Tracking the Evolution of a Tennis Match Using Hidden Markov Models

Ilias Kolonias, William Christmas, and Josef Kittler

Center for Vision, Speech and Signal Processing,
University of Surrey,
Guildford GU2 7XH, UK
{i.kolonias,w.christmas,j.kittler}@surrey.ac.uk
http://www.ee.surrey.ac.uk/CVSSP/

Abstract. The creation of a cognitive perception systems capable of inferring higher-level semantic information from low-level feature and event information for a given type of multimedia content is a problem that has attracted many researchers' attention in recent years. In this work, we address the problem of automatic interpretation and evolution tracking of a tennis match using standard broadcast video sequences as input data. The use of a hierarchical structure consisting of Hidden Markov Models is proposed. This will take low-level events as its input, and will produce an output where the final state will indicate if the point is to be awarded to one player or another. Using hand-annotated data as input for the classifier described, we have witnessed 100% of the points correctly awarded to the players.

1 Introduction

The area of automatic data parsing from multimedia sequences has been one of great research interest for many years now. Such applications can be very useful for a wide range of purposes – most of them concerning automatic annotation and indexing of these sequences according to the criteria one might wish to specify. Especially for today's broadcasters, where the enormous amount of audiovisual information coming in to them can *only* mean more time spent on indexing what they store in their archives, such tools would greatly improve their ability to manage those archives in an efficient manner. Another important reason for this is the fact that automatic annotation of multimedia content will help us preserve it and re-use it through time.

More specifically though, sports events are very common in broadcast television, and also usually have a quite well-defined structure. Sports will either run against the clock and have some distinctive events as reference points (like football or basketball have scoring) or will be solely based on specific events which will define their evolution through time (like tennis or volleyball have scoring). Therefore, the underlying structure of each sport, which is largely dictated by its rules, is strong enough to allow for such approaches to be designed and more efficiently implemented.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1078–1086, 2004.

In this work, we focus on trying to extract high-level information from processing low-level input data for a specific kind of content – video sequences of tennis matches. The choice of a clearly event-driven sport will make things easier in this analysis by not explicitly involving time. Besides, time-constrained sports can be considered as event-driven ones if we consider a set of *'time'* events being triggered at regular intervals throughout the sequence. The layout of this paper is as follows: in the following section, a short summary of some relevant work on scene evolution tracking in video sequences and event recognition in sport videos is presented; Section 3 is a presentation of the proposed system; the results from the application of the proposed system to a hand-annotated data set are presented and discussed in Section 4; and finally, conclusions and ideas for future work are given in Section 5.

2 Related Work

As we can see in literature, HMMs and other forms of Dynamic Bayesian Networks (DBNs) in general are very powerful tools in the area of machine learning. Hence, a great number of cognitive visions applications have deployed such tools in order to perform information extraction from video sequences. Some of those attempts would include those of [1] for automatic annotation of Formula 1 race programs, [2] for recognising various types of strokes from tennis players during a tennis match, [5] for the recognition of general hand gestures, [3] for the recognition of American Sign Language and [4] for general human pose estimation. Apparently, such pieces of work may prove to be extremely useful in our case as well, and they can certainly serve as a guideline of what is feasible in cognitive vision, and what issues in this area are still open.

In the work by Ivanov and Bobick [5], we can see that the authors introduced a framework by which they analyse each complex event into its constituent elementary actions; in one of the examples the authors have used, a gesture is broken down into simple hand trajectories, which can be tracked more successfully via HMMs. Then, they apply *Stochastic Context-Free Grammars* to infer the full gesture. The results reported show that the proposed system performed quite well on real-world data. Such a paradigm can be considered as quite similar to that of a tennis match; if we consider all elementary events leading up to the award of a point in a tennis match to be the equivalent of the elementary gestures in this work, and the tennis rules related to score keeping as an equivalent of the grammar-based tracking of the full gesture the authors have implemented, we can easily see the underlying similarities between the authors' work and reasoning on tennis video sequences.

Another piece of work that deals with sets of body movements that are by definition constrained is that reported by Starner *et al.* in [3]. In this case, the objective was to develop a system that will be capable of recognising a gesture language that is the American Sign Language (ASL). The main idea behind this is the use of Hidden Markov Models to 'learn' a small lexicon of words in ASL, as they are expressed through gestures. The low-level visual information (in this case, the hands that has been segmented out of the input sequence) are separately analysed in terms of their shape, and this information is fed into the HMMs. The output is the word recognised by the system, and it can be seen that (at least for a small vocabulary of gesture-words) the system has performed very well. This piece of work, along with that discussed in the previous paragraph, show us that HMMs can be quite successfully applied in recognising shapes and tracking object trajectories – which are very important parts of analysing tennis video sequences at a lower level.

Moreover, in the work of Rosales and Sclaroff [4], a more general problem concerning human body pose estimation has been addressed. Through the use of an *artificial neural network* (ANN), the authors have initially attempted to classify the various possible configurations of the human body by initially capturing 3-D body positions, which they the project into 2-D (namely, a set of image planes). Given these 2-D projections, they have formulated exclusive subsets via unsupervised clustering, using the *Expectation-Maximisation* (EM) algorithm. The results of this approach have proven to be quite promising as well, showing us that we *can* use reasoning tools to distinguish between different body poses.

In [1], we can see how an overall cognitive vision system for sport videos has been developed and deployed. In this case, the authors have attempted to isolate semantic information from *both* the audio and the visual content of the sequence, and tried to annotate the video sequences processed by detecting events perceived as highly important; for example, and bearing in mind what events can occur in Formula 1 racing, they attempted to cover visual events such as overtaking, cars running out of the road etc. In addition, as the sequences used came from live broadcasts, they also included textual information about the race, like drivers' classification and times; that information was also extracted and used. The audio part of the sequences, since they came from normal broadcasts, was dominated by the race commentary; out of that, features like voice intensity and pause rates were also used. Having performed all these operations, the authors attempted to infer events of semantic importance through the use of Dynamic Bayesian Networks, attempting to infer content semantics by using audio and video information separately or combining this information in the temporal domain; both approaches yielded promising results when tested on simple queries (like finding shots in the Formula 1 race where a car runs out of the race track)

In another piece of work by Petkovic *et al.* [2], we can see a piece of work that is somewhat more constrained that that described in the previous paragraph, but which is still quite useful in terms of understanding the scene described, and which is obviously much more relevant to the present work. In this case, the authors have chosen to use HMMs in order to determine how the player hits the tennis ball; hit types would include forehands, backhands, serves, etc. To do this, the authors initially segmented the players out of the background (that is, the tennis court); then, they used Fourier Descriptors in order to describe the players' body positioning; finally, they trained a set of Hidden Markov Models to recognise each type of hit. The results of this work show that this method can be quite successful in performing the recognition task it was designed for.

Finally, another relevant piece of work is that of Kijak *et al.* [6], where the authors have attempted to analyse the structure of a tennis video through the use of Hidden Markov Models, as well as fuse audio and visual cue data in order to perform reasoning. Their objective was to separate a tennis video sequence into a set of scenes, each of which had to be classified under one of the following categories:

- First missed serve
- Rally
- Replay
- Break – that is, a *commercial* break

The visual cues include a vector of dominant colours and their respective spatial coherencies, and a measure of camera motion, whereas the audio cues form a binary feature vector where speech, applause, ball hits, noise and music are shown to be detected (or not). The results of the semantic segmentation system the authors proposed in this paper also seem to be quite promising.

Nonetheless, this is just a small subset of the applications these inference tools have been tested upon, and we can observe that all of these systems tend to perform the recognition and/or reasoning operations they were designed to do quite well. Therefore, it can safely be assumed that similar methods are capable of producing satisfactory results in other similar problems, like the one we are asked to address in this paper.

3 Proposed Scheme

In our context (the analysis of tennis video sequences), the rules of the game of tennis provide us with a very good guideline as to what events we will have to be capable of tracking efficiently, so as to follow the evolution of a tennis match properly. The full graphical model for the evolution and award of a point in a tennis match is given in the graph of Figure 1.

As we can readily see from this diagram, it is a graphical model where a number of loops exist; the state transitions drawn with bolder lines indicate where these loops close. Moreover, as it has already been mentioned, this graphical model *only* tackles the problem of awarding a single point in the match; there is some more detail to be added to it if we wish to include the awarding of games, sets or the full match. How these stages are going to be implemented will be discussed in more detail later in this section. Finally, this figure also shows us that, in order to address the problem of 'understanding' the game of tennis more effectively and robustly, we will have to convert this complex evolution graph into a set of simpler structures.

In the game of tennis, we have a case of a hierarchical evolution model. That is, we first need to identify and examine elementary events within the tennis sequence. Such events would include, among others:

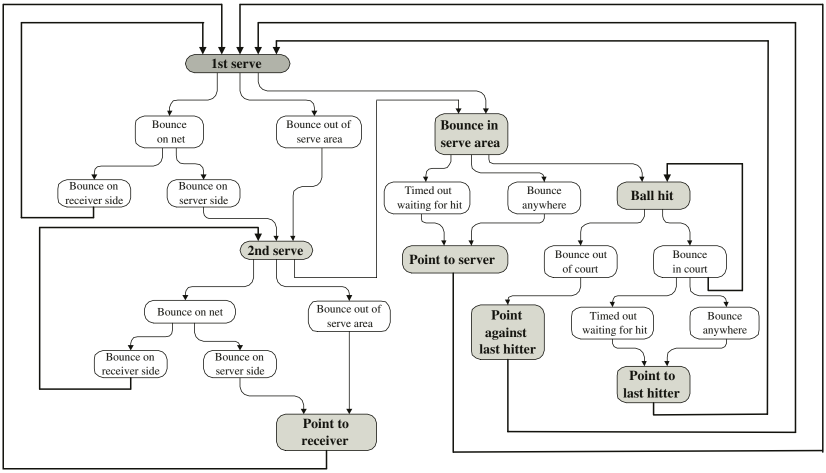

Fig. 1. Graphical model for awarding a point in a tennis match.

- The tennis ball being hit by the players
- The ball bouncing on the court
- The players' positions and shapes (that is, body poses)
- Sounds related to the tennis match (not implemented yet – however, preliminary experiments show that it *does* contain important semantic information as well)

These events can then be used as a basis on which to perform reasoning for events of higher importance, like awarding the current point from the events witnessed during play. Having successfully performed this step, we can then move on to the award of games, sets and the match to the players involved. Obviously, since the detection of such elementary events will be done using machine vision algorithms and techniques; however, we are bound to encounter event detection errors from this process – a fact which leads us to the conclusion that we will need some kind of correction at a higher level of reasoning to address such errors. Since probabilistic reasoning tools (such as HMMs) can address this problem effectively, they are a natural choice to perform reasoning processes where the input data comes *directly* from the low-level event detectors – which is the case in awarding points.

Therefore, the reasoning process described here is best represented by a hierarchical model. Using such a model will allow us to properly decompose the evolution of the tennis match into smaller events, which can be more concisely defined and tracked with greater accuracy. For example, it would be best if we modelled events of lower conceived importance through an HMM, which would then trigger another HMM to infer on more important events within the game; that would also help us prevent spurious data from low-level feature extraction modules from propagating to higher levels of the inference engine.

Moreover, we can also implement parts of this graph by using a *'switch'* variable to select which, among a number of acyclic sub-graphs, will be appropriately modelling the scene at that moment. Those sub-graphs will contain subsets of the initial, full-scale graph we have just seen. In more detail, we can consider the start of a tennis point (that is, the serve) as the starting state of a *set* of Hidden Markov Models, each of which will follow a particular scenario within the game; for example, the serve could either be successful (so we then 'switch' to the model that deals with a rally of balls) or unsuccessful (where we switch to the model dealing with second serves and double faults). Another useful application of a 'switching' model in this context is the fact that the initial state switches from one player serving to the other (or one player hitting the ball, then the other); therefore, a 'switch' variable between two models that are identical in structure but exactly opposite as to which player they address could help simplify the design and training of all models quite considerably.

Thus, we propose to replace the original scene evolution model with a set of smaller models, each one trying to properly illustrate a certain scenario of the match evolution. The most important thing we need to ensure during this procedure is that, when we combine all of the models in this set, we *must* have a model equivalent to the original one (so that it reflects the rules of tennis). The set of sub-graphs proposed to replace the original one is illustrated in Figure 2.

As we can see from the set of models above, we have opted for a more 'perceptual' way of selecting the set of event chains that will formulate the new

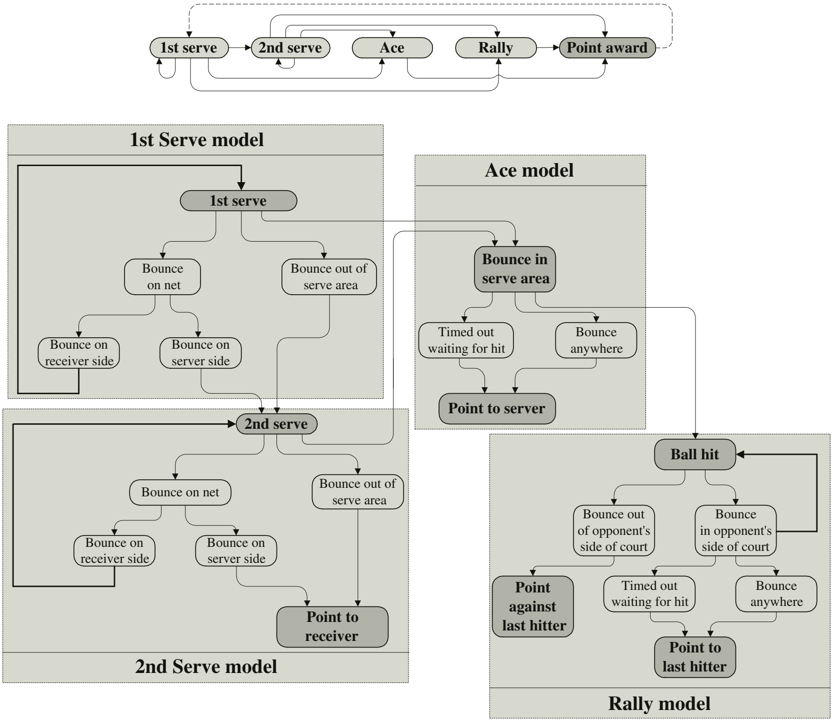

Fig. 2. Switching model and its respective set of sub-models for awarding a point in a tennis match, as separated out from the original graphical model.

model set for our purposes. This design strategy is going to be particularly helpful in the (quite frequent, as it is expected) case where the system receives a query to extract sequences which contain some specific event; since the scene description will consist of a series of events occurring in the match, such queries can be dealt with relatively easily. Moreover, choosing to break the initial graph down to a number of sub-graphs and train each one of them separately will be beneficial in many more different ways. Some of the resulting benefits are:

- Using probabilistic reasoning tools (like HMMs) will allow for correction of wrongly detected elementary events within the game – so we can have a point correctly awarded even if we haven't tracked *all* events accurately from the start.
- Since the models are simpler, we will *not* have to acquire a huge training data set; a relatively small amount of data per model will suffice for our analysis.
- Training will *certainly not* be as time-consuming if we break down the initial graph as it would be if we did not do so. This is because, apart from the training needing more input training data if it were to be trained as a whole, it would also need to calculate more event transition probabilities compared to what would be required if we broke the initial graph to smaller models.
- In some cases, we can *considerably speed up* the training process due to prior knowledge for this type of content – as the amount of statistics available for some events in a tennis match helps us get a very reasonable initial estimate without the need to go through a training procedure – we only need to pick up some existing measurements for this purpose.
- It will be *far easier* for us to determine which specific areas of reasoning need to be improved to boost the overall performance of the system, and which low-level feature extraction modules are more suspect to produce misleading results – so that we can either improve them or altogether discard them.

As soon as we have determined which way the points are going to be awarded, we can move on to the award of games and sets in the match. This can either be done through the use of probabilistic reasoning tools (such as HMMs), or with simpler, rule-based tools – such as grammars. The latter is possible due to the fact that at this level of abstraction in modelling tennis game sequences, it is the rules of the game of tennis that stipulate the evolution of the scene rather than low-level events in it. Anyway, the uncertainty stemming from the successful (or unsuccessful) detection of the elementary events mentioned above is considered to have been effectively addressed in the lower-level stages of the reasoning process – up to the level of point awarding. Therefore, and since it will be an easier and more natural approach to record the rules of tennis via a rule-based tool, we could just as easily opt for using grammars to perform higher levels of reasoning for these video sequences as we could use probabilistic reasoning approaches. An example of the point illustrated in this paragraph could be the way games are awarded in a tennis match out of points won by both sides (assuming that we record the score correctly at all times); if a player has scored *4 or more* points in the current game *and* his/her opponent has *at least 2 points less*, then this player has *won the game* – otherwise the game goes on.

4 Results and Discussion

The architecture described above has been tested on one hour's play from the men's Final of the 2003 Australian Tennis Open. The sequence contained a total of 100 points played – which was the equivalent of approximately one and a half sets of the match. Out of these 100 exchanges, a total of 36 were played on a second serve. The data that was used as input in this experiment were *ground-truth, hand-annotated event chains* from the broadcast match video. Therefore, we have *not* examined the robustness of the proposed method that in this work; we were more interested in its ability to actually model the evolution of the match accurately. To do that, we had to introduce four sets of models – one for every combination of which player serves and which side of the court he/she serves from (left or right).

In those 100 exchanges, we have intentionally left in a few *unfinished* points, so as to examine whether the selection of the hidden states for these models can lead to an accurate representation of the scene at *any* given time – *not only* at the end of the scene. They were 4 in total – 2 leading to a second serve and 2 were cut short while still on play. An overall view of the results is given in the table below.

Table 1. Total results.

	Ground Truth	Correctly Awarded	Wrongly Awarded	Not awarded (still on play)
Near Player Points	59	59	0	0
Far Player Points	37	37	0	0
Unfinished Points	4	4	0	n/a
TOTAL	100	100	0	0

As we can see in Table 1, *all* of the points were successfully tracked by the proposed system. Therefore, it can be easily seen that the performance of this method allows us to use any kind of decision-making scheme (either rule-based or probabilistic) to make decisions about events of higher semantic importance – since that is what is finally intended.

5 Conclusions

As we can readily see from the results shown above, the proposed system has tackled the problem of tracking the evolution of a tennis match very effectively. However, there are still some issues to be addressed in this area. First of all, is is obvious that the excellent performance of the proposed method can partly be attributed to the fact that the input events were came from manual annotation of a tennis video sequence. Thus, it has been verified that they were correct chains of events, either leading to points or not. However, the aim of developing such an automatic evolution tracking system is mainly to be used in conjunction with a set of fully automatic low-level feature extraction tools that will be able to detect the basic events required for input to the proposed system, so that the combined system can be effectively used as an automatic video annotation system. Obviously, as in any fully automatic computer vision system, the low-level feature extraction tools are bound to produce some recognition errors, which will propagate to the proposed decision-making scheme. Therefore, it is essential that the proposed system is tested to effectively address such situations and provide accurate information about the evolution of the game to higher levels of the inference engine.

Moreover, the system can only be considered as the first step in a hierarchical model that will fully describe the evolution of a tennis match – it will only cover the award of a single point in the match. The creation of a full system will include a system to award games, sets and finally the match to the players – which will all rely on the efficiency of this method – on top of it. Therefore, a full system will also have to include a system similar to the one proposed here for point award, which will address that problem effectively as well.

Acknowledgements

This work has been supported by the EU IST-2001-34401 Project VAMPIRE.

References

1. Petkovic, M., Mihajlovic, V., Jonker, W., Djordjevic-Kajan, S.: Multi-modal extraction of highlights from TV Formula 1 programs. In: Proceedings of the IEEE International Conference on Multimedia and Expo. Volume 1. (2002) 817–820
2. Petkovic, M., Jonker, W., Zivkovic, Z.: Recognizing strokes in tennis videos using Hidden Markov Models. In: Proceedings of Intl. Conf. on Visualization, Imaging and Image Processing, Marbella, Spain. (2001)
3. Starner, T., Weaver, J., Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1371–1375
4. Rosales, R., Sclaroff, S.: Inferring Body Pose without Tracking Body Parts. In: Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition (CVPR). Volume 2. (2000) 714–720
5. Ivanov, Y., Bobick, A.: Recognition of Visual Activities and Interactions by Stochastic Parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 852–872
6. Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F.: HMM based structuring of tennis videos using visual and audio cues. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03). (2003)