Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events

M. Petkovic, W. Jonker
Computer Science Department, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands,
Email {milan, jonker}@cs.utwente.nl

Abstract

As amounts of publicly available video data grow, the need to query this data efficiently becomes significant. Consequently, content-based retrieval of video data turns out to be a challenging and important problem. In this paper, we address the specific aspect of inferring semantics automatically from raw video data. In particular, we introduce a new video data model that supports the integrated use of two different approaches for mapping low-level features to high-level concepts. Firstly, the model is extended with a rule-based approach that supports spatio-temporal formalization of high-level concepts, and then with a stochastic approach. Furthermore, results on real tennis video data are presented, demonstrating the validity of both approaches, as well as advantages of their integrated use.

1. Introduction

An increasing number of large video libraries that is becoming publicly available nowadays, results in a demand for techniques that can manipulate the video data based on content. However, traditional database management systems based on a relational or object-oriented data model still do not provide enough facilities for managing and retrieving video contents. As pointed out in [1], three main reasons can be distinguished for this: (1) lack of facilities for the management of spatio-temporal relations, (2) lack of knowledge-based methods for interpreting raw data into semantic contents, and (3) lack of query representations.

This paper addresses these problems with the emphasis on the second one - recognizing semantic content in video data based on visual features. First, we propose a layered video data model that provides a framework for automatic mapping from features to concepts. The model is independent of feature/semantic extractors, providing flexibility in using different video processing and pattern recognition techniques for those purposes. At same time the model is in line with the latest development in MPEG7, differencing video content between diverse categories. Next, we extend the model with object and event grammars. These grammars are aimed at formalizing descriptions of high-level concepts, as well as facilitating their extraction based on features and spatio-temporal reasoning. On the other hand, the model also provides a framework for stochastic modeling of events. Here, we exploit the learning capability of Hidden Markov Models to recognize events in video data automatically. Each approach in isolation, as well as the integrated approach, is validated for the retrieval in the particular domain of tennis matches.

2. State of the art in video retrieval

A rough categorization of the video retrieval approaches from the literature (see [1-3] for review) yields two main classes.

The first class focuses mainly on visual features, such as color histograms, shapes, textures, or motion, which characterize the low-level visual content. Although these approaches use automatically extracted features, representing the video content, they do not provide semantics that describe high-level video concepts, which is much more appropriate for users when retrieving video segments.

The second class concerns annotation-based approaches (such as [4] for example), which use free-text, attribute, or keyword annotation to represent the high-level concepts of the video content. However, this results in many drawbacks. The major limitation of these approaches is that the search process is based solely on the predefined attribute information, which is associated with video segments in the process of annotation. Furthermore, manual annotation is tedious, subjective and time consuming.

Obviously, the main gap lies between low-level media features and high-level concepts. However, as video is a temporal sequence of pixel regions at the physical level, it is very difficult to explore its semantic content. In order to solve this problem, several domain-dependent research

0-7695-1293-3/01 $10.00 © 2001 IEEE

efforts have been undertaken. These approaches take an advantage of using domain knowledge to facilitate extraction of high-level concepts directly from features. In particular, they mainly use information on object positions, their transitions over time, etc., and relate them to particular events (high-level concepts). For example, methods have been proposed to detect events in football [5], soccer [6], and tennis games [7, 8], hunting [9], etc. Motion (for review see [10]) and audio are, in isolation, very often used for event recognition. In [11] for example, extracting highlights from baseball games is based on audio only. Although these efforts resulted in the mapping from features to high-level concepts, still there is a problem of creating the mapping for each domain manually. Furthermore, this might be very difficult, requiring expert knowledge, especially in case of some complex events. In addition, many of these methods are not extensible for detecting new events because they are very dependent on specific artifacts used in the broadcasts of domain programs.

On the other hand, some other approaches use stochastic methods that often exploit automatic learning capabilities to derive knowledge, such as Hidden Markov Models (HMMs), Bayesian belief networks, etc. The first publication addressing recognition of human actions using HMMs [12] describes the application of discrete HMMs in recognizing six different tennis stroke classes in a constrained test environment. Recently, similar techniques have been proposed. Naphade et al [13] used hierarchical HMMs to extract events like explosions. HMMs together with a Bayesian classifier have been used for recognition of human actions in [14]. Structuring of video using Bayesian Networks alone [15] or together with HMMs [16] has been proposed. In [17] a probabilistic model has been used to combine results of visual and audio event detection in order to identify topics of discussion in a classroom lecture environment. Another probabilistic framework that comprises multimedia objects within a Bayesian multinet has been proposed in [18].

The main advantage of our approach compared to the ones mentioned above is providing a framework for integrating deterministic and stochastic approaches. Most importantly, it addresses the problem from the database point of view, allowing users to dynamically create and define new multimedia objects and events.

3. The COBRA video model

In order to overcome the problem of mapping features to high-level concepts we propose the COBRA (COntent-Based RetrievAl) video data model (Fig. 1). It consists of four layers: the raw data, feature, object, and event layer.

The raw video data layer is at the bottom. This layer consists of a sequence of frames, as well as some video

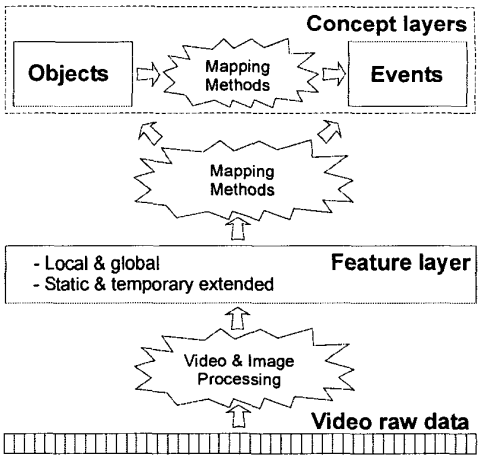


Fig. 1. The layered hierarchy of the COBRA video data model

attributes, such as compression format, frame rate, number of bits per pixel, duration, etc.

The next layer is the feature layer consisting of domain-independent features that can be automatically extracted from raw data, characterizing colors, textures, shapes, and motion. They can be classified as global features that characterize a whole frame, and local features that characterize a region inside a frame.

The object layer consists of entities (logical concepts), characterized by prominent spatial dimension and assigned to regions across frames. A region is a contiguous set of pixels that is homogeneous in texture, color, shape, and motion properties. We define a video object as a collection of regions, which have been grouped together by some criteria defined by the domain knowledge. An object should also satisfy some conditions, such that it is semantically consistent, representing one real-world object, and subject of interest to users or applications. Some examples of video objects are a specific player or the ball in a tennis game or a specific car in a car-race video.

The event layer consists of entities that have a prominent temporal extent, describing the movements and interactions of different objects in a spatio-temporal manner.

The overall structure of our video data model is as follows:

$$V = (\Sigma, A, RVD, F, O, E, S)$$

The signature Σ is defined as a tuple, $\Sigma = (\Sigma_F, \Sigma_O, \Sigma_E, \Sigma_S)$. Σ_F is a set of feature types, Σ_O a set of object types, Σ_E a set of event types, and Σ_S is a set of audio segment types (audio clusters). A is a set of video attributes, such as frame sample rate, audio sample rate, file format, name, creation date, etc.

Raw video data (RVD) is represented by a set of frames and regions. Each frame can comprise zero or more regions (R). Therefore, the RVD set has elements of two types: *i* and *i.Rj*, where integer *i* points to a specific video frame and *i.Rj* to its specific region denoted by integer *j*.

F is a set of video features. Each feature is described by its type descriptor ($\in \Sigma_F$), its value, and a reference to a frame or a region in a frame. Possible types of features include but are not limited to the ones defined in [19, 20].

O is a set of objects. An object is described by its type ($\in \Sigma_O$), identification, a set of regions that compose it, their features, as well as history of its geometry (minimum bounding rectangle) across frames. Video objects are related to particular real-world objects that are modeled in the database.

E is a set of events. An event is defined by its type, identification, a set of object types involved, and the time interval it spans. Hence, event types are parameterized by using different object instances.

An audio track, as one of essential video components, provides a very rich source of information to supplement understanding of a video. Combining audio with other video components can provide much more information than any media alone. Therefore, we integrated audio primitives in the model to provide additional information that might be critical to the perception and understanding of video content. The raw audio data can be divided into speech and non-speech parts through the process of segmentation and classification. A time-aligned transcript of the spoken words can be created using speech recognition, while semantic segmentation based on text can be used for obtaining the meaningful speech segments. The non-speech segments are clustered and associated with textual description. The set S is a set of both types of audio segments.

4. Spatio-temporal formalization of events

In order to facilitate automatic extraction of concepts (objects and events) from visual features, the model has been extended with object and event grammars, which are aimed at formalizing descriptions of these high-level concepts. Therefore, we introduce the following three elements: G_O, G_E, and α.

G_O is an object grammar. It defines the syntax for rules that are used for object type descriptions. Object types can be primitive or compound. If we look at the soccer domain for example, a goalpost can be seen as a primitive object type that is composed of two regions, i.e. white bars and a net. These regions are homogeneous in certain features and several spatial relations exist between them. Despite objects are defined as entities with a prominent spatial dimension, we take advantage of temporality allowing the usage of temporal relations in the rules for object descriptions. Compound objects consist of two or more primitive ones. The rules are defined as follows:

$$\rho_{\text{Oprimitive}}: (2^{\Sigma_{RVD}}, 2^{\Sigma_S}, 2^{\varphi}, 2^{\sigma}, 2^{\tau}) \rightarrow \Sigma_O$$
$$\rho_{\text{Ocompound}}: (2^{\Sigma_O}, 2^{\varphi}, 2^{\sigma}, 2^{\tau}) \rightarrow \Sigma_O$$

The sets of feature operators, spatial and temporal relations (φ, σ, and τ respectively), which are used in the object rules, will be explained later in this section. Instead of the object grammar, it is also possible to use other techniques for object recognition: external extractors for specific objects (e.g. [21]), or MPEG-4.

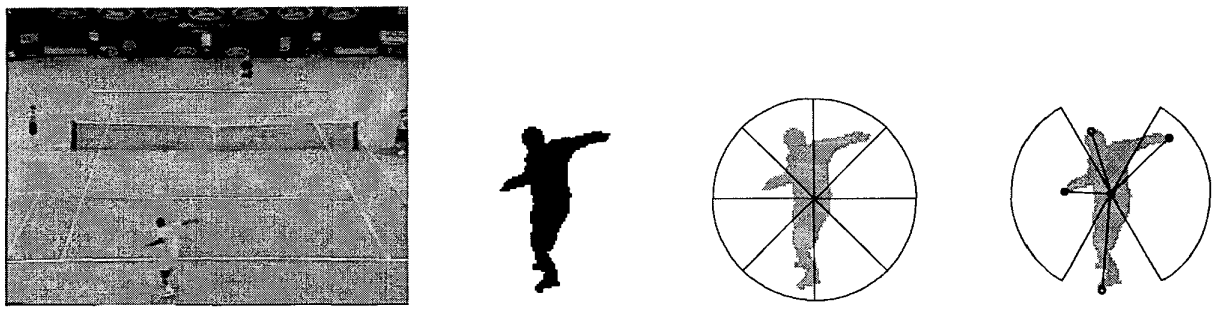
G_E is an event grammar. It consists of rules that are used for event type descriptions. Similar to the object types, event types can be primitive or compound. Two rules for primitive event types exist. The first one defines events using visual features and their spatio-temporal and similarity relations, whereas the second one uses object types instead of features, together with their real-world relations (ω). Audio segment types may also be included. Hence, it is possible to define audio events, and compound audio-visual events. On the other hand, a compound event type is described by a power set of predefined event types, their temporal relations, as well as real-world and spatial relations among their objects. The event grammar rules are defined as follows:

$$\rho_{\text{Eprimitve\_RVD}}: (2^{\Sigma}, 2^{\Sigma}, 2^{\varphi}, 2^{\sigma}, 2^{\tau}) \rightarrow \Sigma_E$$
$$\rho_{\text{Eprimitve\_O}}: (2^{\Sigma_O}, 2^{\Sigma}, 2^{\omega}, 2^{\sigma}, 2^{\tau}) \rightarrow \Sigma_E$$
$$\rho_{\text{Ecompound}}: (2^{\Sigma_E}, 2^{\omega}, 2^{\sigma}, 2^{\tau}) \rightarrow \Sigma_E$$

The last element in this extension is a set of algebraic operators, denoted by α. It includes sets of spatial, temporal, feature, real-world and so-called video operators, i.e. $\alpha = \{\sigma, \tau, \varphi, \omega, \upsilon\}$.

As far as spatial relations are concerned, we use the Minimum Bounding Rectangle (MBR) approximation of the object geometry to increase efficiency. Based on this approximation, we implemented fundamental topological (equal, inside, cover, overlap, touch, disjoint, and two inverse covered_by and contains) and directional relations (north, south, west, east, north-east, north-west, south-east and south-west), as well as the Euclidean distance relation. For the definitions of these relations, as well as interval temporal relations see [22, 23].

As far as temporal relations are concerned, we implemented basic relations of interval algebra (before, meets, overlaps, during, starts, finishes, equal, plus six inverse relations), as well as point temporal algebra (<, =, >). The mapping between them has been solved by introducing aggregates that operate on sets such as *make-intervals*, *start_interval*, *end_interval*, as well as operations on the interval data type such as *duration*,



(a) (b) (c) (d)

Fig. 2. Feature extraction: (a) Original image; (b) Segmented player; (c) Pie features; (d) Skeleton features

intersect, and *union*. Additionally, to provide a precise timing, we added parameters to three of these temporal relations (*before (A, B, n)*, *overlaps (A, B, n)* and *during (A, B, n)*), and define a duration operator. Symbols A and B denote time intervals and n denotes the time difference between starting points of A and B in the case of *before* and *during* relations, but duration of overlapping in the case of the *overlaps* relation.

The set of feature operators φ is used for feature extraction and similarity matching. The set of real-world relations ω is domain dependent. For the tennis domain, it can comprise relations like 'is_right_handed', 'plays_closer_to_camera', etc. The real-world relations can have temporal dimension, so they can be related to a specific time interval. The set of video operators υ consists of operations on video frames and segments: *creation*, *concatenation*, *union*, *intersection*, *difference*, and *s_contains*. The last one checks if objects or events are present in a sequence, while others enable interpretation, composition, and manipulation of video segments.

5. Stochastic event formalization

Although the rule-based approach results in the automatic mapping from features to high-level concepts, we still have a problem of creating object and event rules manually, which might be difficult. As our model is independent of feature/semantic extractors, it provides users with flexibility of using different pattern recognition techniques for that purpose. Hence, instead of the grammars, different stochastic techniques, such as Bayesian belief networks, neural networks, or Hidden Markov Models (HMMs), can be used to map features to high-level concepts.

In our approach, we choose HMMs, because they are effective tools for modeling time-varying patterns with automatic learning capabilities. A major development in the theory of Hidden Markov Models was the maximization technique of Baum et al. [24], which has led to a wide range of theoretical outgrowths. Although continuous [25], and different extensions of discrete HMMs like Pseudo 3-D HMMs [26] are proposed, in most cases first order left-to-right discrete Hidden Markov Models are used. Consequently, the "Forward-backward" algorithm [27] is used for evaluation, and the Baum-Welch method for training. In the training process, a number of iterations of the Baum-Welch algorithm with different initial parameters is used aiming at finding the best model. The model with the highest probability is chosen as the result. In order to avoid computational problems, forward and backward variables are rescaled after each iteration. A flooring method is used to replace probabilities that become zero with small value ε. If the observation sequences are very short, modified re-estimation formula for the training with multiple observation sequences [27] is used.

6. Content-based retrieval of tennis videos

In this section, we describe the feature extraction process and show how each isolated approach (the rule-based spatio-temporal approach or, on the other hand, the stochastic approach) can be used to map features on high-level concepts in the tennis domain. In addition, we point out advantages of the integration of both approaches.

6.1. Feature extraction

We begin from the game scenes (camera observing the whole field as in the Fig. 2a). These scenes can be automatically extracted from the video using a number of global image features and some heuristics, but this is beyond the scope of this paper.

First step in our approach is to segment the player and the court from the background. This is done using the algorithm in [28]. Then, we extract features characterizing the shape of the segmented player's binary representation. Having the specific case of the human figure in this particular application, we extract special parameters trying to maximize their informativeness. Besides the player's position relative to the court ($f_{1\text{-}3}$), the dominant color (f_4),

and standard shape features such as the mass center ($f_{5\text{-}6}$), the area (f_7), the bounding box ($f_{8\text{-}11}$), the orientation (f_{12}), and the eccentricity (f_{13}), we extract the following features:

- The position of the upper half of the mask with respect to the mass center ($f_{14\text{-}15}$), its orientation (f_{16}), and the eccentricity (f_{17}). Those features describe the upper part of the body that contains most of the information.
- For each circle cut out that is centered at the mass center, we count the number of pixels in the mask ($f_{18\text{-}25}$) as shown in the Fig. 2c. This can be seen as a general approximate description.
- The sticking-out parts ($f_{26\text{-}27}$) are extracted similar to [29] by filtering and finding local maximums of the distance from a point on the contour to the mass center. Only certain angles are considered as indicated in Fig. 2d.

6.2. Rule-based approach

In order to be able to answer very detailed complex queries that comprise a combination of feature, spatial, and temporal relations, the Moa/Monet database managements system [30] is enriched with a video extension that includes the COBRA data model, as well as the object and event grammar. In the sequel, we first give an example of an object rule and then present some query examples that include in-line event descriptions.

Video objects can be extracted using the domain knowledge accumulated in the descriptions of the object grammar. For example, an object description that extracts the player closer to the camera from a segmented tennis shot using shape and color features is defined as:

ρ_{Player}: ({r_1: body, r_2: rect(0, 144, 384, 288)}, {700<area(r_1)<1200, dominant_color(r_1)=player.dress}, {contain(r_2, r_1)}, {}).

There are two regions involved: r_1 corresponds to the player and r_2 to the lower part of a frame. A few criteria concerning area (f_7) and dominant color (f_4) features have to be fulfilled. Furthermore, the region r_2 has to contain the region r_1.

Let us look at some query examples. The first query retrieves all video segments where Sampras is playing from close to the net for a given period of time. The new "player_near_the_net" event type is defined in terms of spatio-temporal object interactions assuming that objects like player and net have been already extracted using the object grammar, external extractors for specific objects, MPEG 4, or other techniques. Query 1 also shows how event descriptions can be parameterized (a user might be interested not only in Sampras, but in different players playing close to the net).

Query 1: SELECT vi.frame_seq
FROM video vi
WHERE s_contains (vi.frame_seq, event,
Player_near_the_net = ({o_1: player, o_2: net}, {},{},
{y_distance (o_1, o_2) < 50}, {duration (this) > 60}),
o_1.name = 'Sampras')

Let us consider the rule describing the event type 'Player_near_the_net' in query 1. There are two types of objects involved, player and net. There are no audio types and real-world relations among object types. But, there is one spatial relation (*distance*) defined on features $f_{1\text{-}3}$ and also one temporal relation (*duration*). The temporal relation says that this event type should last a specific period, as well as that the spatial relation should be valid for that period of time.

A user can also reuse already defined event types in order to define compound ones. For example in query 2, possible lobs are retrieved by describing a new event type using the already defined events ('Player_near_the_net' and 'Player_near_the_base-line') and some additional criteria (the interval relation *meets* means that event e_1 has to finish at the same moment when e_2 starts).

Query 2: SELECT vi.frame_seq
FROM video vi
WHERE s_contains (vi.frame_seq, event, Lob =
({e_1: player_near_the_net, e_2: player_near_the_base-line}, {}, {}, {meet (e_1, e_2)}), $e_1.o_1$=$e_2.o_1$)

However, this query will not retrieve all lobs (for example, the ones where the player stays at the net or smash the ball at the service line). To be able to retrieve all lobs, the position of the ball must be included.

In order to simplify the process of describing semantic events, a graphical user interface is developed. Fig. 3 shows how the 'Player_near_the_net' query can be expressed interactively on a screen. The user can also update the meta-data by adding a newly defined event, which will allow him to define more complex compound events that comprise this event. This will also speed up future retrieval of this event, because it is then resolved in the event layer without looking at other layers.

However, this approach is essentially restricted to the extent of recognizable events, since it might become difficult to formalize complex actions of non-rigid objects using rules. If we consider the tennis strokes for example, one can argue that they can be formalized solely using player and ball positions. However, that will not result in reasonable accuracy [8]. On the other hand, introducing other features in the event rules will increase accuracy, but unfortunately, will make these rules to complicated for ordinary user, as we experienced.

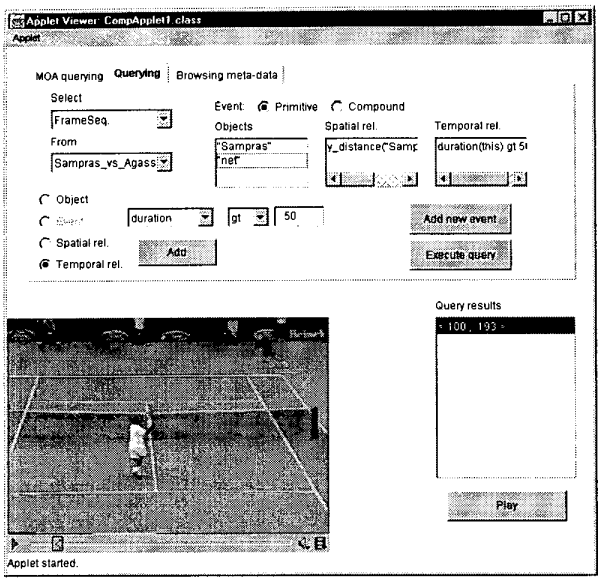


Fig. 3. The user interface for event descriptions

6.3. HMM approach

In order to solve the problem previously described, we have exploited automatic learning capability of HMMs, aiming at increasing recognition accuracy at same time. Consequently, we conducted a few experiments to demonstrate the validity of our approach.

In our experiments we used ordinary TV broadcast tennis videos with different players at different tournaments: Australian Open (Sampras-Agassi, Hingis-Molik, Kournikova-Davenport, etc.), Swisscom Challenge, Vienna Open, etc. Training sequences were manually selected, using a tool that was developed for video annotation and pre-processing. We used first order left-to-right discrete HMMs with 4 to 48 states.

Discrete HMMs require translation of continuous feature parameters (described in section 6.1) into symbols from a predefined set (codebook). In order to design a codebook we used the k-means algorithm [31]. Selection of the codebook size is a trade-off between a smaller quantization error (larger codebook size) and faster HMM operations (smaller codebook size). We tried various codebook sizes in the range of 8-80 symbols.

In the training process, we conducted a number of iterations of the Baum-Welch algorithm with modified re-estimation formula for the training with multiple observation sequences. Training a model with 20 different sequences (each approximately 45 frames) doing 50 iterations of the mentioned algorithm takes from 12 seconds for a model with 4 states and 8-symbol codebook size to approximately 20 minutes for a model with 48 states and 80 symbols on a PC with dual Pentium II.

6.3.1. Experiment 1. In this experiment we aimed at achieving of two goals: (1) determine the best feature set and (2) investigate person independence of different feature sets. Hence, we have performed a number of experiments with different feature combinations. In order to examine how invariant they are on different players including female ones, two series of experiments have been conducted: 1*a* and 1*b*. In the series 1*a*, we used the same player in the training and evaluation sets, while in 1*b* HMMs were trained with one group of players, but strokes performed by other players were evaluated. In both cases, the training set contained 120 different sequences, while the evaluation set contained 240 sequences.

To be able to compare our results with [12], we selected the same six events to be recognized: forehand, backhand, service, smash, forehand volley and backhand volley. In each experiment, six HMMs were constructed - one for each type of events we would like to recognize. Each stroke sequence was evaluated by all HMMs. The one with the highest probability was selected as the result (parallel evaluation). In order to find the best HMM parameters, a number of experiments with different number of states and codebook sizes were performed for each feature combination.

Table 1. Recognition results (%)

Feat.\Ex.	1*a*	1*b*	2
$f_{12\text{-}15}$	82.4	79.3	75.8
$f_{12\text{-}17}$	84.6	82.4	80.5
$f_{12\text{-}13,\ 16\text{-}17}$	81.5	78.6	76.1
$f_{12\text{-}13,\ 26\text{-}27}$	89.3	88.1	87.2
$f_{12\text{-}27}$	86.4	82.1	79.3
$f_{13\text{-}15,\ 26\text{-}27}$	91.2	88.7	88.3
$f_{18\text{-}25}$	85.4	77.8	78.1
$f_{18\text{-}27}$	93.1	87.0	86.4

The recognition accuracies in Table 1 (% of rightly classified strokes using parallel evaluation) show that the combination of pie and skeleton features ($f_{18\text{-}27}$) achieved the highest percentage in the experiment 1*a*. The recognition rates dropped in experiment 1*b* as expected, but the combination of eccentricity, the mass center of the upper part, and skeleton features ($f_{13\text{-}15,\ 26\text{-}27}$) popped up as the most person independent combination, which is nearly invariant on different player constitutions. The optimal result with this combination of features was achieved with the codebook size of 24 symbols and HMMs with 8 states.

Compared to [12], we achieved an improvement of 20% in the recognition accuracy (experiment 1*b*) due to a better training algorithm and mostly from improved, more informative, and invariant features (in first place novel skeleton features and then pie features). The improvement we achieved is certainly more significant taking into

account that we used TV video scenes with a very small player shape compared to the close-ups used in [12].

6.3.2. Experiment 2. In this experiment, we investigated recognition rates of different feature combinations using a regular classification of strokes from tennis literature [32]. There are 11 different strokes: service, backhand slice, backhand spin, backhand spin two-handed, forehand slice, forehand spin, smash, forehand volley, forehand half-volley, backhand volley, and backhand half-volley. The training and the evaluation set remained the same as in experiment 1*b*, only the new classification was applied.

Although some strokes in this new classification are very similar to each other (for example volley and half-volley or backhand slice and spin), the performance (Table 1, last column) dropped only slightly. The majority of false recognitions remained the same as in experiment 1. Nearly 65% comes from forehands recognized as backhands and vice versa, as well as from forehand-volleys recognized as forehands and vice versa. Having, for example, the ball position (an attempt is reported in [33]) would certainly make the distinction between the strokes more robust and significantly increase the recognition rate.

6.4. Integrated approach

The advantage of the COBRA video data model is that it provides a framework for the integrated use of these two approaches. Therefore, we can benefit of using the spatio-temporal event formalization together with the stochastic formalization and answer more detailed complex queries such as "Give me all video sequences where Sampras approaches the net with the backhand slice stroke".

Query 3: SELECT vi.frame_seq
FROM video vi
WHERE s_contains (vi.frame_seq, event,
Appr_net_BSL = ({e_1: Player_near_the_net,
e_2: Backhand_slice}, {},{},{},
{before (e_2, e_1, n), n < 75}),
$e_1.o_1$.name = $e_2.o_1$.name = 'Sampras')

This query comprises a new event type "Appr_net _BSL" that consists of two events. The first one is the "Plyer_near_the_net" event, which is defined in section 6.2 using the spatio-temporal rules, while the second one, the "Backhand_slice" event, is defined in section 6.3 using the HMM approach. For this new event type, we also have one temporal relation that requires e_1 to start at least 75 frames before event e_2.

We can also ask sequences, where Sampras is supposed to play the backhand stroke, but plays the forehand instead. This is another combined query where both approaches have to be used.

Query 4: SELECT vi.frame_seq
FROM video vi
WHERE s_contains (vi.frame_seq, event,
Forehand_instead_of_backhand = ({ e_1:Forehand }, {},{},
{$e_1.o_1$.direction_west> $e_1.o_1$.direction_east}, {}),
$e_1.o_1$.name='Sampras')

7. Conclusions

In this paper, we have introduced our COBRA video data model that provides a framework for using different knowledge-based methods for interpreting raw data into semantic content. The model is independent of feature/semantic extractors, providing flexibility in using different video processing and pattern recognition techniques for that purpose. We have extended the model with the object and event grammars, which aim at formalizing descriptions of high-level concepts, as well as facilitating their extraction based on features and spatio-temporal logic. Next to that, the model also supports stochastic modeling of events (in this case HMMs).

Each approach in isolation, as well as the integrated approach, has been validated for retrieval in the particular domain of tennis game videos. Consequently, a set of novel features and a robust feature extraction scheme have been introduced for this particular domain. A number of experiments with HMMs have been carried out and the results proved that previously described skeleton features are of the greatest importance. They increase the number and the percentage of accurately recognizable strokes in comparison to the methods mentioned in literature. Furthermore, experimental results with regular classification of tennis strokes demonstrated that our HMM approach is promising to realize statistics of tennis games automatically using normal TV broadcast videos. Eventually, we showed the advantage of the integration of the spatio-temporal and the stochastic approach.

We have already implemented the spatio-temporal approach within a prototype of our video database management system. Currently, we are investigating how a stochastic approach, particularly HMMs, can be integrated within the same VDBMS and what the possible advantages are, mainly with respect to feature selection and training tasks.

8. Acknowledgement

The authors would like to thank Zoran Zivkovic and Dejan Mitrovic for providing the tools for player segmentation and HMMs respectively. We also acknowledge Dragan Vuckovic for helpful discussions on

the stroke classification, as well as Roelof van Zwol and Henk Ernst Blok for useful comments.

9. References

[1] A. Yoshitaka, T. Ichikawa, "A Survey on Content-Based Retrieval for Multimedia Databases", *IEEE Transactions on Knowledge and Data Engineering*, 11(1), 1999, pp. 81-93.

[2] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, San Francisco, California, 1999.

[3] W. Al-Khatib, Y. Day, A. Ghafoor, P. Berra, "Semantic Modeling and Knowledge Representation in Multimedia Databases", *IEEE Transactions on Knowledge and Data Engineering*, 11(1), 1999, pp. 64-80.

[4] M.S. Kankanhalli, et al, "Video Modeling Using Strata-Based Annotation", *IEEE MMultimedia*, 7(1), 2000, pp. 68-73.

[5] S. Intille, A. Bobick, "Visual Tracking Using Closed-Worlds", *Tech. Report No. 294*, M.I.T. Media Laboratory, 1994.

[6] Y. Gong, L. T. Sin, C. H. Chuan, H-J. Zhang, M. Sakauchi, "Automatic Parsing of TV Soccer Programs", In *Proc. of IEEE International Conference on Multimedia Computing and Systems*, Washington D.C., 1995, pp. 167-174.

[7] H. Miyamori, S-I. Iisaku, "Video Annotation for Content-based Retrieval using Human Behavior Analysis and Domain Knowledge", In *Proc. of the IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 320-325.

[8] G. Sudhir, J. Lee, A. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval", *IEEE Workshop on Content-based Access and Image and Video Databases*, Bombay, India, 1998, pp. 81-90.

[9] N. Haering, R.J. Qian, M.I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video", *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(6), Sept. 2000, pp. 857-868.

[10] M. Shah, R. Jain (eds), *Motion-Based Recognition*, Kluwer Academic Publishers, 1997.

[11] Y. Rui, A. Gupta, A. Acero, "Automatically Extracting Highlights for TV Baseball Programs", In *Proc. of ACM Multimedia*, Los Angeles, CA, 2000, pp. 105-115.

[12] J. Yamato, J. Ohya, K. Ishii, "Recognizing Human Action in Time-Seuential Images using Hidden Markov Model", In *Proc. of IEEE Computer Vision and Pattern Recognition*, 1992, pp. 379-385.

[13] M. Naphade, T. Kristjansson, B. Frey, T.S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to indexing and retrieval in multimedia systems", In *Proc. of the IEEE ICIP*, Chicago, IL, 1998, vol. 3, pp. 536-540.

[14] D. Moore, I. Essa, M. Hayes, "Exploiting Human Actions and Object Context for Recognition Tasks", In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, Coufu, Greece, 1999, vol. 1, pp. 80-86.

[15] N. Vasconcelos, A. Lippman, "Bayesian Modeling of video editing and structure: Semantic features for video summarization and browsing", In *Proc. of the IEEE ICIP*, Chicago, IL, 1998, vol. 2, pp. 550-555.

[16] A.M. Ferman, A.M. Tekalp, "Probabilistic Analysis and Extraction of Video Content", In *Proc. of the IEEE ICIP*, Tokyo, Japan, 1999, vol. 2, pp. 91-95.

[17] T. Syeda-Mahmood, S. Srinivasan, "Detecting Topical Events in Digital Video", In *Proc. of ACM Multimedia*, Los Angeles, CA, 2000, pp. 85-94.

[18] M. Naphade, T. Huang, "A Probabilistic Framework for Semantic Indexing and Retrieval in Video", In *Proc. of the IEEE Intl. Conf. on Multimedia and Expo (ICME)*, New York, 2000, vol. 1, pp. 475-478.

[19] MPEG Requirements Group, *MPEG-7 visual part of eXperimentation Model 6.0*, ISO/IEC JTC1/SC29/WG11 MPEG2000/N3398, Geneva, CH, June 2000.

[20] MPEG Requirements Group, *Working Draft 2.0 of MPEG-7 Visual*, ISO/IEC JTC1/SC29/WG11 MPEG2000 N3322, Noordwijkerhout, NL, March 2000.

[21] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A System for Video Surveillance and Monitoring" *Tech. report CMU-RI-TR-00-12*, Carnegie Mellon University, 2000.

[22] D. Papadias, Y. Theodoridis, T. Sellis, and M. Egenhofer , "Topological Relations in the World of Minimum Bounding Rectangles: A Study with R-Trees", *SIGMOD RECORD* 24 (2), 1995, pp. 92-103.

[23] J.F. Allen, "Maintaining knowledge about temporal intervals", *Communications of ACM*, 26(11), 1983, pp. 832-843.

[24] L. Baum, T. Petrie, G. Soules, N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Annals of Mathematical Statistics*, 41(1), 1970, pp. 164-171.

[25] T. Starner, J. Weaver, A. Pentland, "Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video", *Pattern Recognition and Machine Intelligence IEEE Trans.on*, 20(12), 1998, pp. 1371-1375.

[26] S. Muller, S. Eickeler, G. Rigoll, "Pseudo 3-D HMMs for Image Sequence Recognition", In *Proc. of the IEEE Intl. Conf. on Image Processing (ICIP)*, Tokyo, Japan, 1999, pp. 237-241.

[27] S. Michaelson, M. Steedman, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[28] Z. Zivkovic, F. vd Heijden, M. Petkovic, W. Jonker, "Image Segmentation and Feature Extraction for Recognizing Strokes in Tennis Game Videos", In *Proc. of the ASCI*, Heijen, The Netherlands, 2001.

[29] H. Fujiyoshi and A. Lipton, "Real-time Human Motion Analysis by Image Skeletonization" In *Proc. of the IEEE WACV*, Princeton NJ, pp. 15-21, 1998.

[30] P. Bonz, A.N. Wilschut, M.L. Kersten, "Flattering an objects algebra to provide performance", In *Proc. of the IEEE Intl. Conf. on Data Engineering*, Orlando, pp. 568-577, 1998.

[31] R. Duda, R. Hart, *Pattern Classification and Scene Analysis*, John Wiley, 1973.

[32] J. Yandell, *Visual Tennis*, Human Kinetics, 1999.

[33] G. Pingali, Y. Jean, A. Opalach, I. Carlbom, "LucentVision: Converting Real World Events into Multimedia Experiences" *IEEE ICME*, New York City, vol.3, pp. 1433 - 1436, 2000