A Hierarchical Semantics-Matching Approach for Sports Video Annotation

Chao Liang, Yi Zhang, Changsheng Xu, Jinqiao Wang, and Hanqing Lu

National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China 100190
{cliang,yizhang,csxu,jqwang,luhq}@nlpr.ia.ac.cn

Abstract. Text facilitated sports video analysis has achieved extensive success in video indexing, retrieval and summarization. A commonly adopted basis in previous work is the separate alignment of timestamps between sports video and game text, which isn't a robust method for generic cross-media analysis. In this paper, we propose a hierarchical semantics-matching approach to annotate sports video. Our key idea is to link video and text with high-level semantics rather than low-level features and find the optimal video-text alignment based on the integral structure rather than individual conditions. For accurate event location, the whole algorithm is implemented in a hierarchical way to generate both refined and accurate video annotation result. Experiments conducted on both basketball and football matches demonstrate that our proposed approach is effective for text facilitated sports video annotation.

Keywords: Video annotation, sports video, semantics-matching.

1 Introduction

Rapid development in multimedia technology has contributed to an amazing growth of sports video content and its increasing popularity among the public. Meanwhile, various multimedia services provided via new media channels such as internet and mobile devices have greatly enriched audiences' watching experience. People are no longer content to passively watch sports programs edited by studio professionals, instead, they are longing for an active manner to enjoy sports matches according to their personalized preferences. For example, a Kobe Bryant's fan may prefer to watch video clips recording Kobe's dunk in the Lakers' match. In this situation, the ability to detect detailed high-level semantics from sports video and locate accurate video segments related to those events is of great value.

Traditional video annotation methods utilize heuristic rules [1] or statistical learning algorithms [2] to infer semantic events from various low-level [3] or mid-level [4] features. These methods can only detect few salient events and annotate them with simple concepts (e.g. goal or foul events) due to the existence of semantic gap. Obviously, such simple event annotation cannot meet audiences' personalized appetite like watching specific sportsman's specific action (e.g. Kobe's slam dunk or Yao's block). In order to obtain more abundant high-level semantics, external textual information is

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 684–696, 2009.

introduced to facilitate video annotation and achieved encouraging results. Babaguchi *et al.* [5] proposed a multimodal strategy using closed caption for event detection and video indexing. Xu *et al.* [6] raised an integrative approach to align text events with match phase information to detect multiple events in soccer video. Xu *et al.* [7] used web broadcasting text from sports websites to detect event semantics and achieved inspiring result. Although these methods utilized different text information, they all adopt 'timestamp' as a key link connecting video and text. Once two identical timestamps are detected in both video and text, their related video content and textual description are aligned as an annotation. Since timestamps are usually matched independently, these approaches cannot utilize the integral structure information to correct local errors. More important, 'timestamp' itself is not an intrinsic and always available connection between sports video and game text, which limits the application scope of such timestamp-based approaches.

In this paper, we propose a hierarchical semantics-matching approach for sports video annotation. The novelty of our approach is to connect video and text using their high-level semantics rather than low-level visual marks and search for the optimal video-text alignment based on the global structure rather than local conditions. We first encode video and text as tag sequences where each tag represents a combination of semantic events detected from video and text. Then we search for a global optimal sequence matching based on tags' semantic similarities. For accurate event segment location, the above matching algorithm is implemented in a hierarchical way, first on the attack-level (to obtain a coarse but correct initial matching) and then the shot-level (to generate a both correct and refined final matching). Finally, every aligned video-text pair represents a video annotation result.

The main contributions of our work can be summarized as follows:

- We link video content and its corresponding textual description through a high-level semantic association, which is an intrinsic and generic linkage across multimedia, and hence can be easily generalized.
- We apply a global optimal sequence matching algorithm to align sports video and game text, which can significantly reduce the final matching errors using integral structure information.
- We implement the matching algorithm in a hierarchical manner (on different granularities), which can accurately locate event segment on the proper scale and hence improve the event location precision in various sports matches.

The rest of the paper is organized as follows. First, a framework of the proposed method is presented in Section 2. Then, the technical details of coarse and refined semantics-matching method are described in the Section 3 and 4 respectively. Finally, Experimental results are reported in Section 5 and our conclusion and future work are given in Section 6.

2 The Proposed Framework

Fig. 1 illustrates the hierarchical framework of our proposed approach. The whole algorithm is implemented in two stages: the attack-based coarse alignment and shot-based refined alignment. In the first stage, we annotate video content on the level of

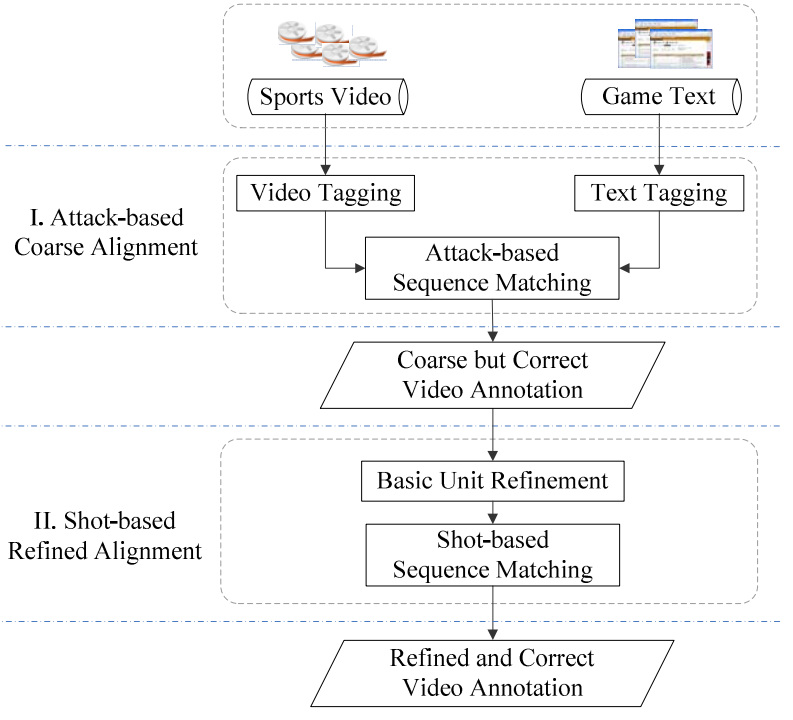


Fig. 1. Hierarchical semantics-matching framework

attack, which is defined as a complete attempt to score a point in sports matches [10]. We firstly encode the video and text as tagged sequences where each tag represents a combination of semantic events happened in an attack. Then we apply a sequence-matching algorithm to search for the global optimal alignment based on tags' semantic relationships. In the second stage, with the generated coarse but correct alignment result, we further divide each attack segment into shot-level clips and reapply the similar encoding and matching procedure to generate both refined and correct video-text alignment results.

3 Attack-Based Coarse Alignment

In the first stage, our approach try to find a proper granularity to annotate video content, so that we can obtain a coarse but correct initial result. In sports video scenario, we find the attack-based basic unit is quite suitable to the above task. On the one hand, attack represents a complete attempt to score goals or win points in sports match, hence it is semantic-related concept that can be accurately identified from both sports video and game text. On the other hand, attack is usually a longer temporal unit than shot, which makes attack-based event location results more robust to local errors than shot-based ones.

3.1 Video Tagging

The function of video tagging module is to generate a semantic tag sequence where each tag represents a combination of events contained in an attack. During this process, three main steps are included: attack-based video segmentation, content-based events detection and video sequence encoding.

Attack-based video segmentation is implemented by dividing adjacent video frames with inconsistent camera motion into different attack attempts. Considering the ubiquitous burring motion in sports video, we utilize the horizontal camera motion [9] and field zone information [6] to codetermine attack segments. Specifically, the initial horizontal motions are first smoothed by the field zone information so that burring motion clips without field zone change can be filtered out. Then, start point of each remaining motion segment is designated as a boundary point between two attack segments. Finally, the attack-based video segmentation is identified with a sequence of boundary points. A realistic example of the above process is illustrated in Fig. 2, in which the horizontal motions are classified into two classes: leftward and rightward motion and the whole field zone is partitioned into three parts: left, mid and right field.

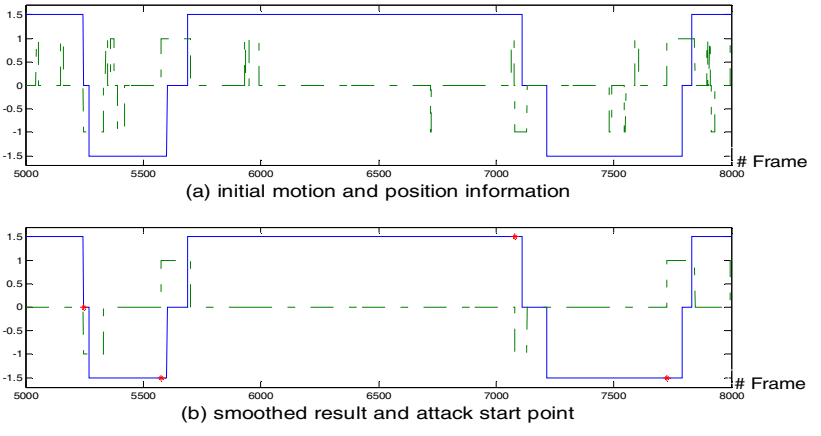


Fig. 2. Attack-based video temporal segmentation with blue solid line represent field zone information (1.5 denotes left field, -1.5 denotes right field and 0 denotes mid filed) and green dash-dot line represent camera motion information (1 denotes left moving, -1 denotes right moving and 0 denotes still). (a) Initial field zone information and burring camera horizontal motion information; (b) Filtered camera horizontal motion with four boundary points (denoted as red star) and three related attack segments (#1: 5210--5613; #2: 5613--7120; #3: 7120--7748.).

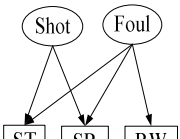
After dividing the video sequence into individual leftward or rightward attack segments, a group of mid-level audio-visual features including shot type transition (*ST*) [10], slow-motion replay (*SR*) [7] and referee whistling (*RW*) [11] are extracted from each attack segment and a simple Bayesian network (shown in Figure 3(a)) is designed to detect shot and foul events as follows:

$$\begin{cases} S^* = \arg\max_{G} \mathrm{P}(S \mid ST, SR, RW) \\ F^* = \arg\max_{F} \mathrm{P}(F \mid ST, SR, RW) \end{cases} \tag{1}$$

where S and F are binary variable representing the existing state of shot and foul events in an attack segment and S^* and F^* are most probable state according to the given observation, with *true* represents the event existent while *false* inexistent.

Finally, for each attack segment, fore detected attack direction and semantic events are further encoded by the combinations of their binary status (shown in Table 1) and the video sequence is converted into a semantically tagged sequence with each tag corresponding to an attack. An example of the tagging process is given in Fig. 3.

Example: Given an observation of shot transition ($ST = true$), slow-motion replay ($SR = true$), referee whistling ($RW = true$) in an rightward (D) attack, Calculate the posterior probability of the combination of Shot (S) and Foul (F) events and related video tag X.



(a) Naïve Bayesian network

[Shot Foul]	P(S, F | ST, SR, RW)
(no, no)	0.34*0.05 = 0.02
(yes, no)	0.66*0.05 = 0.03
(no, yes)	0.34*0.95 = 0.32
(yes, yes)	**0.66*0.95 = 0.63**

$X = \arg\max_{X} \mathrm{P}(X \mid ST, SR, RW)$
$= \arg\max_{X(D,S,F)} P(S, F \mid ST, SR, RW)$
$= X$ (*rightward* attack with *shot* and *foul* events)
$=$ *'d' (ref. Table 1)*

(b) Example of video tagging

Fig. 3. Example of video tagging with Naïve Bayesian network; Given an observation of shot type transition, slow-motion replay and referee whistling in an rightward attack segment, the Bayesian network infers that both shot and foul events are very likely to happen in that attack segment (with an inference probability 0.63). In this condition, we encode the attack segment with a semantic tag '*d*', which represents an rightward attack with both shot and foul events in our code book (*ref.* Table 1).

It is worth noting that the **video tag is intrinsically a discrete random variable with a probability distribution generated by the Bayesian network**. The variable value we used to tag an attack segment corresponds to the most likely events combination inferred from a group of mid-level features. Therefore the whole video tagging process can be express in a concise form as follows:

$$X^* = X(D, S^*, F^*) = \arg\max_{X(D,S,F)} \mathrm{P}(S, F \mid ST, SR, RW) \tag{2}$$

where X^* represents the final video tag encoded from the combination of attack direction (D) and the most probable semantic events (S^* and F^*), which is inferred from a group of mid-level audio/visual features including shot transition (ST), slow-motion replay (SR) and referee whistling (RW).

3.2 Text Tagging

For the ease of semantics extraction, we adopt the web broadcast text as the text source of our experiment. As shown in Fig. 4, web broadcast text usually contains

Table 1. Code book used for video tagging

Sematnic Tag	Attack Direction		Semantic Events	
	Rightward	Leftward	Shot Event	Foal Event
'a'	*true*	*false*	*false*	*false*
'b'	*true*	*false*	*true*	*false*
'c'	*true*	*false*	*false*	*true*
'd'	*true*	*false*	*true*	*true*
'A'	*false*	*true*	*false*	*false*
'B'	*false*	*true*	*true*	*false*
'C'	*false*	*true*	*false*	*true*
'D'	*false*	*true*	*true*	*true*

(a) web broadcast text (b) attack-based text segmentation

Fig. 4. (a) Web broadcast text (HTML); (b) Attack-based text segmentation with the help of team-player affiliation and event attack direction attributes

information such as the development of the match, players and event types descriptions, etc., which is very difficult to be obtained solely from content-based video analysis. Therefore it is an important supplement for the high-level semantics annotation tasks.

Similar to the video tagging, text tagging also includes attack-based segmentation, semantic events detection and sequence encoding. Firstly, domain knowledge is utilized to perform attack-based text segmentation (for algorithm details please reference Table 2). Secondly, semantic events are detected from each text attack segment. Finally, the same code book is used to encode each text attack segment, and the game text is finally converted into a tag sequence which is semantically identical to its video counterpart.

Table 2. Font sizes of headings. Algorithm of attack-based text segmentation.

Algorithm: Attack-based Text Segmentation		**Example:** Bucks vs. Suns (Bucks is the right side)
Input: text record; Output: Is new segment.		Text record: Michael Redd makes layup (Figure 4)
Build a background database to store *player-team affiliation* and *events attack attributes*.		*Player-team affiliation*: Redd belongs to Suns; *Event attack attribute*: layup is an offensive event.
1	Identify involved player and event type;	Player: Redd; Event: layup; (keywords searching)
2	Determine team's attack state based on the detected player name and event type;	Redd→Bucks; layup→offensive event. (database) Team's attack state: Bucks' attack.
3	Obtain current attack direction;	leftward attack (because Bucks is the right side)
4	If current direction is inconsistent with the last one, a new attack segment begins.	If last attack direction is rightward, then a new attack segment begins, or else the last continues.

Although video and text tagging share similar processing modules and output, we argue that these two sequences are intrinsically different. For the former, each tag is a random variable and the finally derived video sequence is composed of the most likely semantic tag given an observation of mid-level features of an attack segment. In contrast, the text sequence is a constant sequence with each tag is identified in a determinate way.

3.3 Attack-Based Sequence Matching

The output of video tagging is a tag sequence with accurate attack boundaries (in terms of video frames) but inaccurate semantic tags (due to the semantic gap), while the output of text tag is another tag sequence with accurate semantic tag (based on textual keywords searching) but no video boundaries information. Therefore, we hope utilize the accurate text tag sequence to label its corresponding video clip so that we can obtain a correct video annotation result on the level of attack. In this paper, we employ the Needleman-Wunsch alignment algorithm [12] to find the global optimal tag-correspondence between video and text sequences.

The Needleman-Wunsch algorithm is intrinsically a dynamic programming algorithm that searches the best matching mode through a multistage decision process. The main algorithm flow is illustrated in Fig. 5. A score matrix is first computed to

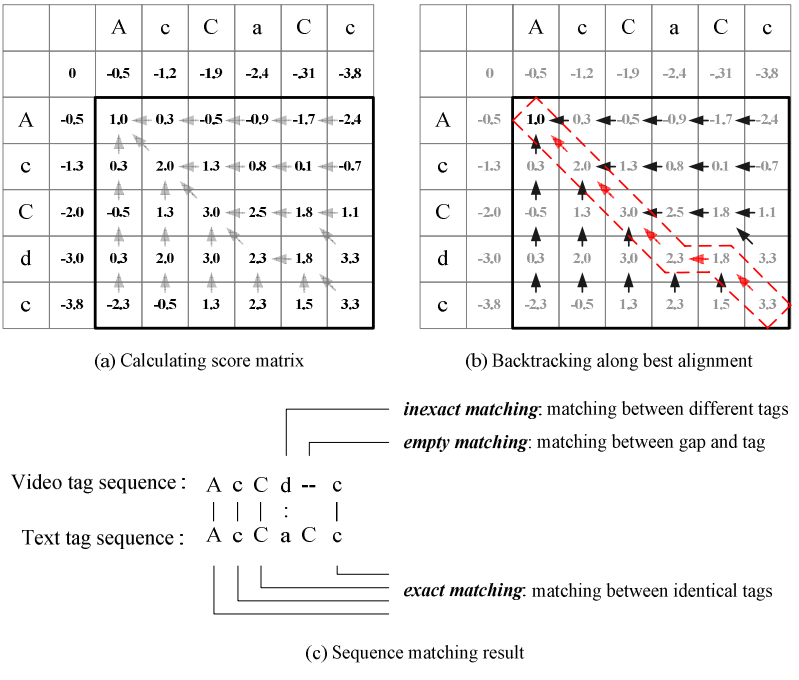


Fig. 5. (a) Calculating score matrix (illustrated in the dark black frame) and storing the optimal local matching directory (marked as arrows); (b) Backtracking along best alignment path (illustrated in the red dashed frame); (c) Sequence matching result with four exact-matching (denoted as 'l'), one inexact matching (denoted as ':') and one empty matching

find the local highest alignment score between two sub-sequences among three different matching directories, which are rightward (gap-tag matching), downward (tag-gap matching) and diagonal (tag-tag matching) respectively. For each obtained similarity score, its related local matching directory is also stored in the corresponding position of an equal-sized backtracking matrix. After the whole backtracking matrix is identified, the best alignment pathway is finally found through tracing back along the stored local matching directory.

In our approach, the forward computation of score matrix M is realized as follows:

$$\mathbf{M}_{i,j} = \max\{\mathbf{M}_{i,j-1} + Pg,\ \mathbf{M}_{i-1,j} + Pg,\ \mathrm{S}_{i,j} + \mathbf{M}_{i-1,j-1}\} \tag{3}$$

where $\mathrm{S}_{i,j}$ is the similarity score between the i^{th} video tag (V_i) and the j^{th} text tag (T_j), Pg denotes the gap penalty given to an empty matching and $\mathbf{M_{i,j}}$ is the local optimal alignment score between sub video sequences $\{V_1 \sim V_i\}$ and text sequence $\{T_1 \sim T_j\}$. Since the video tag is a discrete random variable value corresponding to the largest probability in the distribution generated by Bayesian network, $\mathrm{S}_{i,j}$ can be regarded as the probabilistic similarity between two semantic tags given a group of mid-level audio/visual features and calculated as follows:

$$\mathbf{S}_{i,j} = \frac{\mathrm{P}(X = T_j \mid ST, SR, RW)}{\mathrm{P}(X = V_i \mid ST, SR, RW)} = \frac{\mathrm{P}(X = T_j \mid ST, SR, RW)}{\max \mathrm{P}(X \mid ST, SR, RW)} \tag{4}$$

where X is the video tag variable and ST, SR, RW denote audio-visual features described in Section 3.1. Obviously, the more proximal the inference probabilities between tag V_i and T_j are, the more likely T_j can be used to replace V_i to annotate the related video segment, in other words, the more likely video segment tagged as V_i can be aligned with text group tagged as T_j. In the case of exact matching, the above similarity score is 1, corresponding to its maximum value. As for the gap penalty, it is defined as follows:

$$Pg = [-0.5 - 0.25 * (S^* + F^*)] * \alpha \tag{5}$$

where S^* and F^* represent the most probable existing state of shot and foul events and α is the affine gap cost defined as 1 for the first gap and 0.95 for others in our following experiments. As can be seen from the above equation, the more events contained in an attack, the less likely it cannot find a corresponding text record, hence the severer punishment will be given to its empty alignment, and vise versa.

With a timestamp as a tag, previous timestamp-based methods can be regarded as a special case of our approach. However, two important differences exist. Firstly, tags in our method represent high-level semantics rather than low-level visual features, which is an intrinsic and generic link across multimedia. Secondly, global structure rather than individual condition is utilized to align video and text sequences, which improve our approach's robustness to local errors. Therefore, the proposed semantics-matching approach is more effective for the generic cross-media analysis.

4 Shot-Based Refined Alignment

The output of the attack-based alignment is a coarse but correct annotation result. Each aligned tag pair corresponds to an attack segment related with a group of semantic

shots in the sports video and textual records in the game text. For sports like football where text records are usually scattered in different attack segments, the property relation between shots and text records in one attack is many-to-one. However, for sports like basketball where text records are usually clustered in one attack segment, the above ratio relationship is usually many-to-many. Therefore, a shot-based refinement process is needed to generate more elaborate annotation results (with shot-record ratio approximates to one-to-one).

4.1 Basic Unit Refinement

Based on mass observation, two general broadcasting rules are adopted to facilitate shot-based video and text segmentation respectively.

- Long shots are usually adopted to depict the global situation when the match is in play while short shots are always corresponding to the break state;
- Shot type transition is usually adopted in broadcast sports video when a shot or foul event happens or the ball is out of bound;

Although rule-based segmentation is not as accurate as attack-based result, it can generate more refined annotation on the scale of shot. What's more, our following experiments show that such correctness degeneracy during the refinement process is not significant hence is totally acceptable.

4.2 Shot-Based Sequence Matching

After performing the shot-level segmentation on both attack-based video segments and text groups, the same tagging procedure and sequence matching algorithm are applied to the generated shot-based semantic sequences and a refined alignment result is finally obtained on the shot scale.

5 Experiments

In order to verify the proposed method, we conduct our experiment on typical sports video including three NBA 2008 basketball matches and two Euro-Cup 2004 football matches. The corresponding text records are from ESPN [13] for basketball matches and BBC [14] for football. In average, there are about 400 text events happened in one 100-minute basketball match and 50 text records in one 90-minute football match.

5.1 Attack-Based Coarse Alignment

The coarse alignment result is listed in Table 3 where the first three are basketball matches and the last two are football matches. The inference accuracy denotes the occupation of correctly detected tags in total video tags and alignment accuracy represents the occupation of matched tags in total video tags. The difference between these two indexes lies in the inexact matching.

Table 3. Attack-based Coarse alignment Results

No.	Total Video Tags	Exact Matching	Inexact Matching	Inference Accuracy	Alignment Accuracy
1	192	143	44	74% (143/192)	97% (187/192)
2	196	152	42	78% (152/196)	99% (194/196)
3	188	133	49	71% (133/188)	97% (182/188)
4	55	45	8	82% (45/55)	96% (53/55)
5	50	43	6	86% (43/50)	98% (49/50)

As shown in Table 3, the average inference accuracy is only about 74% for basketball matches and 84% for football matches, which reflect the strong negative effects of semantic gap in Bayesian reasoning. However, in contrast to the limited inference accuracy, the coarse alignment accuracy is still satisfactory (around 98% and 97% respectively). This result demonstrates the strong fault-tolerant ability of the proposed semantics-matching algorithm and can be attributed to the allowance of inexact tag matching and the adoption of global sequence matching.

To further analyze above results and reasons, we take a realistic sequence alignment instance derived from our experiments as an example. In Fig. 6, the video and text sequences are generated from the automatic tagging process. Each alphabetic tag represents an attack with specific semantics (for details please *ref*. Table 1) and the tag '--' denotes the case of empty alignment. Two different matching marks, 'l' and ':', are used to denote exact and inexact matching respectively.

```
Video sequence :  B a C c D -- C d C c D d A d A a A d C d C a A a A a D d A d A c C a C c C d A a D
                  :   |   |   :    |  :   |   |  :  :   |   |   |   |   :   |  :   |   |   :   |   |   :   |   |   |   |   |   |   |   |   :   |  :
Text sequence :   A a C c B a C b C c B b A d A a A c C b C a A c A a D b A d A c C a C c C d C a A
```

Fig. 6. Coarse alignment results of the third quarter match between Suns and Bucks with a detection accuracy 70% and alignment accuracy 100%. The empty alignment is caused by video clip missing during the program making process hence is not the fault of video tagging.

Without loss of generality, let's take the first error video tag '*B*' as an example. In the video tagging process, the Bayesian network judges a shot event happens in the first attack segment based on the observation of shot type transition, but in fact there is nothing happens in that attack. In this condition, our proposed tags' similarity measurement can reasonably depict the replace-ability between tags '*B*' and '*A*' hence retains the possibility of their alignment. In addition, since the sequence matching algorithm searches the optimal alignment based on the global structure, sequence matching with an alignment between '*B*' and '*A*' may achieve higher total score than the one without such inexact alignment. Therefore, the proposed semantics-matching algorithm can effectively utilize the semantic similarity and global structural information to accurately align the related video and text tags even when they have different appearance.

5.2 Shot-Level Refined Annotation

Since events in sports are always overlapping, shot-level annotation is indispensable. In Table 4, we present the refined annotation results on three basketball matches and

Table 4. Refined annotation result (R: recall; P: precision)

No.		Shot	Miss	Block	Stolen	Foul	Rebound
1	R	107/144	73/78	11/11	20/20	44/48	87/89
	P	107/122	73/75	11/11	20/20	44/48	87/90
2	R	112/136	82/91	10/12	16/18	40/45	78/85
	P	112/132	82/84	10/11	16/18	40/43	78/82
3	R	108/134	84/92	13/14	19/22	41/43	92/100
	P	108/126	84/90	13/13	19/21	41/42	92/96

No.		Shot (goal)	Free Kick	Corner	Foul
4	R	9/11	13/15	6/7	16/17
	P	9/10	13/15	6/6	16/16
5	R	12/12	10/12	8/9	15/17
	P	12/13	10/10	8/9	15/16

restrict event types into six main categories: shot, miss, block, stolen, foul and rebound. Similarly, we give the shot-level annotation result on two football matches and restrict event types into four categories: shot (goal), free kick, corner and foul.

As can be seen from table 4, although the refined annotation is not as accurate as the coarse one, it can still locate various events in an acceptable precision. As for the lower accuracy of the shot event in basketball matches, it is mainly due to the occasional irregular photography in free throw events where short shots rather than long shots were adopted by cameramen.

5.3 Comparison

To demonstrate the robustness of our proposed approach, a comparative experiment of our approach with timestamp-based method [4] is conducted on the evaluation data and their F-measure results are shown in Fig. 7. In the ideal condition, timestamp-based approach can achieve very high event detection accuracy if the timestamp can be correctly recognized. However, according to our experiments on both basketball and football matches, the above advantage is either not obvious (Fig. 7 (b)) or even not exists (Fig. 7 (a)). This result can be explained from two aspects: Firstly, timestamp can not be always correctly located and identified in the practical noisy

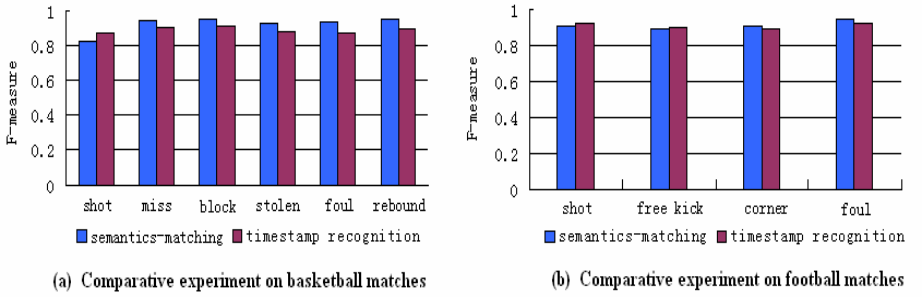


Fig. 7. Comparative results of event detection. The blue bar represents our semantics-matching method and the red timestamp-based method.

broadcast video, which affects the event location precision of timestamp-based methods. Secondly, because of the lack of structural information, local errors in timestamp-based method cannot be corrected by their context environment, which further degrades its comparative performance.

As shown in Fig. 7, the event detection precision on football matches is slightly lower than that of basketball matches, which is due to the performance degradation when the sequence matching algorithm is applied to two sequences with obvious length difference in the shot-level refinement of football matches.

6 Conclusion and Future Work

We have presented a hierarchical semantics-matching approach for sports videos annotation. The key idea of our method is to link sports video and game text in high-level semantics and utilize the structure information to search for a global optimal alignment between semantically tagged video and text sequences. Experiments conducted on three basketball and two football matches validate the robustness and effectiveness of our proposed approach.

In the future, we will extend our current work in two aspects: 1) we will explore the applicability of this semantics-matching approach in other video domains; 2) we will use the rich annotation result to provide personalized video customizaiton.

Acknowledgement

This work is supported by National Natural Science Foundation of China No. 60833006, Natural Science Foundation of Beijing No. 4072025, and 973 Program Project No. 2010CB327900.

References

1. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Trans. on Image Processing 12:7(5), 796–807 (2003)
2. Huang, C.L., Shih, H.C., Chao, C.Y.: Semantic analysis of soccer video using dynamic Bayesian network. IEEE Trans. on Multmedia 8(4), 749–760 (2006)
3. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: Proc. of ACM Multimedia, Los Angeles, CA, pp. 105–115 (2000)
4. Duan, L.Y., Xu, M., Chua, T.S., Tian, Q., Xu, C.S.: A mid-level representation framework for semantic sports video analysis. In: Proc. of ACM Multimedia, Berkeley, USA, pp. 33–44 (2003)
5. Babaguchi, N., Kawai, Y., Kitahashi, T.: Event based indexing of broadcasted sports video by intermodal collaboration. IEEE Trans. on Multimedia 4, 68–75 (2002)
6. Xu, H.X., Chua, T.S.: The Fusion of Audio-Visual Features and External Knowledge for Event Detection in Team Sports Video. In: Proc. of Workshop on Multimedia Information Retrieval, New York, USA, pp. 127–134 (2004)
7. Xu, C.S., Wang, J.J., Lu, H.Q., Zhang, Y.F.: A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans. on Multimedia 10(3) (2008)

8. Wang, L., Lew, M., Xu, G.Y.: Offense Based Temporal Segmentation for Event Detection in Soccer Video. In: Proc. of Workshop on Multimedia Information Retrieval, New York, USA, pp. 259–266 (2004)
9. Dufaux, F., Konrad, J.: Efficient, Robust and Fast Global Motion Estimation for Video Coding. IEEE Trans. on Image Processing 9(3) (2000)
10. Zhang, Y.F., Xu, C.S.: Rui. Y., Wang, J.Q., Lu, H.Q.: Semantic Extraction From Basketball Games Using Multi-modal Analysis. In: Proc. of IEEE International Conference on Multimedia and Expo., Beijing, China, pp. 2190–2193 (2007)
11. Xu, M., Duan, L.Y., Xu, C.S., Kankanhalli, M., Tian, Q.: Event Detection in Basketball Video Using Multiple Modalities. In: Proc. of IEEE Pacific Rim Conference on Multimedia, Singapore, vol. 3, pp. 1526–1530 (2003)
12. Needleman, S.B., Wunsch, C.D.: A General Method Applicable to The Search for Similarities in The Amino Acid Sequence of Two Proteins. J. Mol. Biol. 48(3), 443–453 (1970)
13. http://sports.espn.go.com/nba/
14. http://news.bbc.co.uk/sport