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.

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

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 $\omega$, 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-
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 \( j \) 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 \( j \) the camera shot. Note that the precedes and precedes \( j \) 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 \( \omega \) by means of binary fuzzy Allen relations with binary features, \( f_j \). Where each \( f_j \) is defined as:

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

Where \( \lambda_j(x) \) is a predicate function that checks for a fuzzy Allen relation, and \( \omega' \) 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 \( S \) is computed:

\[
E_p(f_j) = \sum_{x, \omega} \hat{p}(x, \omega)f_j(x, \omega)
\]

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

\[
E_p(f_j) = \sum_{x, \omega} \hat{p}(x)p(\omega|x)f_j(x, \omega)
\]

where \( \hat{p}(x) \) is the observed probability of \( x \) in \( S \). 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 \( T \) equals the expected value of \( f_j \) over \( S \). 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:

\[
H(p) = -\sum_{x, \omega} \hat{p}(x)p(\omega|x) \log p(\omega|x)
\]

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:

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

where \( \alpha_j \) is the weight for feature \( f_j \) and \( Z \) is a normalizing constant, used to ensure that a probability distribution results. The values for \( \alpha_j \) are computed by the Generalized Iterative Scaling (GIS) [5] algorithm. Since GIS relies on
Table 1: Features with fuzzy Allen time interval relations.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Fuzzy Allen</th>
<th>Range (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera work</td>
<td>during</td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td>during</td>
<td></td>
</tr>
<tr>
<td>Close-up</td>
<td>precedes, i</td>
<td>0 - 40</td>
</tr>
<tr>
<td>Goal keyword</td>
<td>precedes, j</td>
<td>0 - 6</td>
</tr>
<tr>
<td>Card keyword</td>
<td>precedes, j</td>
<td>0 - 6</td>
</tr>
<tr>
<td>Substitution keyword</td>
<td>precedes, j</td>
<td>0 - 6</td>
</tr>
<tr>
<td>Excitement</td>
<td>All relations</td>
<td>0 - 1</td>
</tr>
<tr>
<td>Info block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Person block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Referee block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Coach block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Goal block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Card block</td>
<td>precedes, j</td>
<td>20 - 50</td>
</tr>
<tr>
<td>Substitution block</td>
<td>during</td>
<td></td>
</tr>
<tr>
<td>Block length</td>
<td>during</td>
<td></td>
</tr>
</tbody>
</table>

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 [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...
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. 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.

![Figure 4: The Goalgle soccer video search engine.](http://example.com/figure4.png)

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


<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Maximum Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Duration</td>
</tr>
<tr>
<td>Goal</td>
<td>12</td>
</tr>
<tr>
<td>Yellow Card</td>
<td>24</td>
</tr>
<tr>
<td>Substitution</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
</tr>
</tbody>
</table>

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