TIME INTERVAL MAXIMUM ENTROPY BASED EVENT INDEXING IN SOCCER VIDEO

Cees G.M. Snoek and Marcel Worring

Intelligent Sensory Information Systems, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{cgmsnoek, worring}@science.uva.nl

ABSTRACT

Multimodal indexing of events in video documents poses problems with respect to representation, inclusion of contextual information, and synchronization of the heterogeneous information sources involved. In this paper we present the Time Interval Maximum Entropy (TIME) framework that tackles aforementioned problems. To demonstrate the viability of TIME for event classification in multimodal video, an evaluation was performed on the domain of soccer broadcasts. It was found that by applying TIME, the amount of video a user has to watch in order to see almost all highlights can be reduced considerably.

1. INTRODUCTION

Effective and efficient extraction of semantic indexes from video documents requires simultaneous analysis of visual, auditory, and textual information sources. In literature several of such methods have been proposed, addressing different types of semantic indexes, see [12] for an extensive overview. Multimodal methods for detection of semantic events are still rare, notable exceptions are [3, 7, 8, 10]. For the integration of the heterogeneous data sources a statistical classifier gives the best results [12], compared to heuristic methods, e.g. [3]. In particular, instances of the Dynamic Bayesian Network (DBN) framework, e.g. [8, 10]. Drawbacks of the DBN framework are the fact that the model works with fixed common units, e.g. image frames, thereby ignoring differences in layout schemes of the modalities, and thus proper synchronization. Secondly, it is difficult to model several asynchronous temporal context relations simultaneously. Finally, it lacks satisfactory inclusion of the textual modality.

Some limitations are overcome by using a maximum entropy framework. Which has been successfully applied in diverse research disciplines, including the area of statistical natural language processing, where it achieved state-of-the-art performance [4]. More recently it was also reported in video indexing literature [7], indicating promising results for the purpose of highlight classification in baseball. However, the presented method lacks synchronization of multimodal information sources. We propose the Time Interval Maximum Entropy (TIME) framework that extends the standard framework with time interval relations, to allow proper inclusion of multimodal data, synchronization, and context relations. To demonstrate the viability of TIME for detection of semantic events in multimodal video documents, we evaluated the method on the domain of soccer broadcasts. Other methods using this domain exist, e.g. [2, 14]. We improve on this existing work by exploiting multimodal, instead of unimodal, information sources, and by using a classifier based on statistics instead of heuristics.

The rest of this paper is organized as follows. We first introduce event representation in the TIME framework. Then we proceed with the basics of the maximum entropy classifier in section 3. In section 4 we discuss the classification of events in soccer video, and the features used. Experiments are presented in section 5.

2. VIDEO EVENT REPRESENTATION

We view the problem of event detection in video as a pattern recognition problem, where the task is to assign to a pattern x an event or category ω, based on a set of n features $(f_1, f_2, \ldots, f_n)$ derived from x. We now consider how to represent a pattern.

A multimodal video document is composed of different modalities, each with their own layout and content elements. Therefore, features have to be defined on layout specific segments. Hence, synchronization is required. To illustrate, consider figure 1. In this example a video document is represented by five time dependent features defined on different asynchronous time scales. At a certain moment an event occurs. Clues for the occurrence of this event are found in the features that have a value within the time-window of the event, but also in contextual features that have a value before or after, the actual occurrence of the event. As an example consider a goal in a soccer match. Clues that indicate this event are a swift camera pan towards the goal area before the goal, an excited commentator dur-

This research is sponsored by the ICES/KIS MIA project and TNO.

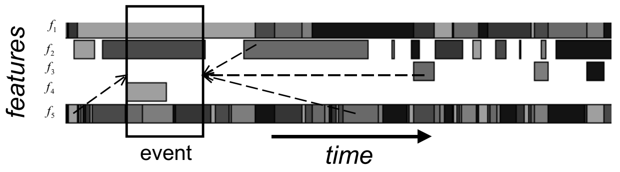


Figure 1: *Feature based representation of a video document with an event (box) and contextual relations (dashed lines).*

ing the goal, and a specific keyword in the closed caption afterwards. Hence, we need a means to express the different visual, auditory, and textual features into one fixed reference pattern without loss of their original layout scheme. For this purpose we propose to use binary fuzzy Allen time interval relations [1]. A total of thirteen possible interval relations, i.e. *precedes*, *meets*, *overlaps*, *starts*, *during*, *finishes*, *equals*, and their inverses, identified by *_i* at the end, can be distinguished. A margin is introduced to account for imprecise boundary segmentation, explaining the fuzzy nature. By using fuzzy Allen relations it becomes possible to model events, context, and synchronization into one common framework. When we choose a camera shot as a reference pattern, a goal in a soccer broadcast can be modelled by a swift camera pan that *precedes* the current camera shot, excited speech that *finishes* the camera shot, and a keyword in the closed caption that *precedes_i* the camera shot. Note that the *precedes* and *precedes_i* relations require a range parameter to limit the amount of contextual information that is included in the analysis.

Thus, we choose a reference pattern, and express a co-occurrence between a pattern x and category ω by means of binary fuzzy Allen relations with binary features, f_j. Where each f_j is defined as:

$$f_j(x,\omega) = \begin{cases} 1, & \text{if } \lambda_j(x) = true \text{ and } \omega = \omega'; \\ 0, & \text{otherwise;} \end{cases} \tag{1}$$

Where $\lambda_j(x)$ is a predicate function that checks for a fuzzy Allen relation, and ω' is one of the categories, or events.

3. PATTERN CLASSIFICATION

Having defined the feature based pattern representation, we now switch to classification using the Maximum Entropy framework [4].

For each feature the expected value over the training set $\mathcal{S}$ is computed:

$$E_{\tilde{p}}(f_j) = \sum_{x,\omega} \tilde{p}(x,\omega) f_j(x,\omega) \tag{2}$$

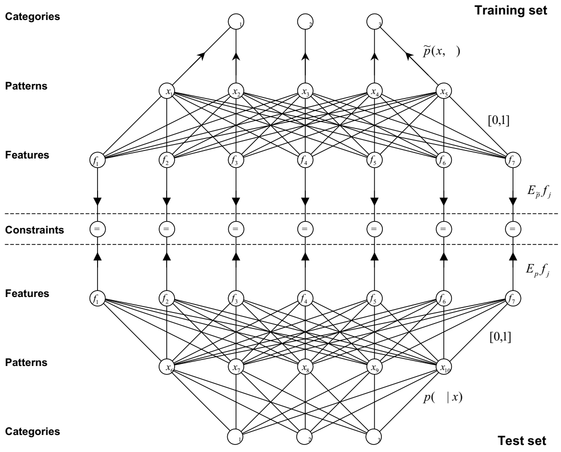


Figure 2: *Simplified visual representation of the maximum entropy framework.*

Where $\tilde{p}(x,\omega)$ is the observed probability of x and ω in $\mathcal{S}$. This creates a model of $\mathcal{S}$. To use this model for classification of unseen patterns, i.e. the reconstructed model $p(\omega|x)$, we require that the constraints for $\mathcal{S}$ are in accordance with the constraints of the test set $\mathcal{T}$. Hence, we need the expected value of f_j with respect to the model $p(\omega|x)$:

$$E_p(f_j) = \sum_{x,\omega} \tilde{p}(x) p(\omega|x) f_j(x,\omega) \tag{3}$$

where $\tilde{p}(x)$ is the observed probability of x in $\mathcal{S}$. The complete model of training and test set is visualized in figure 2. We are left with the problem of finding the optimal reconstructed model $p^*(\omega|x)$. This is solved by restricting attention to those models $p(\omega|x)$ for which the expected value of f_j over $\mathcal{T}$ equals the expected value of f_j over $\mathcal{S}$. From all those possible models, the maximum entropy philosophy dictates that we select the one with the most uniform distribution. Assuming no evidence if nothing has been observed. The uniformity of the conditional distribution $p(\omega|x)$ can be measured by the conditional entropy, defined as:

$$H(p) = -\sum_{x,\omega} \tilde{p}(x) p(\omega|x) \log p(\omega|x) \tag{4}$$

The model with maximum entropy, $p^*(\omega|x)$, should be selected. It is shown in [4] that there is always a unique model $p^*(\omega|x)$ with maximum entropy, and that $p^*(\omega|x)$ must have a form that is equivalent to:

$$p^*(\omega|x) = \frac{1}{Z} \prod_{j=1}^{n} \alpha_j^{f_j(x,\omega)} \tag{5}$$

where α_j is the weight for feature f_j and Z is a normalizing constant, used to ensure that a probability distribution results. The values for α_j are computed by the *Generalized Iterative Scaling* (GIS) [5] algorithm. Since GIS relies on

Feature	*Fuzzy Allen*	*Range (s)*
Camera work	*during*	
Person	*during*	
Close-up	*precedes_i*	0 - 40
Goal keyword	*precedes_i*	0 - 6
Card keyword	*precedes_i*	0 - 6
Substitution keyword	*precedes_i*	0 - 6
Excitement	*All relations*	0 - 1
Info block	*precedes_i*	20 - 80
Person block	*precedes_i*	20 - 50
Referee block	*precedes_i*	20 - 50
Coach block	*precedes_i*	20 - 50
Goal block	*precedes_i*	20 - 50
Card block	*precedes_i*	20 - 50
Substitution block	*during*	
Block length	*during*	

Table 1: *Features with fuzzy Allen time interval relations.*

both $E_{\tilde{p}}(f_j)$ and $E_p(f_j)$ for calculation of α_j, an approximation is used that relies only on $E_{\tilde{p}}(f_j)$ from $\mathcal{S}$ [9]. This allows to construct a classifier that depends on the training set only. Hence, by using the maximum entropy classifier we can focus on what features to use, since relative importance of each feature is computed automatically.

4. EVENT INDEXING IN SOCCER BROADCASTS

To demonstrate the viability of the TIME framework for event detection in multimodal video, we consider the domain of soccer. Typical highlight events that occur in a soccer match are goals, penalties, yellow cards, red cards, and substitutions. We take as a basic pattern a camera shot, since this is the most natural candidate for retrieval of events. In what follows, we will highlight several multimodal features used for modelling those events. Features were chosen based on reported robustness and training experiments. The parameters for individual detectors were found by experimentation. The features with fuzzy Allen relations are summarized in table 1.

4.1. Static information

Game related information, like the players who played during the match, name of the coaches and referees and so on, can be found on the UEFA web site. This information was extracted with a web spider and stored in a game database. This information is used to improve a visual feature detector that is explained later on.

4.2. Textual features

The teletext (closed caption) provides a textual description of what is said by the commentator during a match. This information source was analyzed for presence of informative keywords, like *yellow, red, card, goal, 1-0, 1-2*, and so on. In total 30 informative stemmed keywords were defined for the various events.

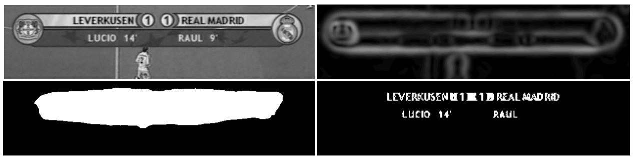


Figure 3: *Different steps in overlay segmentation: color edge detection, marked watershed, and Video OCR.*

4.3. Visual features

From the visual modality we extracted several features. The type of camera work [13] was computed for each camera shot. A face detector [11] was applied for detection of persons. The same detector formed the basis for a close-up detector. Close-ups are detected by relating the size of detected faces to the total frame size. Often, a director shows a close-up of a player after an event of importance. One of the most informative pieces of information in a soccer broadcast are the visual overlay blocks that give information about the game. For segmentation of the overlay blocks we used a color invariant edge detector [6] combined with a marked watershed algorithm. The segmented region was the input for a Video Optical Character Recognition (VOCR) module [13], see figure 3. Results of VOCR are noisy, but by using the game database and fuzzy string matching we were able to reliably detect team names, player names, coach names, referee names, and descriptive text like: *misses next match* or *3 goals in 6 matches*. This information is fused to classify an overlay block as either info, person, referee, coach, goal, card, or substitution. The duration of visibility of the overlay block is also used, as we observed that substitution and info blocks are displayed longer on average.

4.4. Auditory features

From the auditory modality the excitement of the commentator is a valuable feature [10]. For such a feature to work properly, we require that it is insensitive to crowd cheer. This can be achieved by using a high threshold on the average energy of a fixed window, and by requiring that an excited segment has a minimum duration of 4 seconds.

5. EVALUATION

For the evaluation of TIME we digitized 8 live soccer broadcasts from TV, about 12 hours in total. The videos were digitized in 704 × 576 resolution MPEG-2 format. The audio was sampled at 16 KHz with 16 bits per sample. The time

	Ground truth		*Maximum Entropy*	
	Total	Duration	Relevant	Duration
Goal	12	3:07	10	10:14
Yellow Card	24	10:35	22	26:12
Substitution	29	8:09	25	7:36
$\sum$	65	21:51	57	44:02

Table 2: *Evaluation results, duration in minutes:seconds.*

stamped teletext was recorded with a teletext receiver. We used a representative training set of 3 hours and a test set of 9 hours. We focussed on 3 events, $\omega \in \{$*yellow card, substitution, goal*$\}$, red card and penalty were excluded from analysis since there was only one instance of each in the data set. We manually labelled all the camera shots as either belonging to one of four categories: yellow card, goal, substitution, or unknown. We defined the different events as follows:

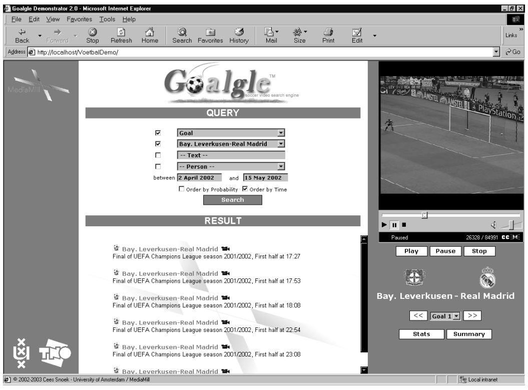
- *Goal*: begin until end of the camera shot showing the actual goal;
- *Yellow card*: begin of the camera shot showing the foul until the end of the camera shot that shows the referee with the yellow card;
- *Substitution*: begin of the camera shot showing player that goes out, until the end of the camera shot showing player that comes in;

Since events can cross camera shot boundaries, adjacent events are merged. Hence, we cannot use precision and recall as an evaluation measure. From a users perspective it is unacceptable that events are missed. Therefore, we strive to find all events. Since it is difficult to exactly define the start and end of an event in soccer video, we introduce a tolerance value T (in seconds) with respect to the boundaries of detection results. We used a T of 7 s. for all events. Results are visualized in table 2. Note that almost all events are found, and that the amount of video that a user has to watch before finding those events is only two times longer compared to the best case scenario.

The weights computed by GIS indicate that for *goal* and *yellow card* specific keywords in the closed captions, excitement with *during* and *overlaps* relations, and the presence of an overlay nearby are important features. For *substitution* the auditory modality is less important.

6. CONCLUSION

We combined multimodal information sources into a common framework for the purpose of event detection in video documents. The presented Time Interval Maximum Entropy framework allows for proper modelling of events, synchronization, and asynchronous contextual information relations. Our method was evaluated on the domain of soccer.


Goalgle Demonstrator 2.0 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Size Print Edit
Links
Address
Go
MediaMill
Goalgle™
soccer video search engine
QUERY
Search
RESULT
Play Pause Stop
<< >>
Stats Summary
Paused
26328 / 84991
Local intranet

http://localhost/VoetbalDemo/

☑ Goal
☑ Bay. Leverkusen-Real Madrid
□ -- Text --
□ -- Person --
between 2 April 2002 and 15 May 2002
□ Order by Probability ☑ Order by Time

Bay. Leverkusen-Real Madrid
Final of UEFA Champions League season 2001/2002, First half at 17:27

Bay. Leverkusen-Real Madrid
Final of UEFA Champions League season 2001/2002, First half at 17:53

Bay. Leverkusen-Real Madrid
Final of UEFA Champions League season 2001/2002, First half at 18:08

Bay. Leverkusen-Real Madrid
Final of UEFA Champions League season 2001/2002, First half at 22:54

Bay. Leverkusen-Real Madrid
Final of UEFA Champions League season 2001/2002, First half at 23:08

Bay. Leverkusen-Real Madrid

Bay. Leverkusen - Real Madrid
Goal 1

© 2002-2003 Cees Snoek - University of Amsterdam / MediaMill

Figure 4: *The Goalgle soccer video search engine.*

Results show that a considerable reduction of watching time can be achieved. The indexed events were used to build the *Goalgle* soccer video search engine, see figure 4.

7. REFERENCES

[1] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *IJDAR*, 2002.

[2] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using HMMs. In *IEEE ICME*, 2002.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. on Multimedia*, 4(1):68–75, 2002.

[4] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[5] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[6] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *IEEE TPAMI*, 23(12), 2001.

[7] M. Han, W. Hua, W. Xu, and Y. Gong. An integrated baseball digest system using maximum entropy method. In *ACM Multimedia*, 2002.

[8] M. Naphade and T. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia*, 3(1):141–151, 2001.

[9] OpenNLP Maxent. `http://maxent.sf.net/`.

[10] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV formula 1 programs. In *IEEE ICME*, 2002.

[11] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE TPAMI*, 20(1):23–38, 1998.

[12] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*. To appear.

[13] The Lowlands Team. Lazy users and automatic video retrieval tools in (the) lowlands. In *TREC*, 2001.

[14] D. Yow, B. Yeo, M. Yeung, and B. Liu. Analysis and presentation of soccer highlights from digital video. In *ACCV95*.