Multimedia Event-Based Video Indexing Using Time Intervals

Cees G. M. Snoek, *Student Member, IEEE,* and Marcel Worring, *Member, IEEE*

***Abstract*—We propose the time interval multimedia event (TIME) framework as a robust approach for classification of semantic events in multimodal video documents. The representation used in TIME extends the Allen temporal interval relations and allows for proper inclusion of context and synchronization of the heterogeneous information sources involved in multimodal video analysis. To demonstrate the viability of our approach, it was evaluated on the domains of soccer and news broadcasts. For automatic classification of semantic events, we compare three different machine learning techniques, i.c. C4.5 decision tree, maximum entropy, and support vector machine. The results show that semantic video indexing results significantly benefit from using the TIME framework.**

***Index Terms*—Context, multimodal integration, semantic event classification, statistical pattern recognition, synchronization, time interval relations, video indexing.**

I. Introduction

MANAGEMENT of digital video documents is becoming more and more problematic due to the ever growing size of content produced. For easy management a semantic index describing the different events in the content of the document is indispensable. Since manual annotation is unfeasible, because of its tedious and cumbersome nature, automatic video indexing methods are necessary.

In general, automatic indexing methods suffer from the *semantic gap* or the lack of coincidence between the extracted information and its interpretation by a user, as recognized for image indexing in [1]. Video indexing has the advantage that it can profit from combined analysis of visual, auditory, and textual information sources. For multimodal indexing, two problems have to be unravelled. First, when integrating analysis results of different information channels, difficulties arise with respect to synchronization. The synchronization problem is typically solved by converting all modalities to a common layout scheme [2], e.g., camera shots, hereby ignoring the layout of the other modalities. This introduces the second problem, namely the difficulty to properly model context, i.e., how to include clues that do not occur at the exact moment of the semantic event of interest? When synchronization and context have been solved, multimodal video indexing might be able to bridge the semantic gap to some extent.

Manuscript received August 9, 2003; revised March 5, 2004. This work was supported by the ICES/KIS MIA project and by TNO. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. David Forsyth.

The authors are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands (e-mail: cgmsnoek@science.uva.nl).

Digital Object Identifier 10.1109/TMM.2005.850966

Existing methods for multimodal integration can be grouped into knowledge-based approaches [3], [4] and statistical approaches [5]–[9]. The former approaches typically combine the output of different multimodal detectors into a rule-based classifier. In [3], for example, the authors first analyze the textual channel for the occurrence of specific keywords that have a relation with a semantic event in American football. This results in a time interval where a possible event has taken place. The visual information of this time interval is then used for final classification. The drawback of this two stage approach is the dependence on the first stage. If the textual stream detector fails, no event is detected. To limit this model dependency, and improve the robustness, a statistical approach seems more promising. Various statistical frameworks can be exploited for multimodal integration. Recently there has been a wide interest in applying the dynamic Bayesian network (DBN) framework for multimodal integration [6], [8]. Other multimodal statistical frameworks that were proposed include the use of C4.5 decision trees [9], maximum entropy (MaxEnt) [5], and support vector machines (SVMs) [7]. However, all of these frameworks suffer from the problems of synchronization and context, identified above. Furthermore, they lack satisfactory inclusion of the textual modality. Therefore, a new framework is needed.

In this contribution we propose the time interval multimedia event (TIME) framework which explicitly handles context and synchronization and, as it is based on statistics, yields a robust approach for multimodal integration.

To demonstrate the viability of our approach for video indexing of semantic events we provide a systematic evaluation of three statistical classifiers, using TIME, and discuss their performance on the domains of soccer and news broadcasts. The soccer domain was chosen because events occur infrequently and in an unpredictable manner. Hence, contextual clues are important for reliable detection. In contrast to soccer, the news domain is far more structured. Here, synchronization of the different information sources is more important than context for accurate event detection.

The rest of this paper is organized as follows. First, we discuss related work, with respect to the domains we consider. Then we proceed with the introduction of the TIME framework in Section III, discussing both representation and classification. In Section IV, we discuss the detectors used for classification of various semantic events in soccer and news video. Experimental results are presented in Section V.

II. Related Work in Soccer and News Analysis

The classification methods introduced in the introduction have been used in various applications. For an extensive

overview we refer to [2]. We focus here on the soccer and news domain.

In literature several methods for automatic soccer analysis have been proposed, e.g., [10]–[13]. Most methods are based on analysis of the visual modality only. One of the first reported methods was presented in [13]. The authors focus on visualization of ball and player tracks using mosaics. However, no experiments in semantic event detection were demonstrated. More recently, methods were proposed that try to narrow the semantic gap based on a correlation between advanced visual detectors and semantic concepts. In [10] and [12], camera-based detectors are proposed, exploiting the relation between the movement of the ball and the camera. A slow-motion replay detector, among others, is proposed in [11] as a strong indicator for an event of importance that happened beforehand. For combination of the visual detectors a statistical DBN is used in [10], [12], whereas [11] exploits a knowledge-based approach.

In contrast to soccer event detection methods, which are still mainly based on visual analysis, the state-of-the-art in news analysis is already based on multimodal analysis [14]–[16], [7]. In [14], anchor shots and graphical shots are detected based on similarity and motion. The remaining shots are classified as news footage and are annotated with text extracted from a video optical character recognition module and a speech recognition module. A similar approach is proposed in [16], besides anchors, graphics, and report events, they detect gathering and walking events by exploiting face statistics. Manually added captions are processed with a named entity recognizer to attach more semantics to the detected events. By exploiting the fixed structure of a news broadcast in combination with similarity, motion, and audio detectors, the authors of [15] are able to detect anchors, monologues, report footage and weather forecasts. Weather reports are also detected in [7]; the authors combine text and image detectors and exploit combination strategies to improve classification accuracy. For the integration phase, again, a differentiation between knowledge-based [14], [16] and statistical methods [15], [7] can be made.

For both domains, problems arise when contextual information is to be included in the analysis and the various information sources have to be synchronized. In soccer for example, contextual clues like replays and distinguishing camera movement do not appear at the exact moment of the event, therefore the timing has to be estimated. In news, on the other hand, there is a clear relation between the visibility moment of overlayed text and the introduction of a speaker, i.e., it is unlikely that the overlay will appear at the end of the camera shot that views the speaker. Hence, their synchronization should be relative to each other. To tackle the problems of proper synchronization and inclusion of contextual clues for multimodal video analysis we propose the statistical TIME framework.

III. Multimedia Event Classification Framework

We view a video document from the perspective of its author [2]. Based on a predefined semantic intention, an author combines certain multimedia layout and content elements to express his message. For analysis purposes this authoring process should be reversed. Hence, we start with reconstruction

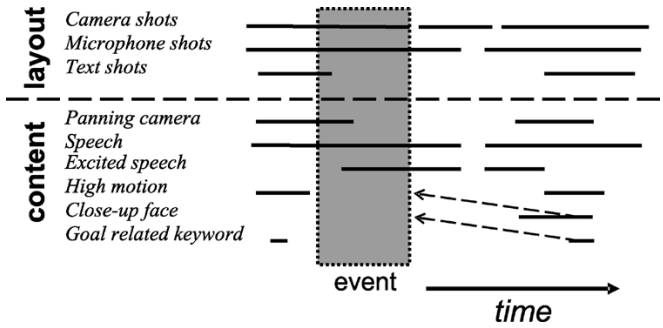


Fig. 1. Detector-based segmentation of a multimodal soccer video document into its layout and content elements with a goal event (box) and contextual relations (dashed arrows).

of layout and content elements. To that end, discrete detectors, indicating the presence or absence of specific layout and content elements, are often the most convenient means to describe the layout and content. This has the added advantage that detectors can be developed independently of one another. To combine the resulting detector segmentations into a common framework, some means of synchronization is required. To illustrate, consider Fig. 1. In this example, a soccer video document is represented by various time dependent detector segmentations, defined on different asynchronous layout and content elements. At a certain moment a goal occurs. Clues for the occurrence of this event are found in the detector segmentations that have a value within a specific position of the time-window of the event, e.g., excited speech of the commentator. But also in contextual detector segmentations that have a value before, e.g., a camera panning toward the goal area, or after the actual occurrence of the event, e.g., the occurrence of the keyword *score* in the time stamped closed caption. Clearly, in terms of the theoretical framework, it does not matter exactly what the detector segmentations are. What is important is that we need means to express the different visual, auditory, and textual detector segmentations into one fixed representation without loss of their original layout scheme.

Hence, for automatic classification of a semantic event ω, we need to grasp a video document into a common pattern representation. In this section, we first consider how to represent such a pattern, x, using multimodal detector segmentations and their relations, then we proceed with statistical pattern recognition techniques that exploit this representation for classification using varying complexity.

A. Pattern Representation

Applying layout and content detectors to a video document results in various segmentations, we define:

Definition 1 (TIME Segmentation): Decomposition of a video document into one or more series of time intervals τ, based on a set of multimodal detectors.

To model synchronization and context, we need means to express relations between these time intervals. Allen showed that 13 relationships are sufficient to model the relationship between any two intervals. To be specific, the relations are: *precedes*, *meets*, *overlaps*, *starts*, *during*, *finishes*, *equals*, and their inverses, identified by adding *_i* to the relation name [17]. For practical application of the Allen time intervals two problems

occur. First, in video analysis, exact alignment of start- or endpoints seldom occurs due to noise. Second, two time intervals will always have a relation even if they are far apart in time. To solve the first problem, a fuzzy interpretation was proposed in [18]. The authors introduce a margin T_1 to account for imprecise boundary segmentations, explaining the fuzzy nature. The second problem only occurs for the relations *precedes* and *precedes_i*, as for these the two time intervals are disjunct. Thus, we introduce a range parameter, T_2, which assigns to two intervals the type *NoRelation* if they are too far apart in time. Hence, we define the following.

Definition 2 (TIME Relations): The set of 14 fuzzy relations that can hold between any two elements from two segmentations, τ_1 and τ_2, based on the margin T_1 and the range parameter T_2.

Obviously the new relations still assure that between two intervals one and only one type of relation exists. The difference between standard Allen relations and TIME relations is visualized in Fig. 2.

Since TIME relations depend on two intervals, we choose one interval as a reference interval and compare this interval with all other intervals. Continuing the example, when we choose a camera shot as a reference interval, the goal can be modeled by a swift camera pan that *starts* the current camera shot, excited speech that *overlaps_i* the camera shot, and a goal-related keyword in the closed caption that *precedes_i* the camera shot within a range of 6 s. This can be explained because of the time lag between actual occurrence of the event and its mentioning in the closed caption. Although a panning camera, excited speech, and a goal-related keyword are possible important cues for a goal event, it is their combination with specific TIME relations that makes it key information with respect to the semantics. Also note that the interval-based TIME relations have a clear advantage over point-based representations, since the relative ordering of segmentations is preserved, and the relations do not suffer from variable lengths between various segmentations. Moreover, by combining TIME segmentations and TIME relations it becomes possible to represent events, context, and synchronization into one common framework. Hence, we define the following.

Definition 3 (TIME Representation): Model of a multimedia pattern x based on the reference interval τ_{ref}, and represented as a set of n TIME relations, with d TIME segmentations.

In theory, the number of TIME relations n is bounded by the number of TIME segmentations, d. Since, every TIME segmentation can be expressed as a maximum of 14 TIME relations with the fixed reference interval, the maximum number of TIME relations in any TIME representation is equal to $14(d-1)$. In practice, however, a subset can be chosen, either by feature selection techniques [19], experiments, or domain knowledge.

With the TIME representation we are able to combine layout and content elements into a common framework. Moreover, it allows for proper modeling of synchronization and inclusion of context as they can both be expressed as time intervals.

B. Pattern Classification

To learn the relation between a semantic event ω, and corresponding pattern x, we exploit the powerful properties of statistical classifiers. In standard pattern recognition, a pattern is represented by features. In the TIME framework a pattern is represented by related detector segmentations.

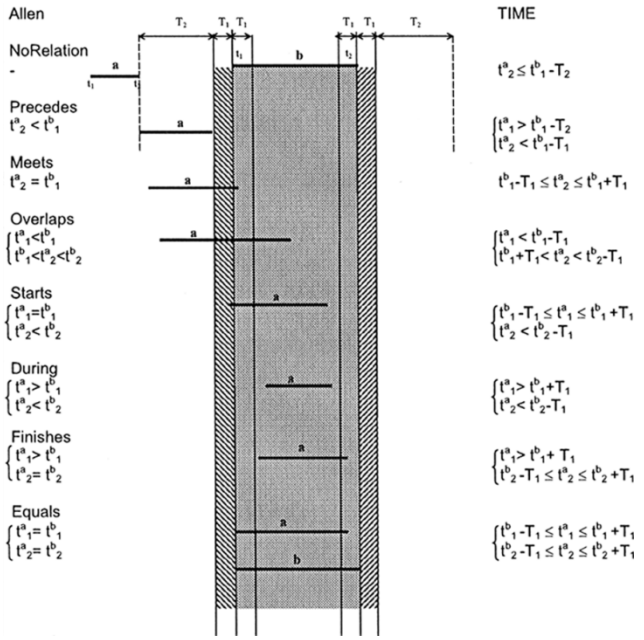


Fig. 2. Overview of the differences between exact Allen relations and TIME relations, extended from [18].

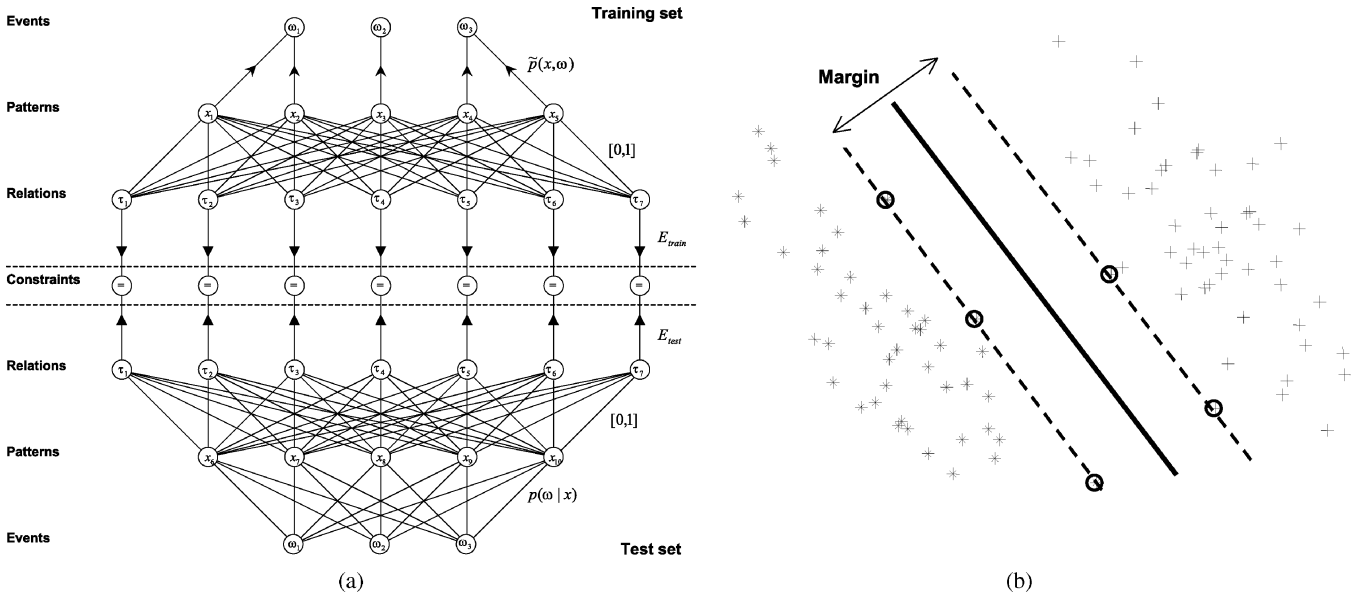
The statistical classification process is composed of two phases: training and testing. In the first phase, the optimal pattern configuration of relations is learned from the training data. In the second phase, the statistical classifier assigns the most probable event to a pattern based on the detected segmentations and their TIME relations. To prevent overtraining of the classifier, patterns in the testing phase should be drawn from an independent data set.

In literature, a varied gamut of statistical classifiers is proposed—see [19] for an excellent overview. For our purpose, classification of semantic events in video documents, a classifier should adhere to the following principles.

- *Binary representation*: since TIME relations are binary by default, the statistical classifier should be able to handle a binary pattern representation.
- *No independence assumption*: since there is a clear dependency between clues found in different modalities, a statistical classifier should not be based on an independence assumption.
- *Learn from few examples*: since the events of importance in a video can be limited, the statistical classifier should be able to learn from few examples.

Three statistical classifiers with varying complexity, adhering to the predefined principles, will be discussed. We start with the C4.5 decision tree [20], then we proceed with the MaxEnt framework [21], [22], and finally we discuss classification using a support vector machine (SVM) [23].

1) C4.5 Decision Tree: The C4.5 decision tree learns from a training set the individual importance of each TIME relation by computing the gain ratio [20]. Based on this ratio, a binary tree is constructed where a leaf indicates a class, and a decision node



(a) (b)

Fig. 3. (a) Simplified visual representation of the MaxEnt framework. Constraints, imposed by the relations, for the training set should be in accordance with those for the test set. From all possible models the one with MaxEnt is chosen. (b) Visual representation of the SVM framework. Here, a two-dimensional relation 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.

chooses between two subtrees based on the presence of some TIME relation. The more important a TIME relation is for the classification task at hand, the closer it is located near the root of the tree. Because the relation selection algorithm continues until the entire training set is completely covered, some pruning is necessary to prevent overtraining. Decision trees are considered suboptimal for most applications [19]. However, they form a nice benchmark for comparison with more complex classifiers and have the added advantage that they are easy to interpret.

2) MaxEnt: Whereas a decision tree exploits individual TIME relations in a hierarchical manner, the MaxEnt framework exploits the TIME relations simultaneously. In MaxEnt, first a model of the training set is created, by computing the expected value, E_{train}, of each TIME relation using the observed probabilities $\tilde{p}(x, \omega)$ of pattern and event pairs, [22]. To use this model for classification of unseen patterns, we require that the constraints for the training set are in accordance with the constraints of the test set. Hence, we also need the expected value of the TIME relations in the test set, E_{test} [22]. The complete model of training and test set is visualized in Fig. 3. We are left with the problem of finding the optimal reconstructed model, p^*, that finds the most likely event ω given an input pattern x, and that adheres to the imposed constraints. From all those possible models, the MaxEnt philosophy dictates that we select the one with the MaxEnt. It is shown in [22] that there is always a unique model $p^*(\omega \mid x)$ with MaxEnt, and that $p^*(\omega \mid x)$ has a form equivalent to

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

where α_j is the weight for TIME relation τ_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) [24] algorithm. Since GIS relies on both E_{train} and E_{test} for calculation of α_j, an approximation proposed by [25] is used that relies only on E_{train}. This allows to construct a classifier that depends completely on the training set. The automatic weight computation is an interesting property of the MaxEnt classifier, since it is very difficult to accurately weigh the importance of individual detectors and TIME relations beforehand.

3) SVM: The SVM classifier follows another approach. Each pattern x is represented in a n-dimensional space, spanned by the TIME relations. Within this relation space an optimal hyperplane is searched that separates the relation space into two different categories, ω, where the categories are represented by $+1$ and -1, respectively. The hyperplane has the following form: $\omega|(\mathbf{w} \cdot x + b)| \geq 1$, where $\mathbf{w}$ is a weight vector, and b is a threshold. A hyperplane is considered optimal when the distance to the closest training examples is maximum for both categories. This distance is called the margin—see the example in Fig. 3.

The problem of finding the optimal hyperplane is a quadratic programming problem of the following form [23]:

$$\min_{\mathbf{w},\xi} \left\{ \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\left(\sum_{i=1}^{l} \xi_i\right) \right\}. \tag{2}$$

Under the following constraints

$$\omega|(\mathbf{w} \cdot x_i + b)| \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l \tag{3}$$

where C is a parameter that allows to balance training error and model complexity, l is the number of patterns in the training set, and ξ_i are slack variables that are introduced when the data is not perfectly separable. These slack variables are useful when analyzing multimedia, since results of individual detectors typically include a number of false positives and negatives.

IV. Multimodal Video Analysis

We consider two domains for analysis, namely soccer and news. These domains were chosen because they allow to evaluate both the importance of context and proper synchronization.

Important events in a soccer game are scarce and occur more or less random. Examples of such events are goals, penalties, yellow cards, red cards, and substitutions. We define these events as follows.

- *Goal*: the entire camera shot showing the actual goal.
- *Penalty*: beginning of the camera shot showing the foul until the end of the camera shot showing the penalty.
- *Yellow card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the yellow card.
- *Red card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the red card.
- *Substitution*: beginning of the camera shot showing the player who goes out, until the end of the camera shot showing the player who comes in.

These events are important for the game and therefore the author adds contextual clues to make the viewer aware of the events. For accurate detection of events, this context should be included in the analysis.

In contrast to soccer, a news broadcast is far more structured. Each episode, the author carefully edits the layout and content elements, strictly adhering to the predefined format of events in the news show. Most important events in a news broadcast are the news stories. However, due to large variability in content, they are hard to model. Therefore, we focus on events that are more uniform in content and are useful for analysis of news structure. Examples of such events are reporting anchors, monologues, split-view interviews, and weather reports. We define these events as follows.

- *Reporting anchor*: the entire camera shot showing a news anchor talking to the camera.
- *Monologue*: the entire camera shot showing a single person, not a reporting anchor or weather reporter, talking for a while.
- *Split-view interview*: the entire camera shot showing both a news anchor and an on-site reporter in dialogue.
- *Weather report*: the entire camera shot showing a weather reporter talking about the weather forecast.

For analysis, the careful editing of the events should be taken into account by means of proper synchronization.

In this section, we will elaborate on the TIME segmentations and TIME relations used for both soccer and news analysis. Some of the detectors, used for the segmentation, are domain specific. It allows to integrate domain knowledge, but as these are learned and not strict they are more robust than domain knowledge captured in rules. Other detectors were chosen based on reported robustness and training experiments. The parameters for individual detectors were found by experimentation using the training set. Combining all TIME segmentations with all TIME relations results in an exhaustive use of relations, we therefore use a subset to prevent a combinatory explosion. The subset was tuned on the training set and exploits domain knowledge. For all events, all mentioned TIME segmentations and TIME relations are used, i.e., we used the same TIME representation for all events from the same domain. For both domains, we use a fixed value of 0.5 s for the margin T_1. We first discuss the soccer representation, and then proceed with the news representation.

A. Soccer Representation

The teletext (European 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.

On the visual modality we applied several detectors. The type of camera work [26]was computed for each camera shot, together with the shot length. A face detector [27]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, an author 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. We subdivided each detected overlay block as either info, person, referee, coach, goal, card, or substitution block [28], and added some additional statistics. For example, the duration of visibility of the overlay block, as we observed that substitution and info blocks are displayed longer on average. Note that all detector results are transformed into binary output before they are included in the analysis.

From the auditory modality, the excitement of the commentator is a valuable resource. For the proper functioning of an excitement detector, 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 s.

We take the result of automatic shot segmentation as a reference interval. An overview of the TIME representation for the soccer domain is summarized in Table I.

B. News Representation

The news events we want to classify are dominated by talking people. Most detectors that we propose are based on this observation. In the auditory modality we look for speech segments. This is simply achieved by using the previously discussed excitement detector with a lower threshold.

In the visual modality, we detected faces [27] and several derived statistics, like position, number, and camera distance used. We also detected the dominant camera work used during the shot, since the events we try to classify are typically shot using a static camera. For each shot, we furthermore computed the average motion, number of flashes, length, and whether it was preceded or succeeded by an effect. Text localization [26] was applied to detect regions of overlayed text. We differentiated between presence of a single region and parallel regions, e.g., one in the top of the image frame and on the bottom.

TABLE I
TIME REPRESENTATION FOR SOCCER ANALYSIS. T_2 INDICATES THE CONTEXTUAL RANGE USED BY THE PRECEDES AND PRECEDES_I RELATIONS

TIME segmentation	*TIME relations*	T_2 *(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 statistics	*precedes_i*	20 - 80
Person block statistics	*precedes_i*	20 - 50
Referee block statistics	*precedes_i*	20 - 50
Coach block statistics	*precedes_i*	20 - 50
Goal block statistics	*precedes_i*	20 - 50
Card block statistics	*precedes_i*	20 - 50
Substitution block statistics	*during*	
Shot length	*during*	

TABLE II
TIME REPRESENTATION FOR NEWS ANALYSIS. T_2 INDICATES THE CONTEXTUAL RANGE USED BY THE PRECEDES AND PRECEDES_I RELATIONS

TIME segmentation	*TIME relations*	T_2 *(s)*
Camera work	*during*	
Effect	*precedes, precedes_i*	0 - 4
Block length	*during*	
Camera distance	*during*	
Face left	*during*	
Face right	*during*	
Face center	*during*	
Number of faces	*during*	
Number of flashes	*during*	
Kinetic Energy	*during*	
Speech	*All relations*	0 - 1
Closed caption	*All relations*	0 - 1
Overlayed text	*All relations*	0 - 1
Parallel overlayed text	*All relations*	0 - 1
Studio keyword	*during*	
Weather keyword	*during*	

For each detected text region, we recognized the text and tried to match it, using fuzzy string matching, with the city name where the news studio is located. The presence of closed caption segments was used as an additional indicator for speech. Moreover, they were scanned for presence of weather related keywords like *sunny, snow, degree, west,* and so on.

Again, we take the result of automatic shot segmentation as a reference interval. The TIME representation for the news domain is summarized in Table II. When comparing both Tables I and II, one can clearly see that Table I includes more context, whereas Table II is more concerned with synchronization. In the next section, we will evaluate the automatic indexing of events in soccer and news video, based on the presented pattern representation.

V. RESULTS

For the evaluation of the TIME framework, we used soccer and news broadcasts from Dutch national TV. We recorded eight live soccer broadcasts, about 12 h in total. The videos were digitized in 704×576 resolution MPEG-2 format. For the news domain, we recorded 24 broadcasts, again about 12 hours in total, in 352×288 resolution MPEG-1 format. The audio was sampled at 16 kHz with 16 bits per sample for both domains. The time stamped teletext was recorded with a teletext receiver. For soccer analysis we used a representative training set of 3 h and a test set of 9 h. For news, a training and test set of 6 h each was used. In this section, we will first present the evaluation criteria used for evaluating the TIME framework, then we present the classification results obtained. After presenting two prototype systems, we end with a discussion on the results.

A. Evaluation Criteria

The standard measure for performance of a statistical classifier is the error rate. However, this is unsuitable in our case, since the amount of relevant events are outnumbered by irrelevant pieces of footage. We therefore use the precision and recall measure adapted from information retrieval. Let $|R|$ be the number of relevant camera shots, i.e., camera shots containing the specific event one is looking for. Let $|A|$ denote the answer set, i.e., the number of camera shots that are retrieved by the system. Let $|R \cap A|$ be the number of camera shots in the intersection of the sets R and A. Then, precision is the fraction of retrieved camera shots (A) which are relevant

$$\text{Precision} = \frac{|R \cap A|}{|A|} \tag{4}$$

and recall is the fraction of the relevant camera shots (R) which have been retrieved

$$\text{Recall} = \frac{|R \cap A|}{|R|}. \tag{5}$$

This measure gives an indication of correctly classified events, falsely classified events, and missed events. For the evaluation of news classification, results will be plotted in a precision-recall curve.

For the evaluation of soccer we used a different approach. Since events in a soccer match can cross camera shot boundaries, we merge adjacent camera shots with similar labels. As a consequence, we loose our arithmetic unit. Therefore, precision and recall can no longer be computed. As an alternative for precision, we relate the total duration of the segments that are retrieved to the total duration of the relevant segments. Moreover, since it is unacceptable from a users perspective that scarce soccer events are missed, we strive to find as many events as possible in favor of an increase in false positives. Finally, because it is difficult to exactly define the start and end of an event in soccer video, we introduce a tolerance value T_3 (in seconds) with respect to the boundaries of detection results. We used a T_3 of 7 s. for all soccer events. A merged segment is considered relevant if one of its boundaries plus or minus T_3 crosses that of a labeled segment in the ground truth.

Besides a comparison of individual classifiers, we also compare the influence of TIME on the final result. Since the benefit

TABLE III
EVALUATION RESULTS OF THE DIFFERENT CLASSIFIERS FOR SOCCER EVENTS, WHERE DURATION IS THE TOTAL DURATION OF ALL SEGMENTS THAT ARE RETRIEVED

	Ground truth		*C4.5*		*MaxEnt*		*SVM*	
	Total	Duration	Relevant	Duration	Relevant	Duration	Relevant	Duration
Goal	12	3^m07^s	2	2^m40^s	10	10^m14^s	11	11^m52^s
Yellow Card	24	10^m35^s	13	14^m28^s	22	26^m12^s	22	12^m31^s
Substitution	29	8^m09^s	25	15^m27^s	25	7^m36^s	25	7^m23^s
$\sum$	65	21^m51^s	40	32^m35^s	57	44^m02^s	58	31^m46^s

of using TIME for domains relying on context is obvious, we only show this result for the news domain.

B. Event Classification

For evaluation of TIME on the soccer domain, we manually labeled all the camera shots as either belonging to one of four categories: yellow card, goal, substitution, or unknown. Red card and penalty were excluded from analysis since there was only one instance of each in the data set. For all three remaining events, a C4.5, MaxEnt, and SVM classifier[1] was trained. Results on the test set are visualized in Table III.

When analyzing the results, we clearly see that the C4.5 classifier performs worst. Although it does a good job on detection of substitutions, it is significantly worse for both yellow cards and goals when compared to the more complex MaxEnt and SVM classifiers. When we compare results of MaxEnt and SVM, we observe that almost all events are found independent of the classifier used. The amount of video data that a user has to watch before finding these events is about two times longer when a MaxEnt classifier is used, and about one and a half times longer when an SVM is used, compared to the best case scenario. This is a considerable reduction of watching time when compared to the total duration, 9 h, of all video documents in the test set. With the SVM we were able to detect one extra goal, compared to MaxEnt. Analysis of retrieved segments learned that results of MaxEnt and SVM are almost similar. Except for goal events, where nine events were retrieved by both, the remaining classified goals were different for each classifier.

For the news domain, we used the same classification approach as for soccer. But we are now focusing on four events, namely: reporting anchor, monologue, split-view interview, and weather report. Again for each event a C4.5, MaxEnt, and SVM classifier was trained. Moreover, we also compared the added value of TIME by inclusion of one run with the SVM classifier where all TIME relations were replaced by *during* relations.

Results of news classification are visualized by means of precision-recall curves in Fig. 4. For the MaxEnt classifier, we varied the threshold on the likelihood for each camera shot computed by (1). For SVM, we varied the threshold on the margin computed by (2) for each camera shot. For C4.5 this is impossible because of its binary nature, we therefore plotted results of five pruning values. When comparing classification results of the different classifiers we observe that SVM outperforms all other classifiers, and that C4.5 achieves comparable classification results when compared with a MaxEnt classifier. MaxEnt performs better on monologues, C4.5 performs better on weather reports, and is even comparable to SVM for this event. The experimental results of SVM with and without TIME clearly show that there is a significant gain in classification results when using the TIME framework. Only for classification of weather report events, an SVM classifier without TIME can achieve comparable results as an SVM with TIME. For all other classes, it is outperformed by the SVM with TIME.

C. Implementation

Based on the current classification result, we have developed the *Goalgle* soccer video search engine[2] and added functionality to the *News RePortal* system, see Fig. 5. In its current form, the web-based prototypes allow to query a selection of broadcasts on keywords, persons and events. Ultimately, this should result in a personalized automatic summary that can be presented on a wide range of pervasive devices.

D. Discussion

When we take a closer look to the individual results of the different classifiers, it is striking that C4.5 can achieve a good result on some events, e.g., substitution and weather report, while performing bad on others, e.g., goal and monologue. This can, however, be explained by the fact that the events where C4.5 scores well, can be detected based on a limited set of TIME relations. For substitution events in soccer, an overlay during the event is a very strong indicator, whereas a weather related keyword in the teletext is very indicative for weather reports. When an event is composed of several complex TIME relations, like goal and monologues, the relatively simple C4.5 classifier performs worse than both complex MaxEnt and SVM classifiers.

To gain insight in the meaning of complex relations in the two domains, we consider the GIS algorithm from Section III-B2, which allows to compute the importance or relative weight of the different relations used. The weights computed by GIS indicate that for the soccer events 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 relations. For the news events reporting anchor and monologue, a close-up face on the left side during the shot, a low average motion during the shot, and overlayed text during the shot were of equal importance. For

[1]For classification, the following open-source toolboxes were used.
S. Ruggieri. *YaDT – Yet another Decision Tree builder*.
J. Baldridge, T. Morton and G. Bierner. *OpenNLP Maxent*.
C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*

[2][Online] Available: http://www.goalgle.com

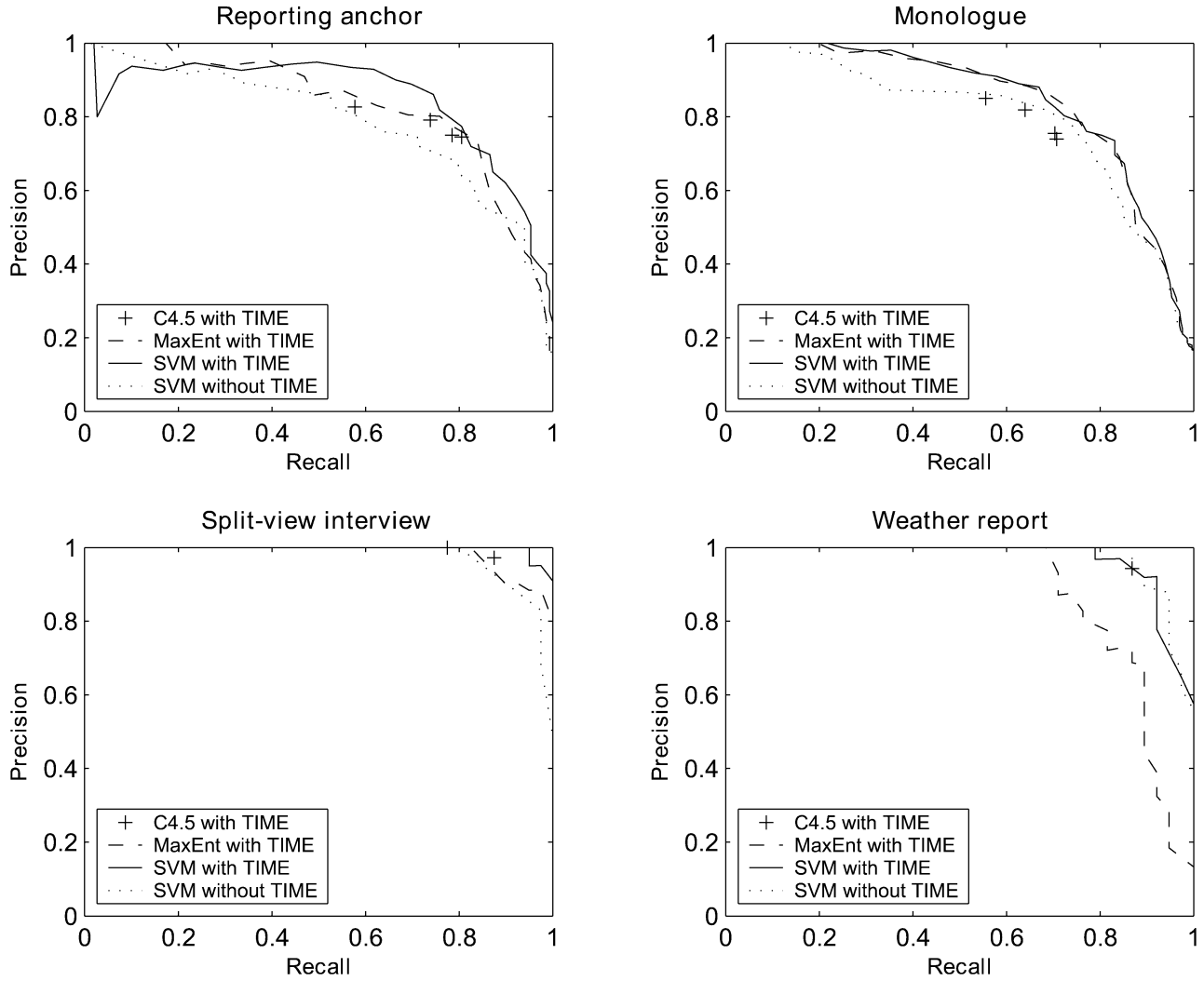
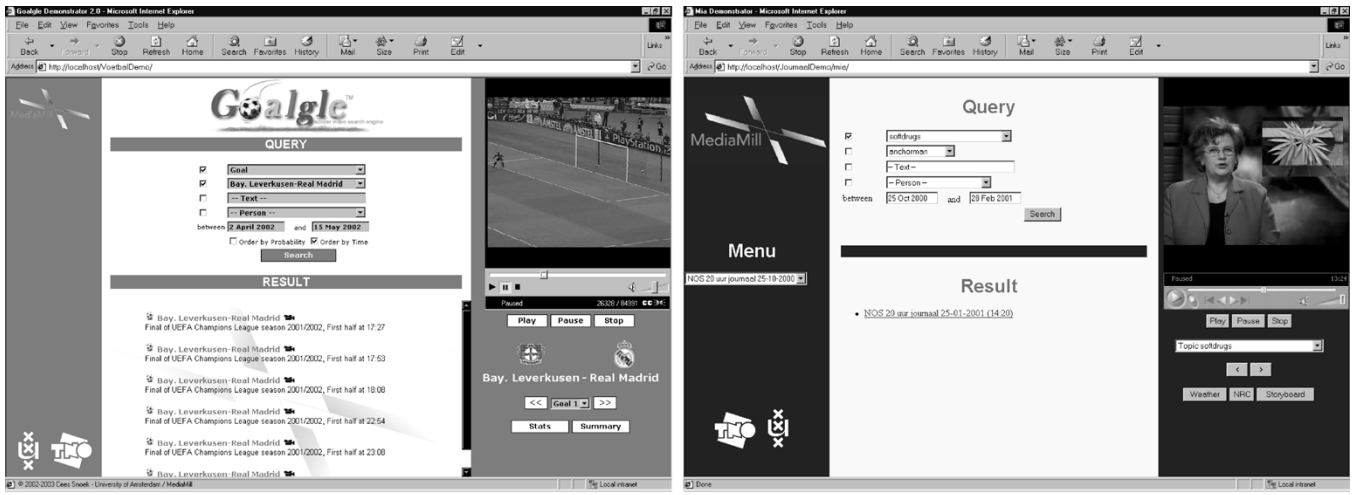


Fig. 4. Precision-recall curves for different semantic events in news broadcasts.



(a) (b)

Fig. 5. Screen dumps of (a) the Goalgle soccer video search engine and (b) the News RePortal system.

reporting anchors, speech that starts the camera shot was important, whereas various relations with overlayed text were important for monologues. The weights for the speech relation for monologues were not high enough to consider it very important, which is quite surprising. This can be explained by the fact that non-Dutch speakers are transcribed by means of overlayed text

in the Dutch news, hence the detection of such overlayed text is more distinguishing than speech for monologues. For split-view interview events, two faces during the camera shot, meets and equals relations with overlayed text showing the location of the two speakers, overlapping and during speech relations, and the identification of a city keyword in the overlay text were important. For weather reports, besides keywords in the teletext, a long shot camera distance during the camera shot, and overlayed text with start and finish relations are of importance.

When combining the weights, MaxEnt sometimes fails to profit from multiple information sources. This is best observed in the precision-recall curve for weather reports. Overall, the SVM classifier achieves comparable or better results than MaxEnt. When we analyze false positives for both classifiers, we observe that these are caused because some of the important relations are shared between different events. For soccer this mostly occurs when another event is indeed happening in the video, e.g., a hard foul or a scoring chance. For news this especially occurs for classification of reporting anchors and monologues. Often a monologue is classified as anchor and vice versa. We also found that close-ups of people in report footage with voice-overs, and reporting anchor's that were filmed from less usual camera positions were often falsely classified. False negatives are mainly caused by the fact that a detector failed. By increasing the number of detectors and relations in our model, we might be able to reduce these false positives and false negatives. Another option is to use a cascade of classifiers, so instead of classifying each event individually, first classify events on which you can do a good job, e.g., split-view interviews, and apply another classifier on the negative results of the first classifier, and so on. This should yield better indexing results.

VI. Conclusion

To bridge the semantic gap for multimedia event classification, a new framework is required that allows for proper modeling of context and synchronization of the heterogeneous information sources involved. We have presented the TIME framework that accommodates these issues, by means of a time interval-based pattern representation. Moreover, the framework facilitates robust classification using various statistical classifiers.

To demonstrate the effectiveness of TIME it was evaluated on two domains, namely soccer and news. The former was chosen because of its dependency on context, the latter because of its dependence on synchronization. We have compared three different statistical classifiers, with varying complexity, and show that there exists a clear relation between narrowness of the semantic gap and the needed complexity of a classifier. When there exists a simple mapping between a limited set of relations and the semantic concept we are looking for, a simple decision tree will give comparable results as a more complex SVM. When the semantic gap is wider, detection will profit from combined use of multimodal detector relations and a more complex classifier, like the SVM. Moreover, we show that the TIME framework, including synchronization and context, outperforms the "standard" multimodal analysis approaches common in video indexing literature.

In the future, we aim to explore the usage of complex classifier combinations and architectures. Moreover, by inclusion of more textual resources we expect to be able to give a richer description of events in video, ultimately bridging the semantic gap for a large set of events.

Acknowledgment

The authors are grateful to J. Baan from TNO and J.-M. Geusebroek from the University of Amsterdam for their help with some of the detectors. H. Wedemeijer from TNO is acknowledged for developing the news demonstrator, which also formed the core of the Goalgle system.

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[2] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools Applicat.*, vol. 25, no. 1, pp. 5–35, 2005.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 68–75, Mar. 2002.

[4] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proc. ACM Multimedia*, San Francisco, CA, 1995, pp. 295–304.

[5] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Multimedia*, Juan-les-Pins, France, 2002.

[6] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video scene classification based on HMM," in *Proc. IEEE Workshop Multimedia Signal Processing*, Copenhagen, Denmark, 1999.

[7] W.-H. Lin and A. G. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *Proc. ACM Multimedia*, Juan-les-Pins, France, 2002.

[8] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 141–151, Mar. 2001.

[9] W. Zhou, S. Dao, and C.-C. J. Kuo, "On-line knowledge- and rule-based video classification system for video indexing and dissemination," *Inform. Syst.*, vol. 27, no. 8, pp. 559–586, 2002.

[10] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," in *Proc. IEEE Int. Conf. Multimedia & Expo.*, Lausanne, Switzerland, 2002.

[11] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[12] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia*, vol. 9, no. 2, pp. 44–51, Apr.–Jun. 2002.

[13] D. Yow, B.-L. Yeo, M. Yeung, and B. Liu, "Analysis and presentation of soccer highlights from digital video," in *Proc. Asian Conf. Computer Vision*, Singapore, 1995.

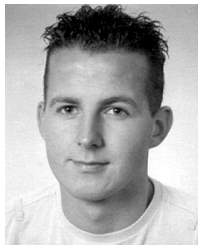
[14] M. Bertini, A. Del Bimbo, and P. Pala, "Indexing for reuse of TV news shots," *Pattern Recognit.*, vol. 35, no. 3, pp. 581–591, 2002.

[15] S. Eickeler and S. Müller, "Content-based video indexing of TV broadcast news using hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Phoenix, AZ, 1999, pp. 2997–3000.

[16] I. Ide, K. Yamamoto, and H. Tanaka, "Automatic video indexing based on shot classification," in *Proc. 1st Int. Conf. Advanced Multimedia Content Processing*, vol. 1554, Lecture Notes in Computer Science, Osaka, Japan, 1999, pp. 87–102.

[17] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, no. 11, pp. 832–843, 1983.

[18] M. Aiello, C. Monz, L. Todoran, and M. Worring, "Document understanding for a broad class of documents," *Int. J. Doc. Anal. and Recognit.*, vol. 5, no. 1, pp. 1–16, 2002.
[19] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[20] J. R. Quinlan, *C4.5: Programs for Machine Learning*. New York: Morgan Kaufmann, 1993.
[21] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.
[22] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computat. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.
[23] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2000.
[24] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Statist.*, vol. 43, no. 5, pp. 1470–1480, 1972.
[25] R. Lau, R. Rosenfeld, and S. Roukos, "Adaptive language modeling using the maximum entropy approach," in *Proc. ARPA Human Language Technologies Workshop*, Princeton, NJ, 1993, pp. 81–86.
[26] J. Baan *et al.*, "Lazy users and automatic video retrieval tools in (the) lowlands," in *Proc. 10th Text Retrieval Conf.*, Gaithersburg, MD, 2001.
[27] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[28] C. G. M. Snoek and M. Worring, "Time interval maximum entropy based event indexing in soccer video," in *Proc. IEEE Int. Conf. Multimedia & Expo.*, vol. 3, Baltimore, MD, 2003, pp. 481–484.



Cees G. M. Snoek (S'01) received the M.Sc. degree in business information systems in 2000 from the University of Amsterdam, Amsterdam, The Netherlands, where he is pursuing the Ph.D. degree in computer science.

Since January 2001, he has been a Research Assistant at the University of Amsterdam. He was a Visiting Scientist at Informedia, Carnegie Mellon University, Pittsburgh, PA, in 2003. His research interests focus on multimedia signal processing and analysis, statistical pattern recognition, and content-based information retrieval, especially when applied in combination for the purpose of semantic multimedia understanding.

Marcel Worring (M'03) received the M.Sc. degree (Hons.) and Ph.D. degree, both in computer science, from the Free University Amsterdam, Amsterdam, The Netherlands, in 1988 and the University of Amsterdam in 1993, respectively.

He is currently an Associate Professor at the University of Amsterdam. His interests are in multimedia information analysis and systems. He leads several multidisciplinary projects covering knowledge engineering, pattern recognition, image and video analysis, and information space interaction, conducted in close cooperation with industry. In 1998, he was a Visiting Research Fellow at the University of California, San Diego. He has published over 50 scientific papers and serves on the program committee of several international conferences.