

Time Interval based Modelling and Classification of Events in Soccer Video

Cees G.M. Snoek

Marcel Worring

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

Keywords: Multimodal video indexing, event classification, time interval modelling, maximum entropy, support vector machine, soccer

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 propose to model events in multimodal video by means of time interval relations, to tackle aforementioned problems. Our approach exploits the powerful properties of statistical classifiers, like the Maximum Entropy and Support Vector Machine classifiers. To demonstrate the viability of our approach for event classification in multimodal video, an evaluation was performed on the domain of soccer broadcasts. It was found that the amount of video a user has to watch in order to see almost all highlights can be reduced considerably. Furthermore, we found that a Support Vector Machine performs better than a Maximum Entropy classifier.

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 [16] for an extensive overview. Multimodal methods for detection of semantic events are still rare, notable exceptions are [4, 10, 12, 13, 14]. For the integration of the heterogeneous data sources a statistical classifier gives the best results [16], compared to heuristic methods, e.g. [4]. In particular, instances of the Dynamic Bayesian Network (DBN) framework, e.g. [13, 14]. 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. Recently proposed methods exploiting a Maximum Entropy classifier [10] and Support Vector Machine (SVM) classifier [12] indicate promising results for the purpose of event detection. However, the way of modelling events in those contributions suffers from the same limitations.

We propose to model events by means of time interval relationships, to allow proper inclusion of multimodal data, synchronization, and context relations. To demonstrate the viability of our approach for detection of semantic events in multimodal video documents, we evaluated the method on the domain of soccer broadcasts using both a Maximum Entropy and SVM classifier. Other methods using this domain exist, e.g. [2, 18]. 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 by means of the time interval relations. Then we proceed with the basics of the Maximum Entropy and SVM 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, \dots, 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 con-

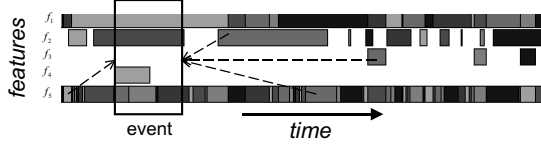


Figure 1: Feature based representation of a video document with an event (box) and contextual relations (dashed lines).

tent 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 during 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. This can be explained because of the time lag between actual occurrence of the event and its mentioning in the closed caption. Therefore, 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:

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

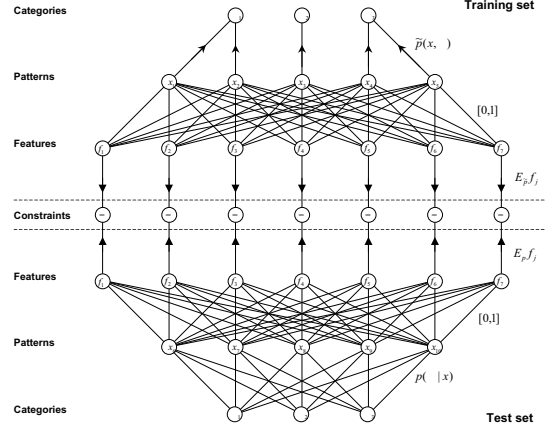


Figure 2: Simplified visual representation of the maximum entropy framework.

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. Two classifiers will be discussed, we start with the Maximum Entropy framework [6], followed by classification using a SVM [17].

3.1 Maximum Entropy

The Maximum Entropy framework starts with computation of the expected value for each feature in the training set \mathcal{S} :

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

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 $p(\omega|x)$:

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

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:

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

The model with maximum entropy, $p^*(\omega|x)$, should be selected. It is shown in [6] 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:

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

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) [8] algorithm. Since GIS relies on both $E_{\tilde{p}}(f_j)$ and $E_p(f_j)$ for calculation of α_j , an approximation proposed by [11] is used that relies only on $E_{\tilde{p}}(f_j)$ from \mathcal{S} . This allows to construct a classifier that depends on the training set only [5]. Hence, by using the maximum entropy classifier we can focus on what features to use, since relative importance of each feature is computed automatically.

3.2 Support Vector Machine

The SVM classifier follows another approach. Each pattern x is represented in a n -dimensional space, spanned by the features f_j . Within this feature space an optimal hyperplane is searched that separates the feature space into two different categories, ω , where the categories are represented by $+1$ and -1 respectively. Each hyperplane has the following form:

$$\omega_i(\mathbf{w} \cdot x_i + b) \geq 1, \forall i \quad (6)$$

where \mathbf{w} is a weight vector, and b is a threshold.

A hyperplane is considered optimal when the Euclidean distance to the closest training examples is maximum from both categories, this distance is called the margin. Consider the example in figure 3. Here a two-dimensional feature space consisting of two categories is visualized. The solid bold line is chosen as optimal hyperplane because of the largest possible margin. The circled data points closest to the optimal hyperplane are called the support vectors.

The problem of finding the optimal hyperplane is a quadratic programming problem of the following form [7]:

$$\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i \right\} \quad (7)$$

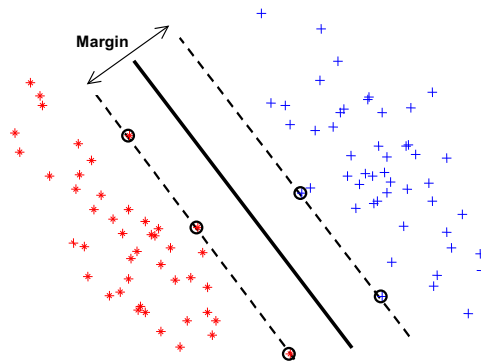


Figure 3: Visual representation of Support Vector Machine framework in two dimensions. The optimal hyperplane is indicated as a thick solid line.

Under the following constraints:

$$\begin{aligned} \omega_i(\mathbf{w}^T \phi(x_i) + b) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, i = 1, \dots, l. \end{aligned} \quad (8)$$

Where C is a parameter that allows to balance training error and model complexity, l is the number of patterns in \mathcal{S} , ξ_i are slack variables that are introduced when the data is not perfectly separable, and $\phi(\cdot)$ is the function that maps the x_j into the n -dimensional feature space. To account for the fact that the number of data in different categories can be unbalanced, we use an adaptation of (7) that allows to correct this [7].

4 Event indexing in soccer broadcasts

To demonstrate the viability of time interval modelling 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.

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.

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.

4.3 Visual features

From the visual modality we extracted several features. The type of camera work [3], i.e. pan, tilt, zoom-in, zoom-out or unknown, was computed for each camera shot. A face detector [15] 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 [9] combined with a marked watershed algorithm. The segmented region was the input for a Video Optical Character Recognition (VOCR) module [3], see figure 4. 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.



Figure 4: Different steps in overlay segmentation: color edge detection, marked watershed, and Video OCR.

4.4 Auditory features

From the auditory modality the excitement of the commentator is a valuable feature [14]. 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 the time interval modelling approach we digitized 8 live soccer broadcasts from Dutch national 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 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, $\omega \in \{\text{yellow card, substitution, goal}\}$, red card and penalty were excluded from analysis since there was only one instance of each in the data set. For evaluation of the ground truth, we manually labelled all the camera shots as either belonging to one of four categories: yellow card, goal, substitution, or unknown. Note that an event can cross camera shot boundaries. 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

	<i>Ground truth</i>		<i>Maximum Entropy</i>		<i>Support Vector Machine</i>	
	Total	Duration	Relevant	Duration	Relevant	Duration
<i>Goal</i>	12	3:07	10	10:14	11	11:52
<i>Yellow Card</i>	24	10:35	22	26:12	22	12:31
<i>Substitution</i>	29	8:09	25	7:36	25	7:23
Σ	65	21:51	57	44:02	58	31:46

Table 2: Evaluation results, where duration is the total duration of all segments that are retrieved. Duration is expressed in minutes:seconds.

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;

For each event both a Maximum Entropy and SVM classifier was trained. 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 in favor of an increase in false positives. For evaluation we compared the total duration of the retrieved events, including false positives, with a ground truth. 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 independent of the classifier used. The amount of video data that a user has to watch before finding those events is about two times longer when a Maximum Entropy classifier is used, and about one and a half times longer when a SVM is used, compared to the best case scenario. This is a considerable reduction of watching time when compared to the total duration, 9 hours, of all video documents in the test set. With the SVM we were able to detect one extra *goal*, compared to Maximum Entropy. Analysis of retrieved segments learned that results of Maximum Entropy and SVM are almost similar. Except for *goal* events, where nine events were retrieved by both, the other ones were different.

As noted in section 3.1, the GIS algorithm allows to compute the importance or weight of the different features used. The weights computed by GIS indicate that for *goal* and *yellow card* specific keywords in the closed captions, excitement with *during* and *overlaps* relations, a close-up afterwards, and the presence of an overlay nearby are important features. For *substitution* the auditory modality is less important.

When we analyze false positives for both classifiers, we observe that those are caused because some of the important features are shared between differ-

ent events. Mostly they occur when another event is indeed happening in the video, e.g. a hard foul or a scoring chance. False negatives are mainly caused by the fact that the feature detector failed. By incorporating more features in our model we might be able to reduce those false positives and negatives.

6 Conclusion

We combined multimodal information sources into a common framework for the purpose of event detection in video documents. The presented time interval modelling approach allows for proper modelling of events, synchronization, and asynchronous contextual information relations. Our method was evaluated on the domain of soccer, using different classifiers. Results show that a considerable reduction of watching time can be achieved. Moreover, the experiments indicate that SVM performs better than a Maximum Entropy classifier. The indexed events were used to build the Goalgle soccer video search engine, see figure 5. The indexed terms allow to query on events, matches, text, and persons. In the future we aim to experiment with other classifiers, e.g. a decision tree, and classifier combination schemes. Furthermore, we will also evaluate the time interval modelling approach on the news domain.



Figure 5: The Goalgle soccer video search engine.

Acknowledgements

The authors would like to thank Jan Baan from TNO-TPD and Jan-Mark Geusebroek from University of Amsterdam for their help with some features. This research is sponsored by the ICES/KIS MIA project and TNO Institute of Applied Physics (TPD).

References

- [1] M. Aiello, C. Monz, L. Todoran, and M. Worrington. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition*, 5(1):1–16, 2002.
- [2] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nenziati, and P. Pala. Soccer highlights detection and recognition using HMMs. In *IEEE International Conference on Multimedia & Expo*, Lausanne, Switzerland, 2002.
- [3] J. Baan, A. van Ballegooij, J-M. Geusebroek, D. Hiemstra, J. den Hartog, J. List, C. Snoek, I. Patras, S. Raaijmakers, L. Todoran, J. Vendrig, A. de Vries, T. Westerveld, and M. Worrington. Lazy users and automatic video retrieval tools in (the) lowlands. In *Proceedings of the 10th Text Retrieval Conference*, Gaithersburg, USA, 2001.
- [4] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, 2002.
- [5] J. Baldridge, T. Morton, and G. Bierner. *OpenNLP Maxent*, 2001. Software available at <http://maxent.sourceforge.net/>.
- [6] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
- [7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [8] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
- [9] J.M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.
- [10] M. Han, W. Hua, W. Xu, and Y. Gong. An integrated baseball digest system using maximum entropy method. In *ACM Multimedia*, Juan-les-Pins, France, 2002.
- [11] R. Lau, R. Rosenfeld, and S. Roukos. Adaptive language modelling using the maximum entropy approach. In *ARPA Human Language Technologies Workshop*, pages 81–86, Princeton, USA, 1993.
- [12] W.-H. Lin and A. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *ACM Multimedia*, Juan-les-Pins, France, 2002.
- [13] M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.
- [14] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV formula 1 programs. In *IEEE International Conference on Multimedia & Expo*, Lausanne, Switzerland, 2002.
- [15] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
- [16] C.G.M. Snoek and M. Worrington. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*. To appear.
- [17] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA, 1995.
- [18] D. Yow, B.-L. Yeo, M. Yeung, and B. Liu. Analysis and presentation of soccer highlights from digital video. In *Asian Conference on Computer Vision*, Singapore, 1995.