video frames are not exploited. The use of a Hidden Markov Model, in conjunction with the annotations computed on a per-image basis is expected to improve performance and is being investigated.

3 The annotation experiment

For the reported experiments, 328 images of size 288 × 360 were selected from a larger set of 1800 images grabbed randomly from 5 digital videotapes of the BBC coverage of three Olympic games (1992, 1996, 2000). A sample of the database is shown in Fig. 1. The test images included frames showing the sought objects from many viewpoints, occluded by the players/crowds and usually viewed in heavily cluttered background. Finally, some of the images show the objects in different times of the day, typically resulting in illumination change. In this paper, we assume that the colour balancing system of the cameras partly compensates for the illumination change, therefore our image processing takes place in the RGB invariant space. All internal parameters of the MNS method were set to default values, that is, no attempt was made to optimise the method for the specific data set. No images were excluded from the original grabbed sequence which included many frames with artifacts and noise, exactly as they were recorded from the cameras.

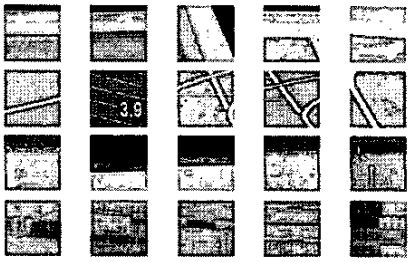

Figure 2. Five examples from those used for computing the object MNS

Table 1. No. of example and test images

Object	Examples	Training	Test
Tennis Court	11	10	11
Athletics Track	11	10	21
Taek-won-Do Tatami	8	35	48
Swimming Pool Lane	6	18	26
Unknown (other)	–	–	113

Four characteristic objects/sports were selected to demonstrate MNS performance. Namely, the tennis court, the athletics track, the taek-won-do tatami and a swimming pool lane marker. Five samples of these examples for each object are shown in Fig. 2. Due to the simple colour structure of the objects used, the number of detectors n was set to 3. The number of images used for each object in the training and test sets are listed in Table 1.

For each object (equivalently sport), the performance of the method was measured as the percentage of correct classifications per sport. The confusion matrix is presented in

Table 2. Classification results: Confusion matrix

	% Estimated label				
True label	Swimming	Taek-won-do	Tennis	Track&field	Unknown
Swimming	**0.9**	0.0	0.0	0.1	0.0
Taek-won-Do	0.0	**0.7**	0.0	0.0	0.2
Tennis	0.0	0.0	**1.0**	0.0	0.0
Track&Field	0.0	0.0	0.0	**0.8**	0.2
Unknown	0.0	0.0	0.1	0.5	**0.5**

Table 2. Good discrimination was achieved in general with an average correct labelling of 85% for the 4 sports. Some false positives in track recognition were mainly due to the presence of many other objects with track-like colours e.g. skin colours, tennis court etc.

4 Conclusions

We proposed a colour-based object recognition approach to video annotation. Labelling a video frame was posed as a classification problem. Object-based measurements were classified as belonging to one of a set of objects which were selected to be representative of a symbolic label useful for the archival/retrieval of the sequence.

The Multimodal Neighbourhood Signature method was used for object modelling. A method for automatic learning of the object representation from multiple example regions was proposed. For matching object representations, a new algorithm was also proposed, using a K-class classifier based on a binary feature vector computed from the object MNS.

The algorithm was tested for annotating sport video keyframes using raw broadcast video material provided by the BBC. Despite the poor quality of some the images and a wide range of appearance variations (occlusion, illumination and viewpoint change, camera noise and cluttered background to name a few), correct (85%) object recognition and sport classification was achieved for a set of selected objects/sports.

Possible extensions of the proposed method include a method for automatic selection of the detector size, a feature selection algorithm and an integrated system that will exploit more visual or other cues and their appearance as a function of temporal information available with a video sequence. Finally, annotation based on the presence and location of the projection of multiple objects in an image is being investigated.

References

[1] http://www.bpe-rnd.co.uk/assavid/.

*The authors acknowledge funding by the EC IST-13082 ASSAVID project. JM was supported by the EC IST-2001-32184 ACTIPRET project.

[2] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, 1999.

[3] S.-F. Chang, W. Chen, H. J. Meng, H.Sundaram, and D. Zhong. VideoQ: An Automated Content Based Video Search System Using Visual Cues. In *ACM Multimedia*, pages 313–324, 1997.

[4] K. Fukunaga and L. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Transactions in Information Theory*, 21(1):32–40, 1975.

[5] D. Gusfield. *The Stable Marriage Problem:Structure and Algorithms*. MIT Press, 1989.

[6] D. Koubaroulis, J. Matas, and J. Kittler. Colour-based Image Retrieval from Video Sequences. In *CIR, 3rd UK Conf. on Image Retrieval*, pages 1–12, 2000.

[7] Z-N. Li, O. Zaiane, and Z. Tauber. Illumination Invariance and Object Model in Content-based Image and Video Retrieval. *Journal of Visual Communication and Image Representation*, 10(3):219–244, 1999.

[8] J. Matas, D. Koubaroulis, and J. Kittler. Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature. In *ECCV*, pages 48–64, 2000 http://www.ee.surrey.ac.uk/Personal/D.Koubaroulis/.

[9] E. Saber, A. Tekalp, R. Eshbach, and K. Knox. Automatic Image Annotation Using Adaptive Colour Classfication. *Journal of Graphical Models and Image Processing*, 58(2):115–126, 1996.

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

[11] I. Witten and T. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory*, 37(4):1085–1094, 1991.

[12] Deng Y., Mukherjee D., and Manjunath S.B. Netra-V: Towards an Object-based Video Representation . In *SPIE*, pages 202–213, 1998.