SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2016, Vol. 59 072103:1–072103:15
doi: 10.1007/s11432-015-5494-4

High-level representation sketch for video event retrieval

Yu ZHANG[1], Xiaowu CHEN[1]*, Liang LIN[2], Changqun XIA[1] & Dongqing ZOU[1]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, School of Compute Science and Engineering, Beihang University, Beijing* 100191*, China;*
[2]*School of Advanced Computing, Sun Yat-Sen University, Guangzhou* 510006*, China*

Received October 22, 2015; accepted December 29, 2015; published online June 15, 2016

Abstract Representing video events is an essential step for a wide range of visual applications. In this paper, we propose the event sketch, a high-level event representation, to depict the dynamic properties of video events composed of actions of semantic objects. We show that this representation can facilitate a novel sketch based video retrieval (SBVR) system, which has not been considered before to the best of our knowledge. In this system, users are allowed to draw the evolutions (e.g. spatiotemporal layouts and behaviors of semantic objects) on a board, and retrieve the events whose semantic objects have the similar evolutions from a database. To do this, event sketches are constructed on both the user queries and database videos, and compared under a novel graph-matching scheme based on data-driven Monta Carlo Markov chain (DDMCMC). To test our approach, we collect a novel dataset of goal events in real soccer videos, which consists actions of multiple players and shows large variability in the evolution process of the events. Experiments on this dataset and the publicly available dataset CAVIAR demonstrated the effectiveness of the proposed approach.

Keywords video retrieval, event representation, event sketch, drawing board, distance function, relevance feedback, DDMCMC

Citation Zhang Y, Chen X W, Lin L, et al. High-level representation sketch for video event retrieval. Sci China Inf Sci, 2016, 59(7): 072103, doi: 10.1007/s11432-015-5494-4

1 Introduction

Representing video events in terms of their evolutions (e.g. the spatiotemporal layouts and behaviours of the semantic objects engaged in the events) is a novel and important topic. For various domains such as video surveillance, film production and digital entertainment, such a representation is essential for effective retrieval of video clips containing plots of users' interest. For example, the inspectors may need to pull out the surveillance videos with events like two people walk together and then split and, also, the coach may want to collect real goal shots of a certain type in soccer events, like a player stands for a while, then shoots the ball to the keeper, with people watching. Illustrations of these two examples are shown in Figure 1. Note that it is much different with the conventional video retrieval problems [1–3], which focused on mapping video content to specific concepts or tags. In contrast, the problem to study in this paper is more complex in the form of its query: it aims to depict the high-level spatiotemporal structures of a specific event, and retrieve video events with the similar evolution patterns.

---

* Corresponding author (email: chen@buaa.edu.cn)

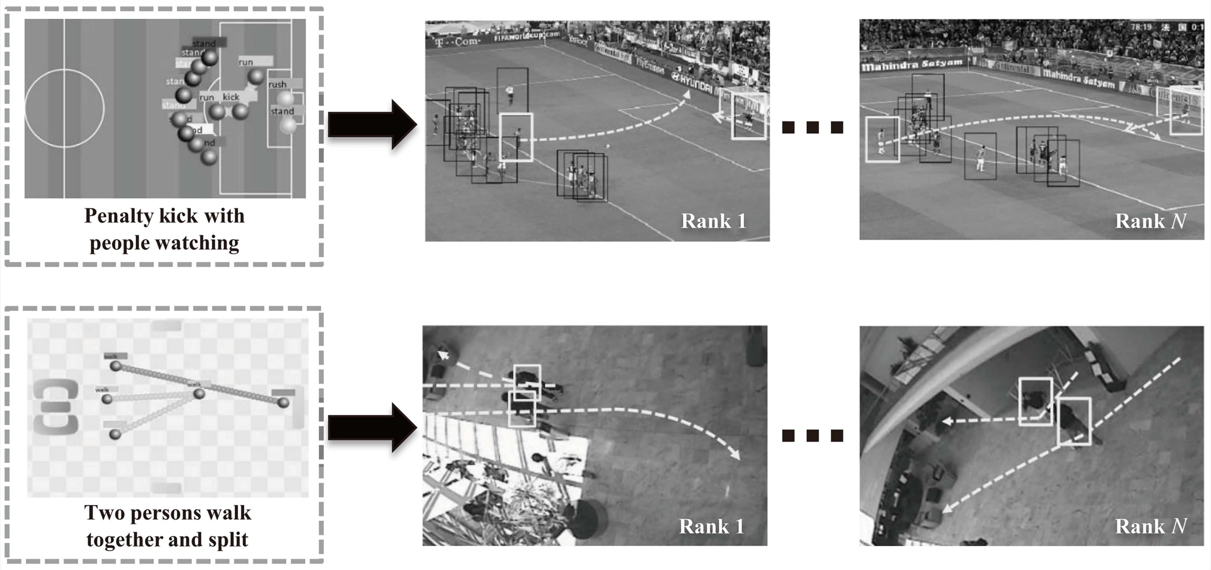

stand
run
rush
kick
stand
Penalty kick with people watching
Rank 1
Rank N
walk
walk
Two persons walk together and split
Rank 1
Rank N

Figure 1 The motivation of this paper. Left: the queries for retrieving detailed event evolutions, formed as dots and lines which represent the agents and their movements, respectively. Right: the retrieved videos containing the desired events. The agents with similar movements with those defined by the queries are marked by boxes and arrows.

To approach this problem, this paper proposes a novel event representation, named event sketch, to depict the high-level properties of events. The concept of event sketch derives from the sketches in conventional sketch based retrieval (SBR) systems for image, video or 3D object retrieval [4–12]. In these approaches, sketches are some simple strokes drawn by users to represent the shapes, colors or motions of the queried objects. Likewise, the proposed event sketch expresses the rough spatiotemporal layouts and the behaviours of objects in a specific video event. It is formulated as an acyclic graph, whose nodes denote the actions of the semantic objects in the video, and the edges represent the role and contextual relationships among actions, respectively. Each node is associated with some attributes, indicating the semantic label and space-time location of the action.

Furthermore, we present a novel SBR framework to address the video retrieval task shown in Figure 1 based on the proposed event sketch. In our SBR system, users are allowed to query events by specifying the evolutions of video events, which can be achieved through the following two ways. Firstly, the event could be specified using a drawing board (e.g. the left-most column in Figure 1), by drawing the actions of objects and their coarse trajectories. Also, the query can be made by tracking and labeling the objects of interest in an example video. To match the query and the database video, we construct event sketches for them and measure their consensus in event properties of different aspects. This is implemented by a weighted combination of distance measures, where the optimal weights are determined by relevance feedback technique. Given the graph distance, we then find the best match between two event sketches through a DDMCMC based graph matching algorithm. To test our approach, we collect a novel dataset of goal events in real soccer videos, which consists actions of multiple players and shows large variability in the evolution process of the events. Experiments are conducted on this dataset and the publicly available dataset CAVIAR to evaluate the proposed retrieval system and analyse the convergence of the graph matching algorithm. Finally, a user study is performed, which demonstrates the robustness of our retrieval system to handle the variations of the same query given by different users.

The main contributions of this paper are summarized as follows: (1) we propose a novel event representation named event sketch, which captures different high-level event properties. Event sketch is easy to be implemented, and effective for representing evolutions of events; (2) Based on the event sketch, a novel SBR system that can retrieve video events in terms of their evolutions is presented. To the best of our knowledge, this is the first attempt to address video retrieval in such a complex scenario; (3) A new dataset consisting of soccer goal events is compiled. Videos are classified into three categories according to the different evolution processes. However, in the same category, the video events still show large variabilities, making the dataset very challenging for event retrieval tasks.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on video event retrieval. Section 3 gives a brief overview, and explains the novel form of query of the proposed SBR system. In Section 4, we give the formulation of event sketch and the definition of graph distance function, as well as the strategy to learn optimal function parameters. Section 5 explains the system detailedly, including creation of user queries, construction of video database and graph matching algorithm under a DDMCMC framework. Experimental protocols and results are shown in Section 6 and conclusion is drawn in Section 7, respectively.

2 Related work

Video event retrieval has drawn extensive attention in a wide range of research areas, including information retrieval, computer vision and multimedia. According to the forms of user queries, existing approaches can be grouped into three main categories: concept-based, example-based and sketch-based approaches.

Concept-based video retrieval aims to map visual content in videos to semantic concepts. They are widely used when the user queries are given by simple text, like tags and words. For example, Ulges et al. [3] utilized user tags to automatically train concept detectors on YouTube videos, and applied them to video retrieval tasks. Bao et al. [2] combined explicit concepts defined by human experts and implicit concepts automatically learned by machines to improve the retrieval performance. Further, Yuan et al. [1] proposed to explore the relationships between primitive visual concepts by utilizing high-level semantic descriptors.

Example-based video retrieval systems receive example videos given by users as query. The function of these systems is thus to obtain videos with similar content with that of the query. For example, Yu et al. [13] characterized the query video events and video datasets as spatio-temporal interest points, and proposed randomized visual vocabularies to enable fast and robust point matching for retrieval. Lan el al. [14] proposed to match the action context descriptors to feature action retrieval from surveillance videos. In addition to matching the visual descriptors, the approach [15] proposed to match the motion trajectories of the objects in the query and database video. Moreover, Ref. [16] combined the motion models and the appearance models to further enhance the retrieval performance.

Sketch-based video retrieval is derived from sketch-based image retrieval [8–10], which is more powerful to express users' desires than concepts, and more convenient to use than collecting video examples. Although having been extensively studied for image retrieval, SBR systems receive only a few efforts [4–7] for videos. In these approaches, Collomosse et al. [5] developed a system that allows users to draw on a free-hand storyboards to retrieve video objects of the specified shape, color and motion. After that, Hu et al. [4] proposed to improve this work by matching storyboard sketches using trellis-based distance. They continued their work in [6, 7] by introducing semantic labels of video content to develop systems capable of searching simple scenes based on desired shapes, motion paths, and semantic labels of objects presented.

The existing sketch-based video retrieval systems, however, can only address retrieval problems when the events in videos involve a small number of objects and simple motions. In contrast, the SBR system developed in this paper attempts to extend the idea to handle more complex video events, which are typically performed by a large number of objects with more complicated spatiotemporal relationships. The most similar work to this is by Xu et al. [12], which allows users to query web images through drawing semantic concepts and their spatial layouts on a board.

3 A high-level view of the proposed SBR system

The proposed SBR system for video event retrieval is illustrated in Figure 2. Users are allowed to query events by drawing the evolutions on a aboard, or labeling them in the example videos. Event sketches are

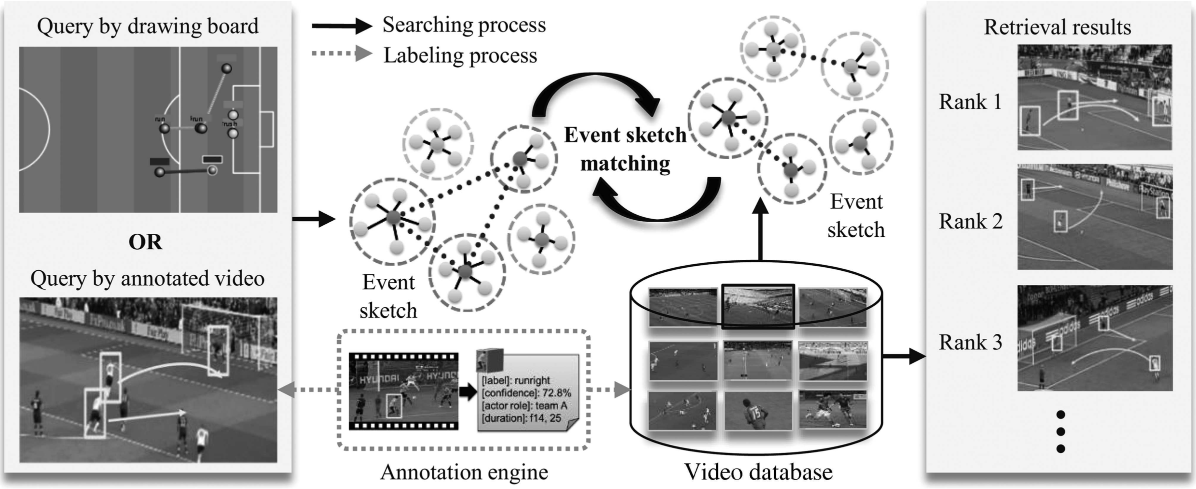


Figure 2 The proposed SBR system for video event retrieval. Two kinds of queries are supported by our system: the events drawn using the provided drawing board, or a realistic video with the annotated agents as well as their movements. An abstract high-level representation called event sketch is constructed for both the query events and database videos, and matched to measure their relevance.

constructed on query event and the events in database videos, and compared to measure their relevance.

In our system, drawing board is a visualized query interface allowing users to customize event processes on space and time axes, which offers necessities of three aspects for video event retrieval. Firstly, it enables to precisely describe video event structures which are hard to be accessed to for traditional text or visual based query interfaces. A query in our system can be viewed as a simulated event and thus can depict event properties comprehensively. Secondly, the events on a board contain rich high-level semantics in space and time, which can be leveraged to model events of users' interest and retrieve results with high relevance without requiring explicit domain knowledge of event structures. Finally, it is easy to be understood and mastered by ordinary users without the need for expertise on domain event description. Detailed usage of drawing board is referred to Subsection 5.2.

Served as alternative query heuristic, annotation engine is also introduced in our system, through which users are able to upload real videos with labeled information, e.g. semantics and tracked motions of objects. Another function of annotation engine lies in the off-line database preprocessing step and will be elaborated in Subsection 5.1.

4 Event sketch

In this section, we will formulate event sketch and define the weighted distance function between two event sketches. To determine the optimal function parameters, we then present how to leverage relevance feedback technique for parameter learning.

4.1 Formulation

An event sketch $G = \langle V, E \rangle$ is an attributed graph shown in Figure 3, where the set of nodes and edges are denoted by V and E, respectively. Each node v in V represents an individual action with label vector $\xi(v) = (i_v, l_v, \tau_v, s_v)$. The first two components, i_v and l_v, denote the identifier of action executor and category, respectively. Each action is associated with the temporal duration $\tau_v = (t_v, I_v)$, encoded by a start anchor time point t_v and the temporal interval I_v. We also record the spatial locations for each action on each frame in its living time, denoted by s_v.

In the event sketch, the edges represent two kinds of different relationships, denoted by dashed edges and solid edges in Figure 3, respectively. The dashed edge constrains that the connected nodes should have the same action executors (have the same identifier of action executors). The solid edge specifies

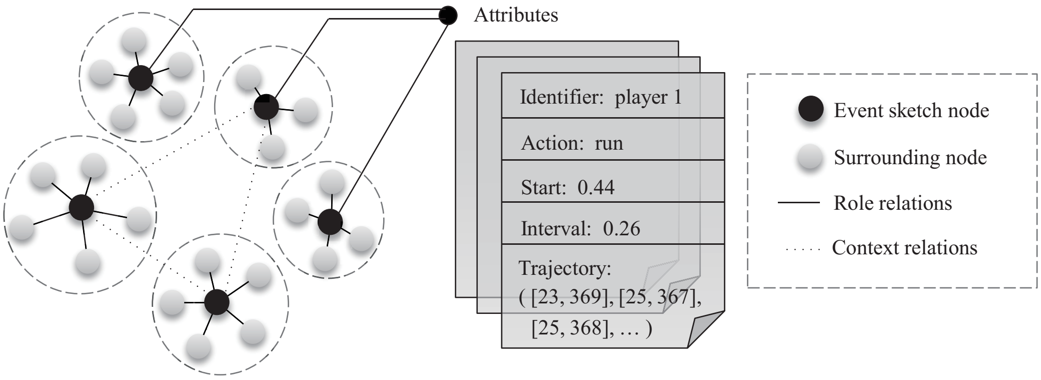


Figure 3 The graph representation of event sketch. The nodes represent the individual actions of the agents, while we use black and grey color to distinguish between the actions that are considered as candidate matches to the query, and the other actions that are spatiotemporally close to the candidate actions. Two kinds of the relationships among the actions, denoted by the dashed and solid lines respectively, are considered in the model. See text for more details.

contextual relationships among actions. That is, if two actions have a spatial distance below than a threshold, then there exists an edge between them. The spatial distance is computed as the average Euclidean distance of the spatial locations in the co-occurrence time period of the given two actions.

4.2 Distance measure

Denote G and G' as the event sketches of the query and the database video, and V and V' as the node set of G and G', respectively. Assume that there exists an injective mapping $\Psi : V \to V'$ between the node set V and V'. The graph distance M_Ψ between G and G' under this mapping is then defined as a weighted combination of five components:

$$M_\Psi = C \cdot M^x_\Psi + (M^s_\Psi, M^t_\Psi, M^l_\Psi, M^c_\Psi) \cdot \omega^{\mathrm{T}}, \tag{1}$$

where C and ω are controlling parameters, ω^{T} is the transpose of ω. M^x_Ψ, M^s_Ψ, M^t_Ψ, M^l_Ψ, and M^c_Ψ stand for similarity measures of event properties in aspects of complexity, semantics, time, layout and context, respectively. We will discuss them detailedly in the rest of this subsection.

Complexity distance M^x_Ψ measures the difference of graph complexity of both event sketches. It is defined by the difference of number of actions if $|V| > |V'|$, or zero otherwise. Here $|\cdot|$ is the cardinality of the set.

Semantic distance M^s_Ψ quantifies the semantic consistencies of the mapped nodes in G and G'. It takes the following form:

$$M^s_\Psi = \frac{1}{|V|}\sum_{v\in V}\mathbb{1}(l_v \neq l_{\Psi(v)}) + \frac{2}{|V|(|V|-1)}\sum_{(v_1,v_2)}\mathbb{1}(i_{v_1} = i_{v_2}) \neq \mathbb{1}(i_{\Psi(v_1)} = i_{\Psi(v_2)}), \tag{2}$$

where $\mathbb{1} \in \{0,1\}$ is the indicator function, (v_i, v_j) denotes a node pair. The first term in the equation above requires that the mapped nodes have the same action labels, while the second term measures role consistencies. If two nodes in V have the same identifier of action executors, then the corresponding mapped nodes in V' should have the same identifier, too.

Time distance M^t_Ψ compares two event sketches in time scale. It is computed as the averaged Euclidean distance between temporal durations of all mapped node pairs. Note that temporal durations of all nodes in an event sketch are normalized in $[0,1]$ relative to the starting time of the first occurred action and ending time of the last occurred action, respectively.

Layout distance M^l_Ψ compares spatial distribution of nodes between two event sketches during event evolution. Let F^m_v denotes a $(|V|-1)\times 2$ displacement vectors from a node $v \in V_E$ to all the other nodes in V at a given time point t, then M^m_Ψ is given by

$$M^l_\Psi = \frac{1}{|V_E||\Gamma|}\sum_{v\in V_E}\sum_{t\in\Gamma}\|F^m_{v,t} \cdot D_{v,t} - F^m_{\Psi(v),t})\|_F, \tag{3}$$

where D is a 2×2 deformation matrix that minimizes the Frobenius norm in (3). The distance measure in M^l_{Ψ} is insensitive to the change of orientation and spatial scale of the action locations. The time point set Γ is uniformly sampled in $[0, 1]$.

Context distance M^c_{Ψ} compares the contextual information of the mapped nodes in both event sketches. The contextual information of an event sketch is denoted by a spatial context histogram, which has 8 components. For a node v, the spatial context histogram is defined as: the proportions of v's surrounding nodes that are near/far from, in the front/back/left/right of node v, the averaged significance/insignificance of v's surrounding nodes. The significance of an action is a scalar predefined based on domain knowledge (e.g. action kick is assigned a higher significance compared to the action walk in soccer videos, and action drop down should draw more attention in comparison with action walk or stand in surveillance videos) and scaled in $[0, 1]$. The contextual similarity is then calculated using χ^2 distance between histograms of each mapped node pair and then averaged.

4.3 Parameter learning

In the last section, we assume that the distance function $M_{\Psi}(G, G'; C, \omega)$ is a linear combination of five terms parameterized by C and ω. In our implementation we set C as a large positive value since a video containing fewer actions than query is unlikely to have a relevant evolution process. As for the choice of ω, we adopt a machine learning approach, named relevance feedback, to automatically learn the parameters, as deployed by [17].

The general idea of relevance feedback is to transform the feature space of relevance functions for search engines, so that the retrieved results can be more relevant to the users' choices [17]. To do this, for a given user query we collect a set of training videos, and ask the user to mark each video as relevant or not to the query. Given a specific instance of ω, we concatenate the weighted distances to the queries of each distance function and ω itself as features, and train a support vector machine classifier [18] using the training labels marked by users. With this classifier, we can then evaluate the classification accuracy on the testing set. A higher accuracy indicates the higher fitness of the current parameters with respect to the users' choices. For each instance of ω, we compute the classification performance for many user queries and average them to obtain the final accuracy. The parameters with the highest accuracy is finally chosen to aggregate the distance function.

5 Video event retrieval

This section introduces the implementation details to apply the event sketches for video retrieval. They include the preprocessing steps, design and usage of drawing board, and the technique to find near-optimal graph matching solution with a DDMCMC framework.

5.1 Preprocessing of database videos

The event sketch described in Section 4 represents the semantic-level information of the video events. Thus, for a raw video we need first to extract such high-level semantics. Specifically, we track each object using the source code of [19] in the raw video and recognize their actions on each frame. However, the strong occlusions and appearance changes in realistic environment render most of existing automatic trackers and detectors unreliable. Instead, we implement an annotation engine to perform this task. In the track stage, users are asked to annotate the object locations on a frame, and the objects are automatically tracked throughout the remaining frames. When strong inconsistencies of the tracked locations between adjacent frames are detected, the annotation engine stops the tracking and asks the users to re-initialize the object locations. To recognize the actions of objects, we implement the motion context descriptors [20] and classify the actions into 6 categories: run, walk, stand, kick, pick and rush. We manually corrected the misclassified actions, in a similar way with the procedure proposed in [21].

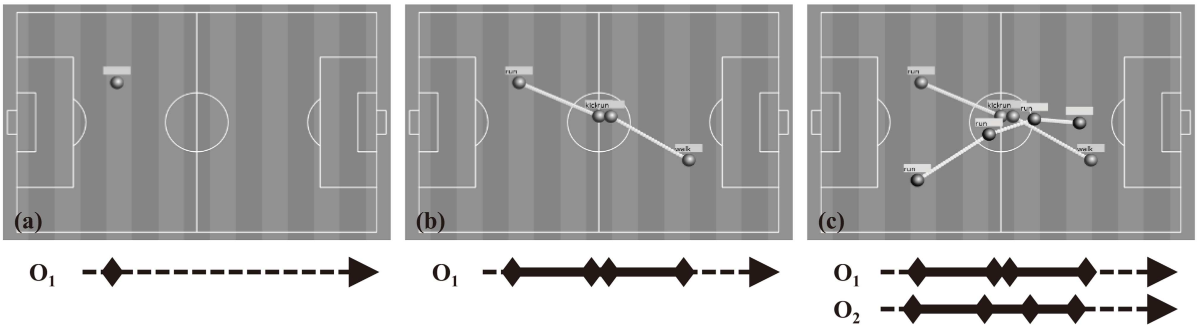


Figure 4 A simple usage of the drawing board. Here, the timelines for each object are listed below the figures. The diamonds along each timeline indicates the change of action labels or trajectories of the corresponding object. As a typical process, (a) the user firstly places an object on the plane, then (b) draws the trajectories of the object and labels the actions on some key time stamps, and (c) places other objects until finished.

After preprocessing, we obtain multiple object trajectories with action labels on each frame for a raw video. The event sketch for this video can be immediately constructed based on these trajectories, according to the definition in Section 4. The limitation of the process introduced above is that it requires necessary human labours to ensure the reliability of the annotations, which would be expensive for large-scale applications. However, under some well-controlled conditions (e.g. the surveillance setting where videos are taken by stable cameras and the occlusions are generally not severe), many recent approaches [22–24] can automatically achieve high tracking accuracy and action recognition rate thus human attentions are not necessarily required. Moreover, a line of crowd-source annotation mechanisms [25–27] were proposed to attract ordinary users to participate in on-line annotation, which could significantly improve the efficiency of the labeling procedure and meanwhile reduce the time and energy cost. However, both of these potential improvements are beyond the scope of this paper and referred here for the readers that need to learn more details.

5.2 Design and usage of the drawing board

Now we explain how to specify an event process using a drawing board. In our scheme, users are asked to customize an event on the space-time axes. This process is performed as follows: (1) add an object on a frame and specify its role and action label on this frame. The editing result is shown in Figure 4(a). Then (2) select the next key frame of this object and edit its new location and action label, as shown in Figure 4(b). (3) Repeat the Steps (1) and (2) until all the objects are finished editing, as shown in Figure 4(c). After editing, we can obtain the object trajectories with action labels as the same as those obtained in Subsection 5.1. The event sketch of this simulated event is thus constructed on these trajectories.

5.3 Event sketch matching

Given the event sketch of the query G and that of the database video G', the objective of matching is to find an injective mapping $\Psi : V \to V'$, under which the distance function M_Ψ can reach the smallest value. The brute-force way to do this is to linearly scan all the possible node mappings. However, the computational cost is growing extremely fast with time complexity $O(\min(|V|, |V'|)!)$. A lot of methods can be used for optimization, among which we follow the DDMCMC framework [28] to sample the searching space. The key intuition of DDMCMC is to incorporate the prior constraints in the data to guide the sampling process, so that to effectively avoid bad local minima.

In the matching process, we define the state variable x to denote the set of current matched node pairs. The state variable evolves in a growing style; it starts with the initial matched node pair x_0, and adds a new matched pair to the set one for a time. After a desired number of steps, we obtain a final mapping. We define the target distribution probability as $\pi(x) \propto \exp(-M_x(G, G')/T)$, where T is the temperature for the simulated annealing process in MCMC, and q is the data-driven transition proposal probability.

Algorithm 1 Event sketch matching

```
Input: Event sketch $G$ and $G'$; initial and final temperature $T_0$ and $T_f$; dropping factor $R$.
Steps:
 1: Select the node $v_0 \in V_E$ with the maximal action importance;
 2: For each mapping $v_0 \rightarrow v' \in G'$, sort the distances $M_{\Psi}(v_0, v')$ in ascending order and select the first $K$ matchings as
    initial state set $X_0$;
 3: repeat
 4:     Sample $x_0$ from $X_0$ randomly;
 5:     $\hat{x} = x_0$, $T = T_0$, $N = 0$;
 6:     while $T \geqslant T_f$ and $N \leqslant \min(|V_E|, |V'_E|)$ do
 7:         Denote the current matched pairs as $x = \{v_1 \rightarrow v'_1, v_2 \rightarrow v'_2, \ldots, v_k \rightarrow v'_k\}$. Let $V_r = V_E - \{v_k\}$, $V'_r = V'_E - \{v'_k\}$
            and $P = \{v \rightarrow v' | v \in V_r, v' \in V'_r\}$. Calculate $q$ using Eq. (4) by adding a new mapped pair $p \in P$ to $x$ one for a
            time;
 8:         Sample $x'$ according $q$ and increase $N$ by 1;
 9:         If random() $\leqslant r(x \rightarrow x')$ and $\pi(x) \leqslant \pi(x')$ then $\hat{x} = x' = x$;
10:         $T = T \cdot R$;
11:     end while
12: until max iteration
13: return $\hat{x}$ with highest target distribution probability.
```

We omit the detailed explanations of these variables here and refer the reader to [28] for more details. Here, we directly give the definition of q in our problem:

$$q(x \rightarrow x') \propto a \cdot \rho(x, x') + b \cdot \theta(x, x') + C_0. \tag{4}$$

In the formula (4), ρ and θ are the gain of action significance and role consistencies of the current state compared with the previous state, respectively. In details, ρ is defined as the difference between the sum of action significance of the matched nodes in x and x'. The role consistency gain θ is defined as follows: for each node pair in V, we find the matched pair in V', and check if the identifiers of action executors for each pair are the same. If the pair in V has the same identifiers while the pair in V' has not (and vice versa), we consider the mapping is not consistent, otherwise it is consistent. For the consistently mapped pair, we increase a counter by one. The θ is the difference of the counter values between the state x and the previous state x'. The controling parameters a and b are both set to 0.5 empirically, and C_0 is a positive constant that prevents the value from being negative. The transition proposal defined in (4) forces the algorithm to choose a state where actions are of high significance and performed by actors consistently.

The matching algorithm is summarized in Algorithm 1. Note that the defination of acceptance ratio r is similar to most MCMC algorithms and referred to [28]. We run many iterations for Algorithm 1 and select the candidate mapping with the highest target distribution probability as the best solution.

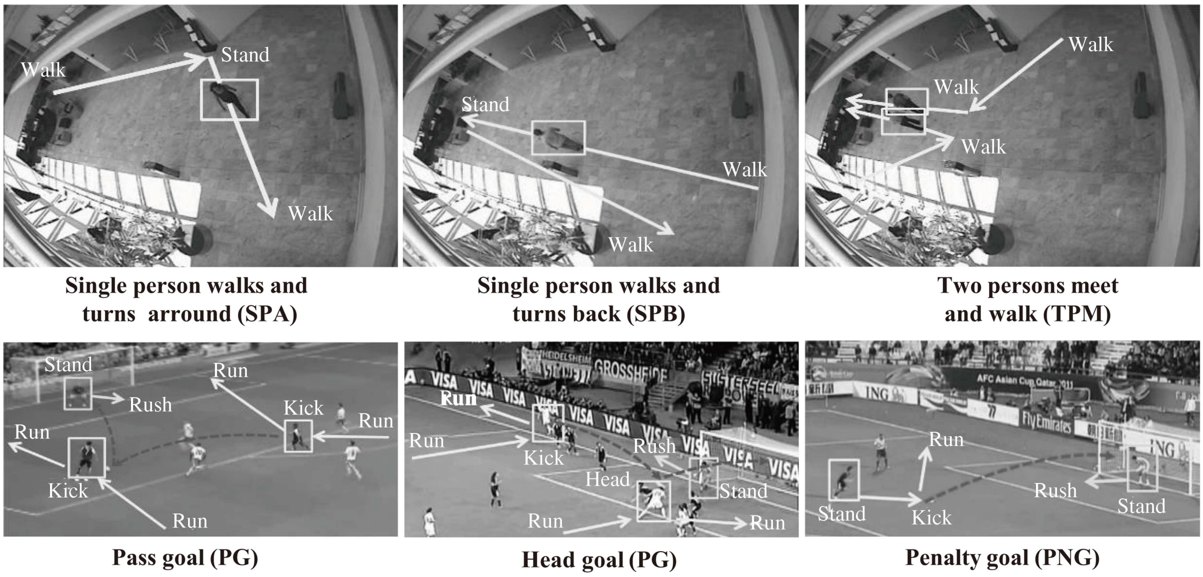
Time analysis. Like those of the general MCMC algorithms, the main computational bottleneck of our algorithm is the process of sampling Markov chains (i.e. Steps 6–11 in Algorithm 1). In the worst case, the time complexity of these sampling steps is $\min(|V_E|, |V'_E|)^3$. On a 3.4 GHz machine, the single-thread implementation of the proposed algorithm would take about 0.6 s for matching a drawing board sketch consisting of 5 actions with the event sketch of a typical database video with 20 actions. However, the sampling stage can be easily parallelized by simply running the Steps 6–11 in Algorithm 1 on many threads since that each iteration is independently performed.

6 Experiments

To evaluate our approach, we conduct experiments on two datasets. The first one is the CAVIAR dataset, which consists of videos scenes performed by several actors taken from a wide view[1]. For evaluation we segment the original 28 long videos to 51 clips and reclassify them into 3 categories: (1) single person walks and turns arround (SPA), (2) single person walks and turns back (SPB) and (3) two people meet

---

[1] CAVIAR project. http://homepages.inf.ed.ac.uk/rbf/CAVIAR.



Single person walks and turns arround (SPA)

Single person walks and turns back (SPB)

Two persons meet and walk (TPM)

Pass goal (PG)

Head goal (PG)

Penalty goal (PNG)

Figure 5 Typical scenes in CAVIAR dataset (top row) and soccer goals dataset (bottom row). Movements of objects are illustrated by solid arrows, and the corresponding action labels are annotated beside. For soccer goals dataset the moving directions of the ball are also visualized using dashed arrows for better understanding of temporal arrangements of the actions.

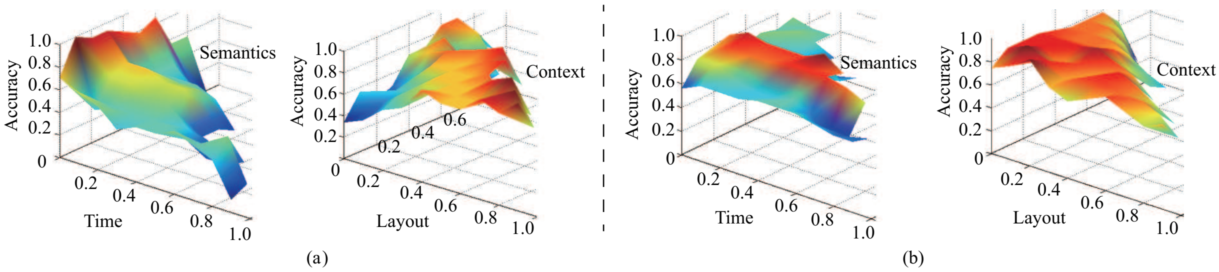


Figure 6 (Color online) Classification suface of (a) CAVIAR dataset and (b) soccer goals dataset.

and walk together (TPM). The typical examples of these categories are shown in the top row of Figure 5. Since that most existing datasets contain only simple events with little variabilities in their evolution processes, we additionally compiled a dataset of 50 goal events in real-world soccer videos. We group the videos into 3 categories according to the types of goals: (1) pass goal (attacker recieves the ball from his teammate and shoots for a goal, PG), (2) head goal (attacker recieves the ball and heads for a goal, HG) and (3) penalty goal (PNG). The bottom row of Figure 5 shows one instance for each of the three categories. Soccer goals dataset is suitable for our purpose, since each event category can be realized through a wide range of different evolution processes. Ground-truth annotations are provided in terms of bounding boxes and action labels around the players, as well as the category labels of the scenes.

6.1 Parameter selection

We set N and K in Algorithm 1 to 30 and 5 respectively in our experiments. To learn distance function parameters, drawing board queries of predefined event categories for the two datasets. We conduct relevance feedback as proposed in Subsection 4.3 to obtain classification accuracies of all combinations of parameters with the range from 0 to 1 at 0.1 increments. Figure 6 shows the visualized classification accuracy surface. For CAVIAR dataset the optimal parameter combination is $0.1, 0, 0.7, 0.2$ (which denote the relative weights of distance measures of semantics, time, layout and context, from left to right), reflecting that layout of object locations are the most discriminative compared to the other measures. For soccer goals dataset the choice is $0.3, 0.5, 0.1, 0.1$, where spatial information of players is weighted

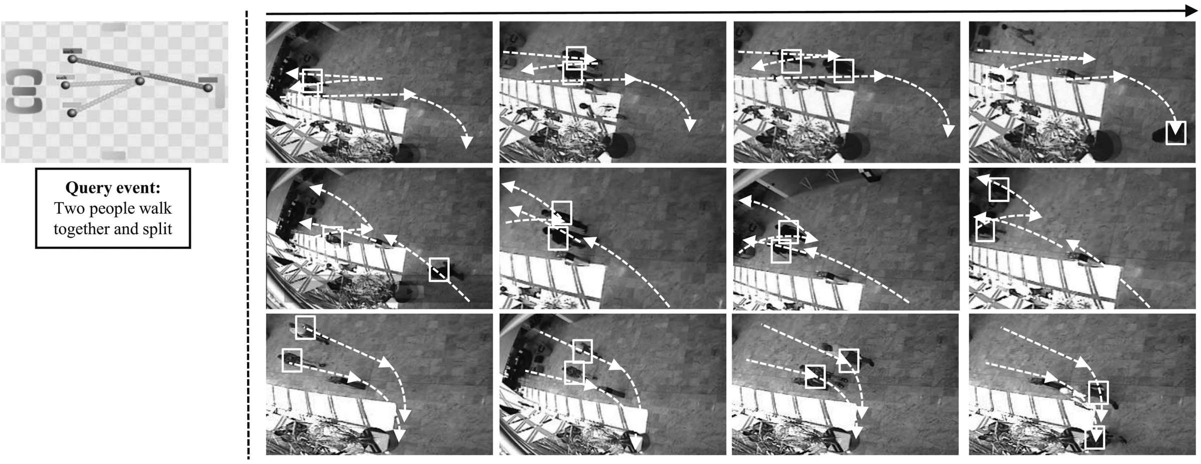


Figure 7 The top ranked results of query event two people walk together and split under combination of distance measures. The first row: first ranked result when all distance measures are considered. The second row: first ranked result when all except context distance measures are used. Third row: first ranked result when all except context and layout are used.

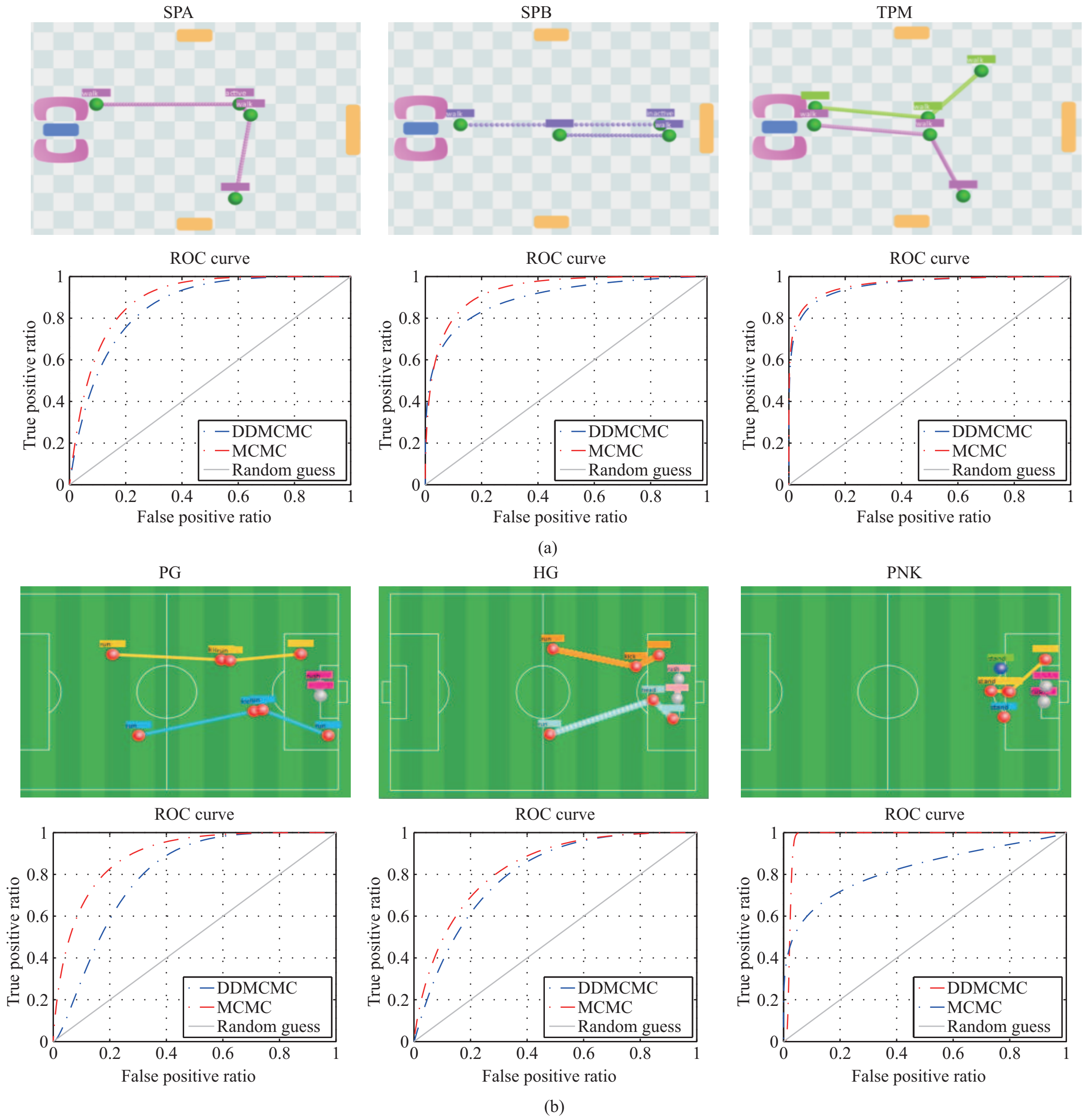
Table 1 Retrieval performance of CAVIAR dataset (left) and soccer goals dataset (right)

Method	SPA	SPB	TPM	Mean	Method	PG	HG	PNG	Mean
Proposed	92.2%	93.7%	96.3%	94.1%	Proposed	87.2%	78.0%	100.0%	88.4%
Baseline	88.9%	86.3%	96.3%	90.5%	Baseline	57.5%	57.1%	68.4%	61.0%

low while temporal property takes high weight, since that goal events vary greatly with regard to player locations but exhibit similar temporal arrangements of player behaviours. Moreover, from Figure 6 we observe that missing one or more distance measures will lead to unsatisfactory results in all conditions (this can be concluded from the fact that the most heated regions rarely appear along the edges of the surface in Figure 6). To clarify the necessity of the proposed distance measures, we designed a retrieval task as shown in Figure 7. In this example, we aim to query the event two people walk together and split (walk to the opponent direction). In the first row, all the distances are used and the relevant results are successfully retrieved. In the second row where the context distance is neglected, we see that the retrieved results have similar spatiotemporal layouts, while the directions are wrong due to missing context (in the query the two agents are spliting, while in the retrieved video agents are approaching). In the third row where both the context and layout distance are missed, the first ranked event would be irrelevant to the query since that the temporal arrangement of actions in this event is much more similar to the query than other relevant videos.

6.2 Retrieval performance

To quantitatively analyse the performance of our retrieval system, we compute the ROC curves for each kind of query and show them in Figure 8. The mean average precision (mAP) is summarized in Table 1. To evaluate effects of the DDMCMC graph matching algorithm, we implement a baseline by replacing the transition proposal probabilities to uniform probabilities, which result in the general MCMC algorithm. It is easy to observe that the proposed DDMCMC algorithm can achieve much better average precision. Particularly, for the event category with little variabilities in structures (e.g. penalty kick), the improvements can be very significant. To understand the superiority of the proposed algorithm, we illustrate the averaged distance between the query and the database videos with respect to the number of iterations in Figure 9. It can be seen that the proposed approach converges faster than the baseline, and can reach better local minima.



(a)

(b)

Figure 8 ROC curves on two datasets. (a) CAVIAR dataset; (b) soccer goals dataset.

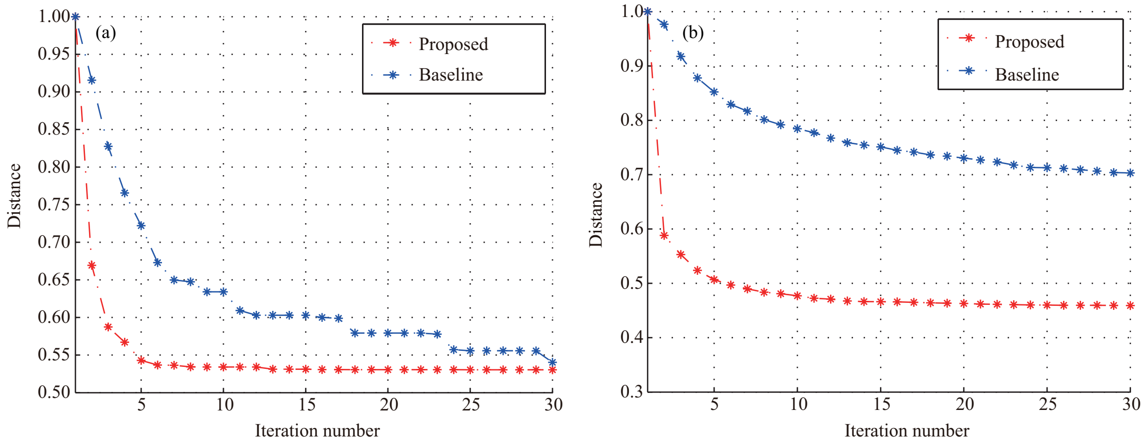


Figure 9 Study of algorithm convergence. (a) CAVIAR dataset; (b) soccer goals dataset.

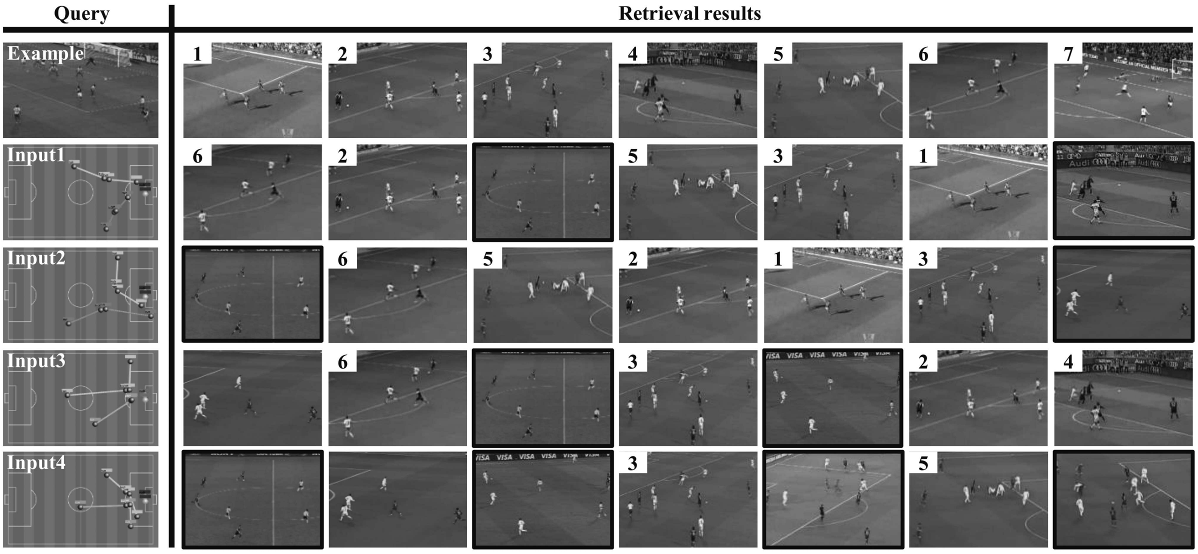


Figure 10 User study of our method. The first input was made by the dedicated labeler, while the others were from ordinary users. For each result retrieved by user queries, the corresponding rank in the results retrieved by video example is shown in the top-left corner of the frame. False alarms are marked by black boxes around each image.

6.3 User study

We conducted the user study to evaluate the robustness of the proposed retrieval system in handling the same query given by different users. To this end, we selected a video example from PG class and manually labeled the objects in it as query, and retrieved the top seven results. Then, we asked a professional labeler and three ordinary users to watch this example video and draw the evolution processes on a board, resulting in four drawing board queries of the same event. The retrieved results of these queries are shown in Figure 10. The query input1 is made by the professional labeler, and the retrieved results can match well with those for the example video. The query input2 has a similar layout with that of the first one, but incorrectly draws the overall orientation of the event. In this scenario, our model can still sucessfully retrieve a large proportion of the relevant results. The other queries input3 and input4 are much more challenging since that they may lack enough input actions and may strongly deviate from the actual event layout. In these cases, our approach can still retrieve several relevant videos, however the failures indicate that different styles of user queries still contain challenging subtleties that are not easily captured by the proposed system.

For a clear observation, we illustrate the retrieved videos for two queries from the SPA category in CAVIAR dataset and the PG category in soccer goals dataset in Figure 11, respectively.

7 Conclusion

In this paper, we propose a novel event representation model named event sketch to capture high-level event properties of different aspects, and incorporated it into a SBR system that allows users to draw events on a board as queries. Event sketches are constructed on both the queries and database videos and compared using a weighted distance function, whose parameters are learned through relevance feedback. Experiments conducted on datasets with large variability of object movements and diversity of event categories demonstrate the effectiveness of the proposed approach.

As a preliminary study on sketch-based retrieval of complex video events which does not receive enough attention before, our method suffers time cost when the size of database is large and the query is complex. For acceleration, it would be interesting to develop indexing mechanisms and distance measures that are more efficient to compute. A promising direction in this context is to design coding techniques of the redundant semantic attributes in the video to facilitate inference, borrowing from the idea of low-level

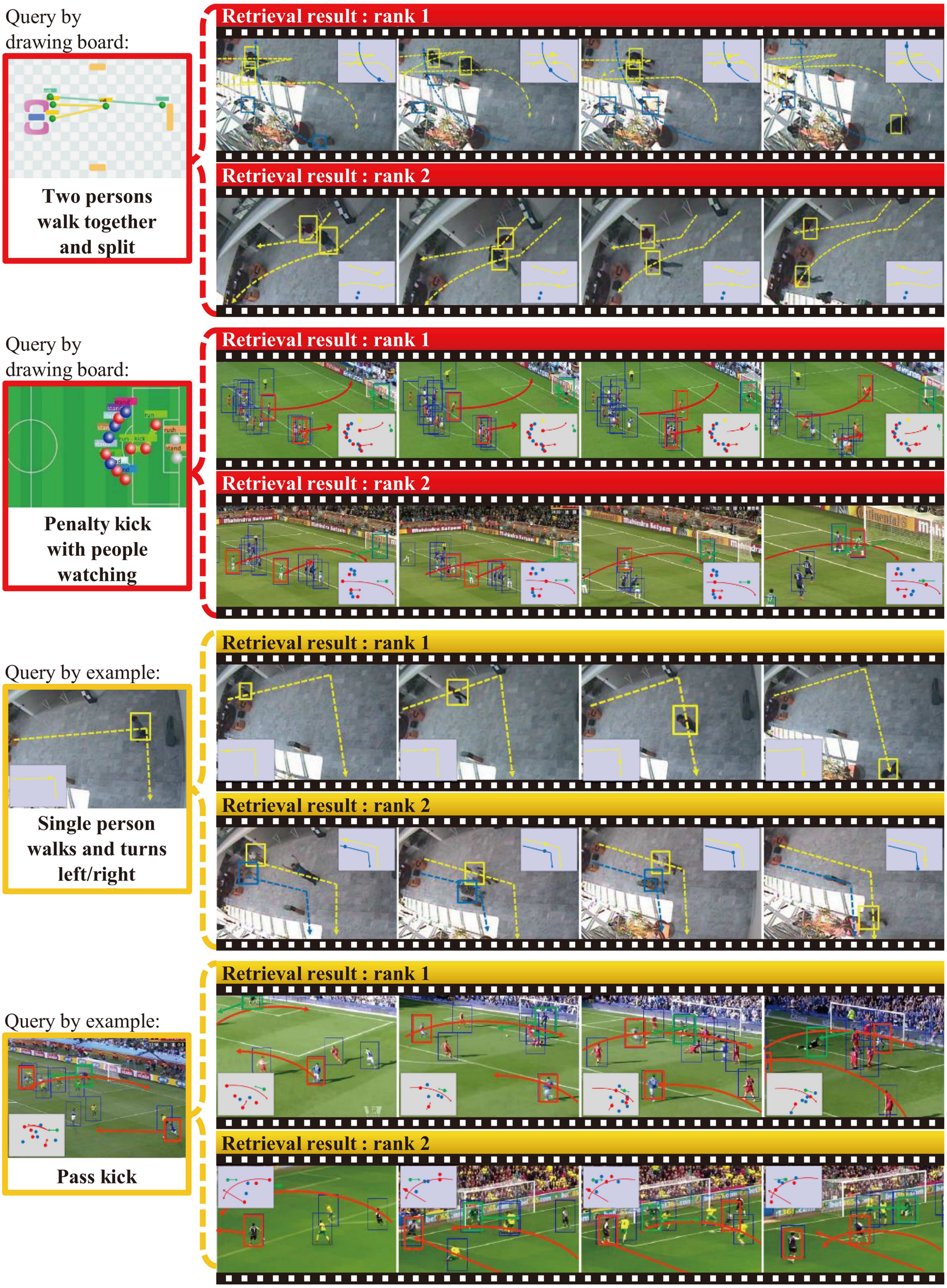

Query by drawing board:
Two persons walk together and split
Retrieval result : rank 1
Retrieval result : rank 2
Query by drawing board:
Penalty kick with people watching
Retrieval result : rank 1
Retrieval result : rank 2
Query by example:
Single person walks and turns left/right
Retrieval result : rank 1
Retrieval result : rank 2
Query by example:
Pass kick
Retrieval result : rank 1
Retrieval result : rank 2

Figure 11 Top 2 ranked retrieval results of two video queries with their corresponding drawing board queries. Retrieved event details are marked by yellow arrows and boxes for CAVIAR dataset. For soccer goals dataset, movements of interest players are marked red (attacker) and green (keeper) respectively. Vertical views are provided in the corner of each frame.

video coding. Also, it would be interesting to combine the high-level representation with low-level visual cues extracted from the videos to further reduce the human efforts as well as to improve the retrieval performance.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 61532003, 61325011, 61421003), National High Technology Research and Development Program of China (Grant No. 2013AA013801), and Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20131102130002).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Yuan J, Zha Z J, Zheng Y T, et al. Learning concept bundles for video search with complex queries. In: Proceedings of International Conference on Multimedia, Scottsdale, 2011. 453–462

2 Bao L, Cao J, Zhang Y, et al. Explicit and implicit concept-based video retrieval with bipartite graph propagation model. In: Proceedings of International Conference on Multimedia, Firenze, 2010. 939–942

3 Ulges A, Schulze C, Koch M, et al. Learning automatic concept detectors from online video. Comput Vis Image Underst, 2010, 114: 429–438

4 Hu R, Collomosse J. Motion-sketch based video retrieval using a trellis levenshtein distance. In: Proceedings of International Conference on Pattern Recognition, Istanbul, 2010. 121–124

5 Collomosse J P, McNeill G, Qian Y. Storyboard sketches for content based video retrieval. In: Proceedings of International Conference on Computer Vision, Kyoto, 2009. 245–252

6 Hu R, James S, Collomosse J. Annotated free-hand sketches for video retrieval using object semantics and motion. In: Proceedings of the 18th International Conference on Advances in Multimedia Modeling. Berlin: Springer, 2012. 473–484

7 Hu R, James S, Wang T, et al. Markov random fields for sketch based video retrieval. In: Proceedings of International Conference on Multimedia Retrieval, Dallas, 2013. 279–286

8 Zhou R, Chen L, Zhang L. Sketch-based image retrieval on a large scale database. In: Proceedings of International Conference on Multimedia, Nara, 2012. 973–976

9 Eitz M, Hildebrand K, Boubekeur T, et al. Sketch-based image retrieval: benchmark and bag-of-features descriptors. IEEE Trans Vis Comput Graph, 2011, 17: 1624–1636

10 Cao Y, Wang C, Zhang L, et al. Edgel index for large-scale sketch-based image search. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, Colorado, 2011. 761–768

11 Lu D, Ma H, Fu H. Efficient Sketch-based 3D shape retrieval via view selection. In: Proceedings of Advances in Multimedia Information Processing–PCM, Nanjing, 2013. 396–407

12 Xu H, Wang J, Hua X S, et al. Interactive image search by 2D semantic map. In: Proceedings of International Conference on World Wide Web, Raleigh, 2010. 1321–1324

13 Yu G, Yuan J, Liu Z. Action search by example using randomized visual vocabularies. IEEE Trans Image Process, 2013, 22: 377–390

14 Lan T, Wang Y, Mori G, et al. Retrieving actions in group contexts. In: Proceedings of the 11th European Conference on Trends and Topics in Computer Vision–Volume Part I. Berlin: Springer, 2012. 181–194

15 Ma X, Chen X, Khokhar A, et al. Motion trajectory-based video retrieval, classification, and summarization. In: Video Search and Mining. Berlin: Springer, 2010. 53–82

16 Cheng Z, Qin L, Huang Q, et al. Human group activity analysis with fusion of motion and appearance information. In: Proceedings of International Conference on Multimedia, Scottsdale, 2011. 1401–1404

17 Fisher M, Savva M, Hanrahan P. Characterizing structural relationships in scenes using graph kernels. ACM Trans Graph, 2011, 30: 34

18 Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Tech, 2011, 2: 27

19 Pérez P, Hue C, Vermaak J, et al. Color-based probabilistic tracking. In: Proceedings of European Conference on Computer Vision, Copenhagen, 2002. 661–675

20 Tran D, Sorokin A. Human activity recognition with metric learning. In: Proceedings of European Conference on Computer Vision, Copenhagen, 2008. 548–561

21 Jiang K, Chen X, Zhang Y, et al. Video event representation and inference on and-or graph. Comput Animat Virtual Worlds, 2012, 23: 145–154

22 Ribeiro P C, Santos-Victor J. Human activity recognition from video: modeling, feature selection and classification architecture. In: Proceedings of International Workshop on Human Activity Recognition and Modelling, Oxford, 2005. 61–78

23 Ben Shitrit H, Berclaz J, Fleuret F, et al. Tracking multiple people under global appearance constraints. In: Proceedings of International Conference on Computer Vision, Barcelona, 2011. 137–144

24 Xie Y, Chang H, Li Z, et al. A unified framework for locating and recognizing human actions. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, Colorado, 2011. 25–32

25 Hua X-S, Qi G-J. Online multi-label active annotation: towards large-scale content-based video search. In: Proceedings of International Conference on Multimedia, Vancouver, 2008. 141–150

26 Ahn L-V, Dabbish L. Labeling images with a computer game. In: Processings of SIGCHI Conference on Human Factors in Computing Systems, Vienna, 2004. 319–326

27 Sorokin A, Forsyth D. Utility data annotation with amazon mechanical turk. In: Workshops of International Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–8

28 Lee J, Cho M, Lee K M. A graph matching algorithm using data-driven markov chain monte carlo sampling. In: Proceedings of International Conference on Pattern Recognition, Istanbul, 2010. 2816–2819