ADAPTIVE RULE-BASED RECOGNITION OF EVENTS IN VIDEO SEQUENCES

V. Tzouvaras, G. Tsechpenakis, G. Stamou and Stefanos Kollias

Image, Video and Multimedia Laboratory
Department of Electrical and Computer Engineering
National Technical University of Athens, Greece
{tzouvaras, gtsech}@image.ntua.gr, {gstam, stefanos}@softlab.ntua.gr

ABSTRACT

Knowledge-based fuzzy inference and neural learning are used in this paper in order to model the event recognition task in semantic video analysis. The advantage of their use is the symbolic nature of the representation of the knowledge concerning the events to be recognized. Moreover, this knowledge can be adapted with the aid of data taken from video sequences. The proposed system has been tested in soccer video sequences for detecting some complex predetermined (and represented in the form of rules) events.

1. INTRODUCTION

Lately, there is a growing interest in the automatic annotation of *multimedia* content [6]. For the multimedia content description in the form of metadata information, the Moving Pictures Expert Group (MPEG) has developed the Multimedia Content Description Interface (MPEG-7). The goal of this standard is to develop a rich set of standardised tools to enable machines to generate and understand audiovisual descriptions for retrieval, categorisation and filtering purposes. Although for the *description* of audiovisual information there is lately a significant progress, the pace on *automatic extraction* of this description (especially of its semantic part) is rather low. This is an effective of the limitations of state of the art semantic video analysis systems that mainly characterised by recognition of a small number of specific objects, static behaviour, lack of contextual recognition and long computational time [1], [5].

Machine learning systems are lately examined for the modelling and implementation of semantic video analysis systems and overcome the limitations described above, mainly for two reasons. First, built-in knowledge about entities and adaptation permits robustness and uncertainty handling. Second, learning allows updating of knowledge, i.e. adaptation to environmental changes.

In this paper, a generic hybrid neurofuzzy architecture is examined for this purpose. This architecture combines subsymbolic and symbolic processing, providing rule-

based inference and adaptation using the ideas of neural learning and fuzzy inference. Artificial neural networks are well suited for the adaptive analysis of subsymbolic information, while knowledge-based fuzzy logic techniques are capable of including and refining rules that form the basis for symbolic descriptions. Over the past years, neural network approaches have been successfully combined with fuzzy and symbolic processing techniques. Hybrid connectionist-symbolic models constitute a promising approach towards developing more robust and versatile intelligent systems [3]. The main focus of the present work is the symbolic descriptions of events, in terms of *rules* employing fuzzy hypotheses (the degree of belief of the presence of specific objects and their interrelations), extracted from the video sequence. For the above representation, a neurofuzzy structure is used in order to preserve the symbolic nature of the information, also providing learning capabilities.

2. DETECTION OF FUZZY PREDICATES

Let $y = [y_1, y_2, \dots, y_m]$ denote a fuzzy set defined on the set of events to be recognized. Actually, each y_i represents the degree in which the i -th event is recognized. The input of the proposed system is a fuzzy set $x = [x_1, x_2, \dots, x_n]$ defined on the set of the fuzzy predicates (objects and relations like ball, player, close(ball, player) etc), with each x_i representing the degree in which the i -th predicate is detected. The proposed system extracts these fuzzy predicates from the video sequence and represents the association $f: X \rightarrow Y$ in a neurofuzzy structure.

In this section we briefly discuss the main key-points of the algorithm used for the detection of fuzzy predicates, i.e. for the extraction of the main mobile objects (and their interrelations) in video sequences, in the general case of sequences acquired by a moving camera. The proposed method is actually a modification of the approach proposed by Makarov et al [4]. In our work, bounding boxes or polygons are extracted; each

one of them is the minimum one entirely including only one mobile object and thus can be used as approximation of its shape and position in each frame of the sequence. In order to improve the extraction of the moving objects, as well as to eliminate the appearing noise, a multiresolution approach is also proposed. The main advantage of this approach is that it retains a low computational cost, given that all the operations are performed on the edge maps and not the images themselves.

The algorithm is based on background updating, where the slow insertion of updated areas of the observed scene into a reference background image is interpreted as a newly detected static feature of the scene; otherwise any observed alterations are considered either as noise or as a moving object. The main advantage of this algorithm is the ability to reject changes that occur due to the presence of excessive temporal clutter. In the following the specific steps of the algorithm are briefly described.

When a mobile camera is used, we first obtain the edge-maps and extract a reference background image. In order to achieve that, two criteria are used: (a) a frame counter that has high values at static pixels and lower values at transient ones and (b) a gradient change counter that indicates at which pixels the image gradient's direction significantly changes and thus these pixels are considered as transient. By merging these two counters into one decision module, and thresholding it appropriately (adaptation time), we obtain a binary reference background image. The inverse of this image is combined with the current frame edge map, through the logical operator *AND*, and results in a binary one containing the transient pixels (or edges).

The next step is to cluster the extracted transient pixels, if more than one mobile object exist in the sequence. In this module a block-matching motion estimation scheme is used, only for the pixels that have been considered as transient and the motion vectors are then clustered, using a fuzzy c-means scheme.

In the final step, the bounding polygon of each main mobile object is extracted. The basic idea is to estimate the minimum polygon for every point set, since more than one polygon may contain a main mobile object. The Graham Scan method is used for that purpose, due to its low computational cost [9].

In order to obtain more accurate bounding polygons closer to the main moving objects' contour and without any significant noise, the above approach is implemented in different image resolutions. For each one of them, one bounding polygon for each main moving object is estimated. In the cases of low resolution, the background's undesirable details are eliminated, but also worse bounding polygons are obtained. On the other hand, in high-resolution cases, the bounding polygons are estimated more accurately, but also other undesirable pixels appear. By combining the results of all resolutions

(in our experiments two resolution levels proved to be sufficient), we can obtain accurate polygons, close to main mobile objects' contours without any significant noise.

Once the bounding polygons have been extracted, it is not difficult to extract the fuzzy predicates using simple heuristic criteria (for each specific case), concerning some features (colour, shape etc). It is mentioned that the matching of the fuzzy predicates is a fuzzy number representing the degree of belief that this object appears in the video, at the specific time and location.

3. THE EVENT RECOGNITION PROCESS

After the evaluation of the fuzzy predicates, the events represented in the knowledge of the system can be recognized with the aid of fuzzy systems' reasoning [2]. One of the widely used ways of constructing fuzzy inference systems is the method of approximate reasoning which can be implemented on the basis of compositional rule of inference. Fuzzy inference systems can be implemented with the aid of max-min compositions of fuzzy sets [2]. The need for more general research lead to the representation of fuzzy inference systems on the basis of generalized sup-t-norm compositions [8]. A t-norm (triangular norm) is a function $t:[0,1] \times [0,1] \rightarrow [0,1]$ satisfying for any $a, b, d \in [0,1]$ the next four conditions:

1. $t(a,1) = a$ and $t(a,0) = 0$
2. $b \leq d$ implies $t(a,b) \leq t(a,d)$
3. $t(a,b) = t(b,a)$
4. $t(a,t(b,d)) = t(t(a,b),d)$

Using the definition of t-norms, an additional operator $\omega_t:[0,1] \times [0,1] \rightarrow [0,1]$ is defined as:

$$\omega_t(a,b) = \sup\{x \in [0,1] : t(a,x) \leq b\}$$

While the t-norm t may be interpreted as logical conjunction, the corresponding ω_t operation may be interpreted as logical implication. With the aid of the above operators, compositions of fuzzy relations can be defined. These compositions are used in order to construct fuzzy relational equations and represent the rule-based symbolic knowledge with the aid of fuzzy inference [7].

Let X, Y, Z be three discrete crisp sets with cardinalities n, m and l respectively, and $A(X,Y), B(Y,Z)$ be two binary fuzzy relations. The definitions of sup- t and inf- ω_t compositions are given by

$$(A \circ^t B)(i,j) = \sup_{k \in N_m} t\{A(i,k), B(k,j)\}, i \in N_n, j \in N_l$$

$$(A \circ^{\omega_t} B)(i,j) = \inf_{k \in N_m} \omega_t\{A(i,k), B(k,j)\}, i \in N_n, j \in N_l.$$

Let us now proceed to a more detailed description of the proposed neurofuzzy architecture (Figure 1). Its

general structure is shown in Fig. 1. It consists of two layers of *compositional* neurons which are extensions of the conventional neurons [7]. While the operation of the conventional neuron is described by the equation:

$$y = \alpha \left(\sum_{i=1}^n w_i x_i + \mathcal{G} \right)$$

where α is non-linearity, $\mathcal{G}$ is threshold and w_i are the weights, the operation of the sup- t compositional neuron is described by the equation:

$$y = \alpha' \left\{ \sup_{j \in N_n} t(x_j, w_j) \right\}$$

where t is a fuzzy intersection operator (a t-norm) and α' is the following activation function:

$$\alpha'(z) = \begin{cases} 0 & x \in (-\infty, 0) \\ x & x \in [0, 1] \\ 1 & x \in (0, +\infty) \end{cases}$$

A second type of compositional neuron is constructed using the ω_i operation. The neuron equation is given by:

$$y = \alpha' \left\{ \inf_{j \in N_n} \omega_j(x_j, w_j) \right\}$$

Figure 1: The proposed neurofuzzy system

The proposed architecture is a two-layer neural network of compositional neurons. The first layer consists of the $\inf-\omega_i$ neurons and the second layer consists of the sup- t neurons. The system takes as input, the fuzzy predicates extracted with the procedure described in the previous section and gives to the output the recognized events. The first layer computes the antecedents of the fuzzy rules, while the second implements the fuzzy reasoning using the modus ponens schema.

The rules describing the events are generally of the form “if *fuzzy_predicate* (1) and ... and *fuzzy_predicate*

(n) then *event* (i)”. Each rule consists of an antecedent (the *if* part of the rule) and a consequence (the *then* part of the rule). The set of the rules of the system is given in symbolic form by the expert and is used in order to initialize the neurofuzzy network (giving its initial structure and weights). During the learning process the number of neurons in the hidden layer and the weights of the two layers may change with the aid of a learning with the objective of the error minimization. The learning algorithm that supports the above network is applied in each layer independently. During the learning process, the weight matrices are adapted in order to approximate the solution of the fuzzy relational equation describing the association of the input with the output. Using a traditional minimization algorithm (for example the steepest descent), we cannot take advantage of the specific character of the problem. The algorithm that we use is based on a more sophisticated *credit assignment* that “blames” the neurons of the network using the knowledge about the topographic structure of the solution of the fuzzy relation equation [7]. After the learning process, the network keeps its transparent structure and the new knowledge represented in it can be extracted in the form of fuzzy IF-THEN rules.

4. SIMULATION EXAMPLE

In this section we present some experimental results demonstrating the applicability of the proposed algorithm in video sequences. The examples are illustrated in soccer video sequences, where the moving objects are the players and the ball. That kind of sequences has some special characteristics. These special characteristics, apart from the size and the motion of the moving objects, also concern the different camera views of the game, some fast motion cameras and their frequent zooming. These characteristics are used as constraints for the extraction of the moving objects. An appropriate parameter tuning is performed to detect and localize the objects even if their size is very small or the background is not homogenous, or else different cameras capture an event. The example consists of two soccer video sequences. In each video sequence an event is taking place. The first video is showing the event ‘ball draws away from player’, while the second is showing the event ‘player A passes the ball to player B’. In each case the moving objects in attention, are located with polygons. In order to detect an event, we must first recognize the moving objects and extract them as fuzzy predicates. In the first example, the moving objects are the player and the ball. These predicates are not capable to compose a rule, which characterize the specific event. An event requires spatiotemporal relations between the object predicates. In the event ‘ball drawing away from player’, the relation is the distance between

the two objects, and it is calculated from the position of their metacenter. Furthermore, the timing of the detected predicates is a vital factor for the detection of an event. For example, the predicates 'player', 'ball' and the 'distance between them' cannot be used to detect the event 'ball drawing away from the player'. The distance of the ball from the player must be increasing. The subsymbolic module gives degrees of confidence for each fuzzy predicate every one-second. These degrees are stored into a buffer and fed into the neurofuzzy network. In Fig. 2 the distances in pixels are: (a) 12, (b) 23, (c) 40, (d) 60, (e) 90, (f) 140. These numbers are normalized and transformed in degrees of confidence for the predicate 'distance between the ball and the player'.

Figure 2: Event 'ball draws away from player A'

In Fig. 3, a more complicated event is shown. The ball is drawing away from player A, moving in a certain direction and when a new player appears (player B), it is concluded that the ball is approaching him. The incorporated fuzzy predicates in this example are: 'player A', 'player B', 'ball', 'distance of the ball from player A', 'distance of the ball from player B' and 'player A and player B same team'. The localization of the concerned objects is shown in boxes and for the description of this event some more information, apart from the objects relative positions, is needed. Firstly, the objects included in the event are recognized and the distances are calculated in pixels, showing that the ball is going from player A to player B: (a) player A – ball = 21, (b) player A – ball = 61, ball – player B = 88, (c) player A – ball = 64, ball – player B = 85, (d) player A – ball = 72, ball – player B = 77, (e) player A – ball = 81, ball – player B = 67, (f) player A – ball = 87, ball – player B = 61. In Fig 3(a) the ball is set given the description 'close to player A', in Figs. 2(b-d) the ball is described as 'drawing away from player A', in Figs. 2(e-f) the ball is described as 'moving towards player B' and in Fig. 3(f) the ball is 'close to player B'. Also, if we extract some extra features for the moving objects, like the mean intensity of the regions inside the bounding polygons, we can

conclude that players A and B belong to the same team and thus the event 'player A passes to player B' is detected.

Figure 3: Event 'Player A passes the ball to player B'

It must be noted that during the examination of the events, some extra processing is being done, like image filtering (with morphological filters) to eliminate noise, and the description of each object and specific action or event is given with some degree of uncertainty.

ACKNOWLEDGEMENTS

This work has been partially funded by the EU IST-1999-20502 FAETHON project.

5. REFERENCES

- [1] A. Del Bimbo, Visual Image Retrieval, CA: Morgan Kaufmann, San Francisco, 1999.
- [2] G. Klir and Bo Yuan, Fuzzy Sets and Fuzzy Logic, Theory and Applications, New Jersey, Prentice Hall, 1995.
- [3] C.-T. Lin, C.S. Lee, *Neural fuzzy Systems: A neuro-fuzzy synergism to intelligent systems*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
- [4] A. Makarov, J.M. Vesin and F. Reymond, "Intrusion Detection Robust to Slow and Abrupt Lighting Changes," Real-Time Imaging, SPIE-2661, 1996, 44-54.
- [5] M.R. Naphade and T.S. Huang, Extracting Semantics from Audiovisual Content: The Final Frontier in Multimedia Retrieval, IEEE Transactions on Neural Networks, Vol. 13, Num 4, July 2002.
- [6] S. Santini, Exploratory Image Databases: Content-based Retrieval, New York: Academic, 2001.
- [7] G.B. Stamou, S. G. Tzafestas, "Resolution of composite fuzzy relational equations based on Archimedean triangular norms", Fuzzy Sets and Systems, 120, pp. 395-407, 2001.
- [8] G.B. Stamou, S. G. Tzafestas, "Neural fuzzy relational systems with a new learning algorithm", *Mathematics and computers in simulation*, pp. 301-304, 2000.
- [9] R.L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Inform. Process. Lett.*, 1:132-133, 1972.