Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies

Lamberto Ballan · Marco Bertini · Alberto Del Bimbo · Giuseppe Serra

Abstract In this paper we present a framework for semantic annotation of soccer videos that exploits an ontology model referred to as Dynamic Pictorially Enriched Ontology, where the ontology, defined using OWL, includes both schema and data. Visual instances are used as matching references for the visual descriptors of the entities to be annotated. Three mechanisms are included to support effective annotation: *visual instance clustering*—to cluster instances of similar patterns, *prototype selection*—to select one or more visual representatives of each cluster, *dynamic cluster updating*—to update clusters and prototypes whenever new knowledge is presented to the ontology. Experimental results show the capability of performing semantic annotation of entities that exhibit a variety of complex changes in visual appearance or of events that show complex motion patterns in the same shot. SWRL rules are used to perform rule-based reasoning over both concepts and concept instances, to improve the quality of the annotation.

Keywords Semantic video annotation · Dynamic pictorial ontology · Content descriptor matching · Ontology reasoning · Sports video analysis

L. Ballan · M. Bertini (✉) · A. Del Bimbo · G. Serra
Media Integration and Communication Center,
University of Florence, Florence, Italy
e-mail: bertini@dsi.unifi.it

L. Ballan
e-mail: ballan@dsi.unifi.it

A. Del Bimbo
e-mail: delbimbo@dsi.unifi.it

G. Serra
e-mail: serra@dsi.unifi.it

Published online: 07 August 2009

1 Introduction

Automatic annotation of video content at the semantic level has received a significant attention from the research community in the recent years [24, 41], to face the growing request for search and retrieval by content of interesting elements resulting from the explosive growth of video production. Some of the fields of application where there is the greatest impact on industry are surveillance, news and sport videos. In particular, considering soccer videos, it can be observed that in the last years TV coverage of soccer events has achieved an enormous level of worldwide popularity. Just as an example, the television coverage of the 2006 FIFA World Cup was aired in a total of 43,600 broadcasts across 214 countries and territories, generating a total coverage of 73,072 hours [17]; this is an increase of 76% on the 2002 event and a 148% increase on 1998. Broadcasters commonly produce annotations of relevant events that occur in these videos to assembly video summaries of events for their sport programmes, or to provide the sequences of only certain actions for mobile video services. These annotations are also used for the creation of the broadcasters' archives (the so called *posterity logging*), an operation that is required for proper asset management. Therefore there are strong motivations for the development of tools and systems that automatically detect and recognize the most relevant events (commonly defined as *highlights*) and entities (e.g. players).

Generally speaking, image analysis and pattern recognition has been used to provide the capability to extract low-intermediate level features and perform their classification to ease indexing of video archives. However, indexing based on such features often does not meet the user's information needs due to the *semantic gap* between the information that can be extracted from the visual data, and its interpretation by a user in a given context. Recently, ontologies complemented with image analysis and pattern recognition have been regarded as the appropriate tool to solve the semantic gap and effectively support semantic annotation and retrieval of multimedia content. Ontologies have been originally developed with the aim of providing a common vocabulary that encodes levels of semantics to overcome semantic heterogeneity for information, expressed in a language that supports reasoning. An ontology can be divided into two parts. One part is *schema*, which refers to a structured set of linguistic terms expliciting concepts, concept properties and relations, and their associated definitions so as to provide a formal description of something. The other part is *data* and consists of instances of concepts or individuals. *Generic ontologies* are essentially definitions of concepts in general, such as DOLCE [30] and Upper Cyc Ontology [25] or vocabularies in a wider sense like WordNet [16] and the Large-Scale Concept Ontology for Multimedia (LSCOM) [32] commonly used within the TRECVID benchmark [39]. Instead, *domain ontologies* provide a description of some specific application domain, such as the Dublin Core ontology [13] for multimedia digital libraries, ontologies for paintings [27] and soccer [11, 43], and many others.

In the last years different approaches have been proposed for the use of ontologies for video annotation. Many researchers have built integrated systems where the ontology provides the conceptual view of the domain at the schema level, and appropriate classifiers are used to classify entity or event in the nearest concept of the ontology. Once the observations are classified, the ontology is exploited to have a richer semantic annotation, establishing links to other concepts and disambiguating the results of classification [20, 21, 37, 42, 46, 47, 54]. Other researchers have directly included in the ontology concept instances, so as to incorporate some explicit parametric representation of the visual knowledge [8, 10, 12, 38]. In the attempt of having richer annotations, other authors have explored the usage of reasoning over multimedia ontologies. In this case spatio-temporal relationships between concept occurrences are analyzed so as to distinguish between scenes and events and provide more fitting and complete descriptions [5, 15, 18, 33, 34]. Most of these experiences have nevertheless covered only simple cases, where the event can be asserted from a few consecutive keyframes with static entities. Typically, all these solutions do not apply successfully when entities exhibit a variety of complex changes in shape and visual appearance or events show complex motion patterns in the same shot such as happens in sports videos. An example of this case is the *attack action* event in soccer. Typically it develops on several shots and, for each shot, the action could be displayed in a variety of modes, due to the many possible combinations of the motion pattern of the player, the playground zone in which the action occurs, the number of players involved.

In this paper, we present an ontology-based system for semantic annotation of soccer videos that attempts to solve the above referred problems. In our case, the ontology follows the principles of the Dynamic Pictorially Enriched Ontology model [8] with new additions and improvements. Visual instances are part of the ontology: they provide a matching reference for the visual descriptors of the observed entities, and permit reasoning on instance values as well. Three mechanisms are included in the ontology to support the effective annotation of entities with large intra-shot changes: *visual instance clustering*, *prototype selection*, and *dynamic cluster updating*. The first two mechanisms are used to associate to each concept of the ontology a number of visual instances, each of which is the representative of a cluster of similar patterns in which the concept can manifest itself; the third mechanism is used to update such clusters (and their representatives) with the new observations that are presented to the ontology for their annotation. External classifiers are used to assess most of the concepts that refer to entities with little changes in their appearance, such as playground, playground line, playground zones, etc. Descriptor matching with the visual prototypes of the ontology is used instead to assess entities and events with large changes, such as player faces, advertising billboards, soccer highlights, etc. Since both concepts and concept instances in the ontology are defined using the Web Ontology Language (OWL), the Semantic Web Rule Language (SWRL) has been used to effectively perform reasoning over both concepts and concept instances, so as to disambiguate the results of the classification or derive new semantic annotations of parts of the clips.

The rest of the paper is organized as follows. Section 2 provides an overview of the related works on automatic annotation and retrieval of sports videos. The ontology model, the mechanism for visual instance clustering, prototype selection, cluster updating and spatial/temporal reasoning are discussed in Section 3. In Section 4 we discuss in some detail the feature detectors and classifiers used to create instances of the main concepts of the ontology. In Section 5 we describe in detail the rule-based reasoning used to refine classification or to derive richer semantic annotation. In Section 6 we present experimental results and comment on performance improvements obtained by exploiting visual prototype updating and spatial/temporal reasoning over instances. Conclusions and future research are expounded in Section 7.

2 Related works

Most of the existing works addressing semantic annotation of sports videos have been devoted to parse the video stream in order to detect various entities and events typical of the sport under consideration (like players or shots on goal in soccer videos). Usually, knowledge of the sport domain and of the typical broadcasting production rules are exploited to achieve the detection of a number of domain-specific events. Comprehensive reviews can be found in [24, 53]. Related works on soccer video annotation can be grouped in shot classification, entities recognition (e.g. players and objects, such as billboards), events and highlights recognition.

Shot classification in play/break events has been proposed by Xu et al. [50]. Their method is composed by two steps; in the first step they classify each sample frame into global, zoom-in and close-up views using an unique domain-specific feature, grass-area-ratio. Then heuristic rules are used in processing the sequence of views, and obtain play/break status of the game. In [51] a statistical framework that uses Hidden Markov Models (HMMs) for the recognition of replay in soccer videos is proposed. The method uses a two step process, were shots are initially classified in three types of view and replay, followed by a refinement of replay shots. Xie et al. in [48] improve their previous work [50] including an HMM to perform automatically play/break segmentation. They use a set of features extracted from the compressed domain, such as dominant color ratio and motion intensity. Recently Mei and Hua [31] proposed a mosaic-based framework in which structure and events can be detected. A mosaic is generated for each shot and shots are hierarchically classified in global view and close-up view; global view shots are then analyzed because of the assumption that "most of the events in sports video happen in global view shots".

Object detection and tracking, to describe events of a soccer game, has been proposed by Utsumi et al. [44]. The proposed method uses color rarity and local edge properties to detect various in-field and out-of-field objects, and heuristic rules are used to classify players in order to track them. Players recognition, based on the use of the jersey's numbers, has been proposed in [52]; each image is segmented in color homogeneous regions, and Zernike moments are extracted from pipe-like regions to recognize numbers. In [28] long-range views are recognized using a decision tree algorithm, dominant color and geometric features; players are then detected using a Viola & Jones detector. Players and referee are clustered using the color of the jerseys and a Markov chain Monte Carlo process is used to track the players. Watve et al. [45] proposed a system for the detection and recognition of billboards in soccer videos. An initial shot classification in long-range and close-up shots is performed to adapt, accordingly to the view, set of predefined thresholds used to detect the billboards. Frames are processed to select candidate regions based on vicinity with the playground and presence of hue and gray-level edges; billboard models are described using color and texture features.

The approaches for event and highlight recognition deal with the problem of detecting and recognizing the salient events of a match. Classification of three placed kick highlights (free, corner and penalty kick) using HMMs has been proposed in [1]. A three states left-to-right model for each highlight is used, based on the consideration that the states correspond well to the evolution of the highlights in term of characteristic content. The features used are the view type (very long shot, long shot and medium shot view), camera pan and tilt (quantised in 5 and 2 levels).

Leonardi and Migliorati [26] have used MPEG motion vectors, playground shape and players position with HMMs to detect soccer highlights. In [29] frames and clips are classified in a set of views (close-up, playground center and goal areas, medium views) using Bayesian Networks. Then, in order to identify shot on goals the proposed system groups the clips that are preceding and following the clips classified as showing the goal areas. If a certain pattern of clip views and values of the feature that corresponds to the position of the field end line is found, then a shot on goal is determined to be present. In [2] Finite State Machines (FSMs) have been employed to detect the principal soccer highlights, such as shot on goal, placed kick, forward launch and turnover, from a few visual cues, such as playground position, speed and camera direction, etc. In [14], Ekin et al. performed generic highlight detection in soccer video using both shot sequence analysis and shot visual cues. In particular, they assume that the presence of highlights can be inferred from the occurrence of one or several slow motion shots and from the presence of shots where the referee and/or the goal box is framed. The use of FSMs to model highlights and events has been proposed also in [6]; scored goal, foul and generic play scenes have been modeled using four types of views (e.g. in-field, slow motion, etc.) for the states of the FSMs and transitions are determined by some audio-visual events such as the appearance of a caption or the whistle of the referee. Experiments have been performed using a set of manually annotated views and audio-visual events.

In [22] Dynamic Bayesian Networks are used to model the contextual information provided by the timeline. It is argued that HMMs are not expressive enough when using a signal that has both temporal and spatial information; moreover, DBNs allow a set of random variables instead of only one hidden state node at each time instance. In this work five events are defined: shoots, corner kick, free kick, play and break. Events are modeled considering five types of primitive scenes: close-up, medium view, playground center, playground three-quarter, playground goal area. The percentage of playground area in the frame, the size of players, the presence of midfield lines and end lines are used as observable features. Sadlier and O'Connor proposed [35] a method for event detection in field sports such as soccer, hockey and rugby. The proposed system classifies shots as *eventful* and *noneventful*. Shots are filtered considering a simple rule that checks the presence of close-up shots within a short time after the end of the shot. The classification of shots is performed using a SVM; each shot is represented using a feature vector composed by five values, one for each audio/visual feature used. Snoek et al. [40] performed annotation of sport highlights using statistics derived from a face detector, superimposed captions, teletext and excited speech recognition; temporal relations between the detected concepts are modelled using Allen's logic. They compared three classification techniques and showed that, if the semantic gap between the features used and the concept that is searched is narrow, even the weakest classifier performs well with this approach. Finally, in [49] Xu et al. present a framework for semantic annotation and personalized retrieval of sports video that extend the traditional audio/video approach including a web-casting text analysis. This approach is based on the analysis and alignment of web-casting text with broadcast videos, in particular soccer and basketball, and it provides a personalized retrieval scheme to support both general and specific queries according to user's preference related to particular game, team or player.

3 Video annotation framework

In our framework we implement an ontology schema to describe the soccer video domain; Table 1 reports the most important concepts and hierarchical relations described in our ontology. These concepts can be classified in *concrete* and *abstract*. *Concrete concepts* represent entities and events that have some manifestation in the reality (e.g. players, shot on goal, etc.), while *abstract concepts* represent more immaterial elements, like game or result. We extend the traditional linguistic ontology model with visual information, following the Dynamic Pictorially Enriched Ontology model [8]. Visual data is included in the ontology as visual instances applied to the concrete concepts of the schema; each visual instance includes object identifier, time label, content descriptors and link to the raw data (the images or clips in the repository). Both concepts and concept instances are defined using the standard Web Ontology Language (OWL) so that the ontology can be easily reused and shared.

Concrete concepts can be characterized by little or large changes in visual appearance. Therefore, we have defined two different methods to deal with their representation and automatic annotation; *i)* visual instances of concepts that have little changes in appearance (e.g. playground, text captions, jersey numbers, etc.) are instantiated by external classifiers, while *ii)* visual instances of concepts with large changes are classified using descriptor matching with visual prototypes of the ontology (defined by visual instance clustering and prototype selection). For this reason, all the concrete concepts are assigned to one of this two classes during the creation of the ontology schema. Figure 1 illustrates a partial schema of our ontology, showing the two kind of concrete concepts in different colors. The automatic annotation of a video shot—obtained segmenting soccer videos with the LTD (Linear Transition Detection) algorithm [19]—is performed following the ontology schema; in this way, the shot is initially classified as break or play and then only the related concepts are searched to obtain a richer annotation.

In the following subsections, we describe in details the ontology-based classification methods for all the different kind of concepts, and the proposed reasoning approach used to disambiguate the results of the classification or to derive new semantic annotations.

3.1 Annotation using ontologies

As previously introduced, visual instances of concepts that have little changes in shape, appearance or motion patterns are instantiated by external classifiers. These classifiers are related to concrete concepts such as playground elements, text captions and jersey numbers. Instances of these concepts are included in the ontology, so as to support reasoning over instance values or to be used as indexes to access the raw data in the repository.

Instead, concepts with large changes—like *player's-face*, *advertising-billboard* and *play action*—are described by one or more reference visual instances. These instances, called *visual prototypes*, are created initially (during the creation of the ontology) using a training set of already annotated data, performing unsupervised Fuzzy-C means clustering on their visual descriptors. As a consequence, several clusters in general exist for each concept, and each cluster roughly corresponds to one modality in which that concept can manifest itself in the reality. The visual instance

Table 1 Main concepts of the ontology for soccer game domain

Concept	Description
Video	The video of a soccer match
Shot	A video sequence filmed continuously by a camera
Clip	A video sequence (may comprise several shots)
Superimposed text	Superimposed text that provides information such as players' names or match score
Stadium	Sport arena with tiers of seats for the spectators
Spectators-seats	Stadium benches and seats
Advertising-billboard	Board with advertising media positioned next to the playground
Goal-post	The structure consisting of two posts, crossbar and net into which all goals are scored
Playground	The area where the soccer match is played
Goal-post-area	The area of the playground near the goal post
Playground-line	Line that delimits parts of the playground
Midfield-line	The line that divides the field in half along its width
Face	A human face
Player's-face	The face of a player that has been identified
Unknown-face	The face of a player that has not been identified
Player	The athlete participating to the match
Jersey	Distinctive shirt worn by the players
Jersey-number	Distinctive number assigned to each player
Play action	Event/action that is part of the match, usually filmed by the main camera
Kickoff	Action at the start/restart of the game
Forward-launch	Quick action including a pass of a team towards the field of the opponent team
Attack	Continuous action including a pass of a team towards the field of the opponent team
Shot-on-goal	Action in which a player kicks the ball to the opponent's goal post to score a goal
Scored-goal	Action in which a goal is scored
Placed-kick	Free kick in any part of the field
Placed-kick-near-goal-post	Action played near the goal post, including penalty, corner and free kick
Foul	Action that violates the rules for which a referee assesses a placed-kick
Throw-in	Action where a player throws the ball with two hands, after that the opposite team threw the ball across a sideline
Unknown-play	Non-classified play action
Break-action	Event/action that is not part of the match, usually filmed by side cameras
Crowd	Group of spectators
Players-medium-view	View of a player, framed from a side camera near the playground
Players-group	Group of players
Player-close-up	Player close-up view
Unknown-break	Non-classified break action

The indentation indicates that a concept is a specialization of another concept

at the center of a cluster is taken as the visual prototype of that cluster (*prototype selection*). Automatic annotation of a new shot is performed through comparison between its visual descriptors and visual prototypes. The matching of this shot with a

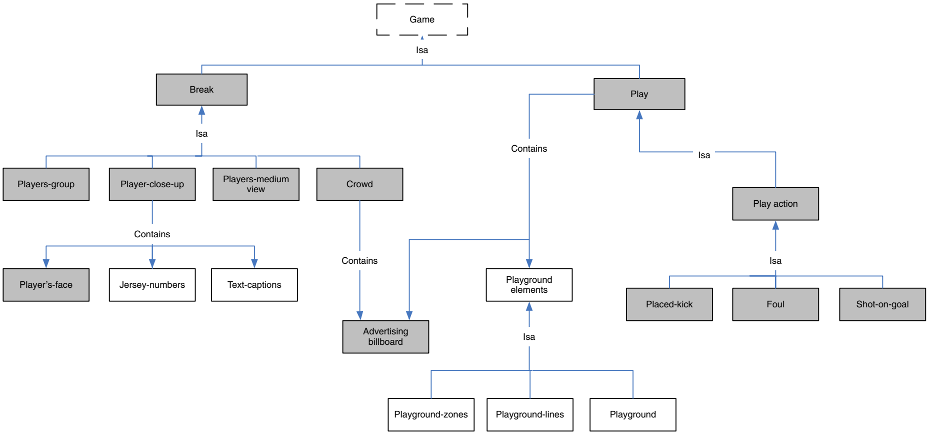

Fig. 1 Partial schema of our soccer video ontology: concrete concepts are depicted using *continuous box* while abstract concepts are depicted using *dashed box*; concepts with large changes in appearance are shown in *gray* while concepts with little changes are shown in *white*

visual prototype determines the instantiation of a new visual instance of the concept related to this visual prototype. A special concept, the *unknown-entity*, is created to include all the instances not yet assigned to any concepts. In consideration of the large variety of modes in which entities or events may appear, any new entity/event that is presented to the ontology for annotation can be regarded as new knowledge that is added to the ontology. According to this, every time a new visual instance is associated to a concept, cluster updating is performed over all the clusters of

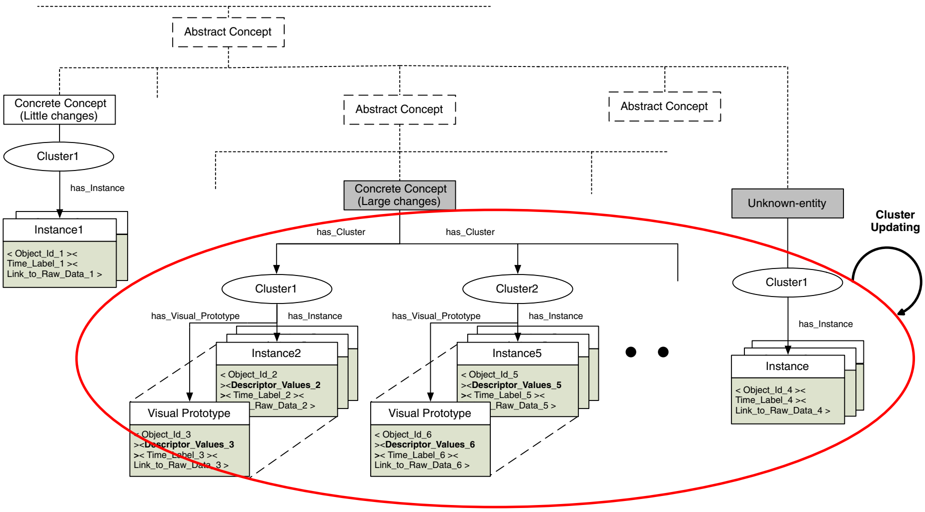
Abstract Concept
Concrete Concept (Little changes)
Abstract Concept
Abstract Concept
Cluster1
has_Instance
Instance1
< Object_Id_1 >< Time_Label_1 >< Link_to_Raw_Data_1 >
Concrete Concept (Large changes)
Unknown-entity
Cluster Updating
has_Cluster
has_Cluster
Cluster1
Cluster2
Cluster1
has_Visual_Prototype
has_Instance
has_Visual_Prototype
has_Instance
has_Instance
Instance2
< Object_Id_2 ><Descriptor_Values_2 >< Time_Label_2 >< Link_to_Raw_Data_2 >
Instance5
< Object_Id_5 ><Descriptor_Values_5 >< Time_Label_5 >< Link_to_Raw_Data_5 >
Instance
< Object_Id_4 >< Time_Label_4 >< Link_to_Raw_Data_4 >
Visual Prototype
< Object_Id_3 ><Descriptor_Values_3 >< Time_Label_3 >< Link_to_Raw_Data_3 >
Visual Prototype
< Object_Id_6 ><Descriptor_Values_6 >< Time_Label_6 >< Link_to_Raw_Data_6 >

Fig. 2 Schematization of the proposed ontology model

that concept and the unknown-entity (*dynamic cluster updating*). This mechanism permits to represent more effectively the variety of appearances and motion patterns of the entities and provides a form of *temporal evolution* for the knowledge in the ontology. In fact, during cluster updating previously unknown labeled entities can be assigned to some cluster, or new clusters are created as a result of the new instance. The schematization of the ontology model is reported in Fig. 2, showing visual instances of both concrete concept classes and mechanisms for *visual instance clustering*, *prototype selection* and *dynamic cluster updating*.

Figure 3a shows an example of cluster updating of face tracks. As a new face sequence is presented, it is associated with the nearest visual prototype and clusters are updated accordingly. In this case, the effect of the re-clustering procedure is that a previous unknown face track is now correctly annotated (see *Cluster 1*). Instead, Fig. 3b shows an example of cluster splitting for highlight shots. As a consequence of the new shot, *Cluster 2* is split in two different clusters described by their new visual prototypes.

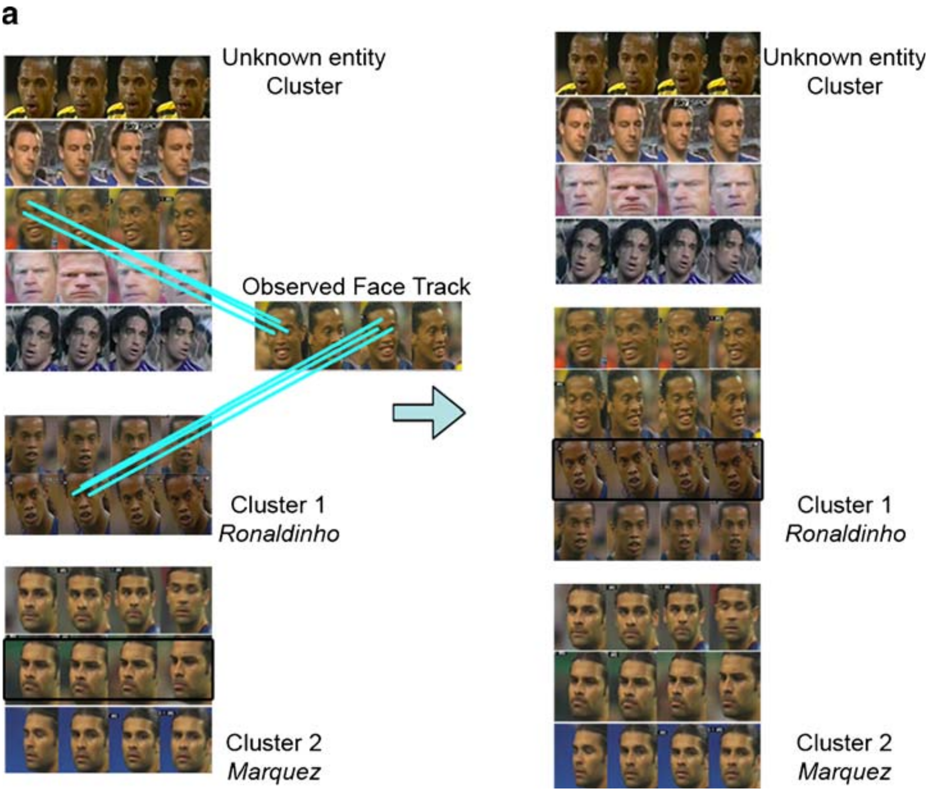

Fig. 3 Two examples of instance clustering and cluster updating, respectively applied to **a** face image sequences, obtained from face detection and tracking, and **b** highlight shots, obtained from video segmentation. As a new face/highlight sequence is presented, it is associated with the nearest instance in the clusters (SIFT matching and highlight descriptor matching are used respectively for faces and highlights, as explained in Section 4) and clusters are updated accordingly. Cluster splitting is shown for highlight clusters as an effect of cluster updating

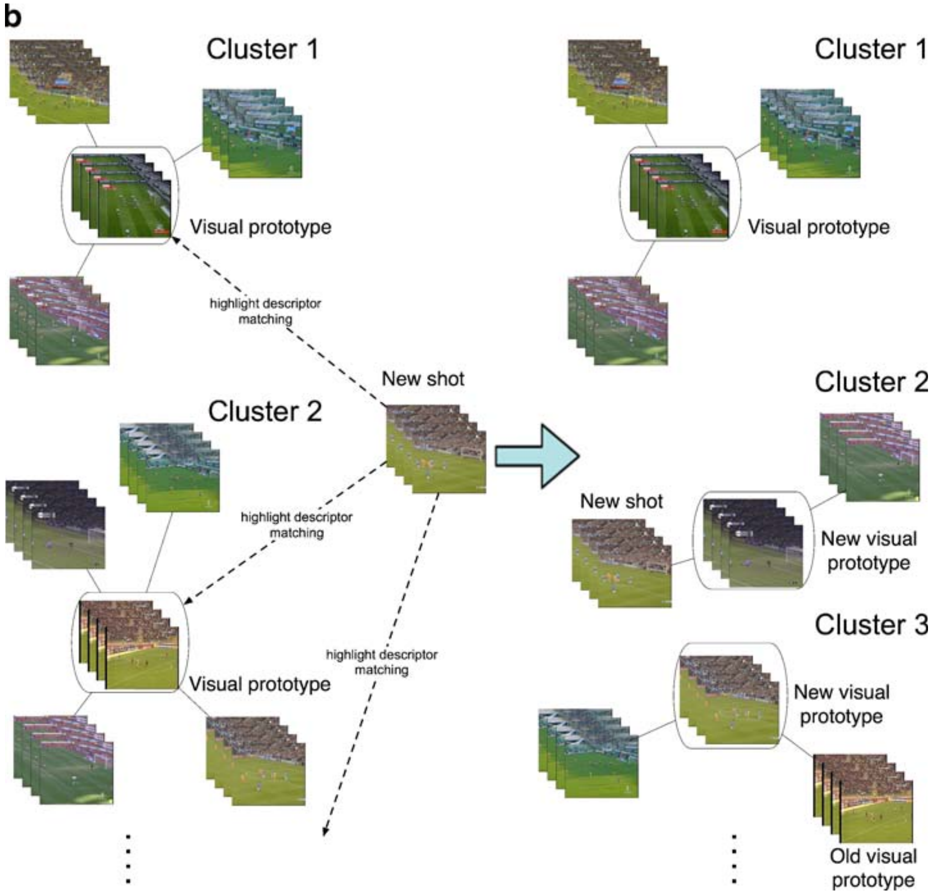

Fig. 3 (continued)

3.2 Annotation using reasoning

For some events, it is illusory to expect that visual descriptors in one shot are sufficient for a correct identification of the event. In many cases, instead, they can be asserted only after that the contents of the preceding and/or following shots have been analysed and classified properly. For example we will observe, in the experimental section, that shots that contain a part of a placed kick action can be classified as unknown action. But the classification of these shots as placed kicks can be inferred exploiting the fact that placed kick shots are often preceded by player-close-up and medium view shots (respectively focusing on the player at the ball and the other players), provided that these are correctly classified. Another example is given by the discovery of advertising billboards. Their detection is usually complicated by the fact that billboards appear with strong perspective distortion, occlusion and, most critical, at long distance from the camera. Reasoning about their proximity of the playground can help their recognition.

We can therefore define patterns of events and exploit spatial/temporal reasoning to improve shot classification. Since our ontology is defined in OWL, these patterns are formally expressed in SWRL rules. SWRL allows to use built-in mathematical, logical, string comparison and temporal operators; these permit an easy definition of rules that may include assertions about the presence of some concept, constraints on the values of the attributes of the concept instances, or their spatial/temporal occurrence.

4 Classifiers and visual descriptor matching

External classifiers and the descriptor matching procedure are applied to raw data to create instances of the concepts of the ontology lexicon, respectively for entities/events with little and large changes. In the following we discuss in details the mechanism used to distinguish between break/play actions; then visual descriptor matching for faces and advertising billboards is presented. Finally we describe the implementation of classifiers for text captions and jersey numbers, playground region, lines and zones.

4.1 Break and play action

Distinguishing between *break* and *play action* is essential for understanding much of the semantics of a soccer video. Break shots may include, among the others, player-close-up and crowd scenes. Play shots typically include play scenes with highlights. We represent the spatio-temporal pattern of the action in each shot as a vector V obtained as the composition of n vector U, as many as the feature descriptors used to describe the action (e.g. MPEG-7 color layout, edge histogram, camera motion, etc.). Each vector U is defined as the concatenation of the f_j feature vector ($j = 1, \ldots, m$, where m is the shot length) that has been extracted from the j^{th} frame of the shot. According to this definition, the length of the descriptor U and V may be different in different clips, depending on the duration of the shots [8].

Break and play action shots are distinguished considering the U components with the MPEG-7 color layout and edge histogram descriptors, and the face descriptors. In fact, break shots (*crowd*, *players-group*, *players-medium-view*, *player-close-up*) are characterized by small playground regions and large regions with strong edges.

The color layout descriptor includes the low-frequency DCT coefficients DY, DCr and DCb of the 8×8 blocks of each frame, in the YCrCb color space. The following weighted distance measure is used for matching [23]:

$$D = \sqrt{\sum_i w_{yi} \left(DY_i - DY'_i\right)^2} + \sqrt{\sum_i w_{bi} \left(DCb_i - DCb'_i\right)^2} + \\ + \sqrt{\sum_i w_{ri} \left(DCr_i - DCr'_i\right)^2}$$

The edge histogram descriptor provides the spatial distribution of five types of edges (four directional and one non-directional) obtained from the computation of edge histograms for 4 × 4 blocks of each frame and quantization of bin values to 3 bits. The L_1 norm is used as distance measure for matching edge histogram descriptors. The face descriptor reports the number of faces detected in the frame and their size.

For play actions the U component includes the main camera motion direction, intensity and acceleration (that approximately model the motion of the ball), the playground zone, and the number of players in the upper and lower part of the playground. The visual features are obtained from MPEG-2 motion vectors and uncompressed video, represented with quantized values (eight values for direction, three for speed, three for players, two for acceleration, twelve for playground zone), and smoothed to eliminate possible outliers. Play actions are distinguished in *kick-off*, *placed-kick*, *foul*, *forward-launch*, *attack-action*, *throw-in* and *shot-on goal*.

To perform matching of the shot descriptors V_s with the V_p descriptors of the visual prototypes, we use a distance function that is a modification of the string edit distance, defined as the sum of all the normalized Needleman-Wunch (NW) distances between the distinct U components:

$$d\left(V_s, V_p\right) = \frac{\sum_U NW\left(U_{V_s}, U_{V_p}\right)}{min\left(length\left(V_s\right), length\left(V_p\right)\right)}$$

This allows to account for the fact that video shots have different length and might include video editing operations such as frame trimming, insertion, and filtering. The algorithm is $O(mn)$ in time and $O(min(m, n))$ in space, where m and n are the lengths of the two strings being compared.

4.2 Face detection and recognition

Face detection is applied to *close-up* shots. To detect faces reliably we use the Viola & Jones detector, with the Adaboost classifier trained with frontal and quasi-frontal faces. We also apply face verification by computing the color autocorrelogram in the bounding box of the hypothesized face [9].

The color autocorrelogram of a face F is calculated as:

$$\alpha_c^{(k)}(F) = Pr_{p_1, p_2 \in I_c}\left[|p_1 - p_2| = k\right]$$

where $k = 35$ is the distance between two pixels p_1 and p_2 and c is one of 64 uniformly quantized colors in the RGB color space. A K-Nearest Neighbor classifier is used to separate good and false detections. Since classification may be unreliable in high-dimensional spaces, we project the correlogram vectors onto a linear subspace obtained from principal component analysis and retain the six principal components. As a face is detected, we compute three 128-length SIFT descriptors centered on the two eyes (a 20 × 20 pixel region, for face width normalized to 80 pixels), and the midpoint between eyes (a 15 × 30 pixel region). These regions provide in fact the most robust SIFT features.

To have a more robust matching of the player face, we use face sequences instead of single face images [7]. Once a face is detected in a frame, it is tracked so as to collect a sequence of face images of the same individual, with different expressions and poses. For tracking, we use particle–filtering applied to face color correlograms with a first-order dynamic model for the tracker state (i.e. we suppose a constant image velocity) [4]. In the ontology, there is one face cluster per player. Face sequences where the players' name is extracted from the text captions or jersey numbers are immediately associated with the appropriate cluster. In the other cases, the association is performed by SIFT matching between the faces in the sequence and the faces in the player face clusters and in the unknown cluster. If U is the face track of the observed player, and L is a face cluster in the ontology, matching is performed according to the minimum distance between the two sets as: $d(U, L) = min_{i,j}\|U_i - L_j\|$, where U_i and L_j are two 384-length vectors and $d(U, L)$ is the L_1 norm. If both faces in a player face cluster and the unknown cluster are matched, the face sequence matched in the unknown cluster is re-assigned to the player face cluster. An example of face clustering and updating is shown in Fig. 3a.

4.3 Advertising trademark billboard

Trademark billboards are observed in *crowd* and *play action* scenes. Differently from faces, trademarks in advertising billboards have not a fixed definite structure. To obtain a matching that is robust to partial occlusions, perspective distortions and scaling, we use keypoints (Difference-of-Gaussians local extrema) as interest points and SIFT descriptors. Detection of trademarks is done by comparing the bag of SIFT feature points of each reference trademark T_j in the ontology with the bag of SIFT descriptors in the V_i video frame [3].

If T_j is represented as a bag of the n_j SIFT feature points: $T_j = \{(x^t_k, y^t_k, s^t_k, d^t_k, \mathbf{O}^t_k)\}$, for $k \in \{1, \ldots, n_j\}$, where x^t_k, y^t_k, s^t_k, and d^t_k are, respectively, the x- and y-position, the scale, and the dominant direction of the kth feature point, and O^t_k is its 128-dimensional local orientation histogram, and each frame V_i of the video shot is similarly represented as a bag of the m_i SIFT feature points detected in that frame, then matching is performed by searching, for each point in T_j its two nearest neighbors in the V_i set:

$$N_1\left(T^k_j, V_i\right) = \min_q ||\mathbf{O}^v_q - \mathbf{O}^t_k||$$

$$N_2\left(T^k_j, V_i\right) = \min_{q \neq N_1(T^k_j, V_i)} ||\mathbf{O}^v_q - \mathbf{O}^t_k||$$

For every point in the frame, we compute a *match score* defined as:

$$M\left(T^k_j, V_i\right) = \frac{N_1\left(T^k_j, V_i\right)}{N_2\left(T^k_j, V_i\right)},$$

and discard the matches whose ratio $M(T^k_j, V_i)$ is greater than a threshold equal to 0.8. The final decision on whether frame V_i contains trademark T_j is made by thresholding the *normalized match score*: $|M^j_i|/|T_j| > \tau$.

The exact localization of the trademark is made with a robust estimation of the point cloud, by iteratively solving for the centroid coordinates (μ_x, μ_y) :

$$\sum_{i=1}^{n} \psi(x_i; \mu_x) = 0, \quad \sum_{i=1}^{n} \psi(y_i; \mu_y) = 0$$

where the influence function ψ used is the Tukey biweight and the scale parameter c is estimated using the median absolute deviation from the median. Figure 4 shows an example of trademark billboard matching and localization in a video frame.

4.4 Text captions and jersey numbers

Text captions and jersey numbers are detected in *close-up* shots. They are useful as they permit to obtain information about the player's name, automatically. We implemented two detectors for text captions and jersey number [9].

To detect text captions, accounting for their very small font size, we use image corners as local salient points. For each corner, we analyze the surrounding area in order to filter out the outliers caused by background objects. Unsupervised clustering is applied to the corners, to identify the bounding box of the cluster. *Maximally Stable Extremal Regions* (MSER) are hence computed to obtain a more precise identification of the regions that contain text captions. We implemented an improved version of the *union find* algorithm [36] to find the connected pixels of the image during the computation of MSERs. Pixels at the intersection of the corner neighborhood with the MSERs are subjected to further processing that includes performing K-fill filtering (to eliminate salt and pepper noise) and color uniformity checking (to eliminate small blobs). OCR is hence applied. The text string obtained from OCR is matched with the names of the players registered in the ontology by approximate string matching.

Official rules of FIFA soccer organization state that jerseys must be decorated with numbers from 1 to 22 on their front, and that their size is within a fixed range. To recognize jersey numbers, we used twenty two Viola & Jones detectors, each one trained for one single number. We found this approach by far more reliable than having classifiers for digits from 0 to 9, because two-digit numbers are not always

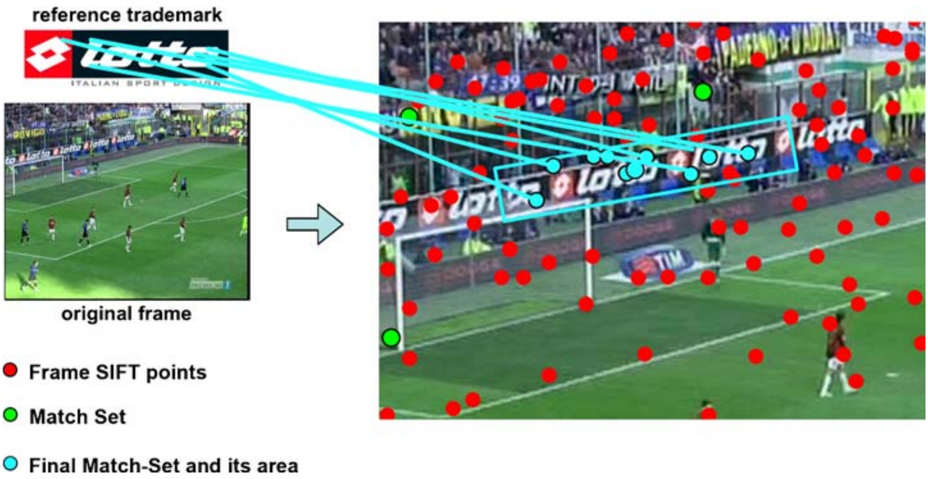

Fig. 4 Advertising trademark matching and localization: original trademark and a video frame with advertising billboards and SIFT point matches into evidence. *Light colored points* correspond to correct SIFT matching while *dark points* are wrong SIFT matches discarded. The area where the trademark has been localized has been included into a bounding rectangle

well separated, and missed detections are therefore highly probable. Each classifier was trained with 50 manually cropped positive and 100 negative examples, randomly selected from images.

4.5 Playground elements

Playground elements are searched in *play action* shots. We implemented distinct classifiers for the basic playground elements, namely playground region, playground lines and playground zones [2].

In a frame, the playground region is detected from color histogram intersection and color information. The region shape is refined with K-fill, flood fill, erosion and dilation, and approximated with a polygonal shape. Template matching is used to classify the playground region into one out of six different shapes corresponding to the most common views framed by the main camera (e.g. the corner views and the central playground views). A special class is used to collect all the unrelevant region shapes.

The playground lines are detected in the playground region, by exploiting edge and color information. Stick growing and merging of close and collinear lines is used to improve the quality of classification. For the midfield line, since it is observed from the main camera, we distinguish three different positions, depending on its detection in the left, central or right part of the frame. For the other playground lines, we distinguish between four different orientations: $[0° - 10°) \cup [170° - 180°)$, $[10° - 60°)$, $[60° - 120°)$, $[120° - 170°)$.

Playground is partitioned in twelve slightly overlapping zones, defined in a way that moving from one zone to the adjacent one indicates a new phase in the play. Each zone is recognized using a dedicated Naïve Bayes classifier, with the playground shape and the orientation of the playground lines as inputs. Since the classifiers have identical structure, large differences in their outputs are likely to be significant. According to this, the classifier with the highest output probability indicates which playground zone is framed.

5 Classification by spatial/temporal reasoning

Spatial/temporal reasoning expressed by SWRL rules is used to refine classification or to derive richer semantic annotation. In particular we used temporal reasoning in play actions. In fact, shots that contain play actions are often classified as *unknown* entities (as we report in the experimental results, Section 6). Moreover we observe that typically these play actions are preceded or followed by break actions; e.g. *placed-kick* shots are often preceded by *player-close-up* and *player-medium-view* shots that show the player at the ball and other players. Therefore, we have defined a few sample patterns for these events so as to verify the improvement in their classification rate, re-classifying the *unknown* entities. The definitions of such patterns are as follows:

– *given a shot classified as unknown action, this is re-classified as shot-on-goal IF two player-close-up shots AND a view to the goal post occur after the shot, with a few seconds of goal post view, within a time interval between 10 and 20 seconds.*

- *given a shot classified as unknown action, this is re-classified as placed-kick IF camera motion is zero for a short time interval at the beginning of the shot, AND the shot is preceded by two player-close-up OR one players-medium-view shots at least, for a total duration of about 50 seconds.*
- *given a shot classified as unknown action, this is re-classified as foul IF camera motion is zero for a short time interval at the end of the shot, AND the shot is followed by two player-close-up OR one players-medium-view shots at least, for a total duration of about 50 seconds.*

Patterns can be also used to derive semantic annotations for parts of video clips from the direct observation of the type and duration of basic concepts. The pattern of a scored-goal can be defined as follows:

- *IF a shot-on-goal is followed by crowd AND at least two player-close-up, AND all this is followed by a score-change event AND the time interval from shot-on-goal to score-change is between 40 and 80 seconds THEN a scored-goal event is asserted for the interval from the shot-on-goal to score-change (included).*

Figure 5 shows an example of a *scored-goal* subsequence within a longer sequence. The SWRL rule for the *scored-goal* pattern is given in Rule 1.

Reasoning can also be performed considering the co-occurrence of concepts or instance values, or spatial constraints between entities. In particular reasoning on co-occurrence of instance values was used to annotate player faces with player names. Due to the fact that players are framed in different orientations, illumination conditions and with different face expressions, even face images of the same player can be very different from each other. Therefore, the *unknown face* cluster is usually densely populated with many face sequences.

We used instead spatial reasoning to improve the discovery of advertising billboards in soccer video and annotate the video with the time interval in which a trademark is displayed. Discovering of billboard trademarks in soccer video by SIFT matching is complicated by the fact that billboards, as said before, appear with strong perspective distortions, occlusion and, most critical, at long distance from the camera. According to this, setting the appropriate threshold for the normalized match score is

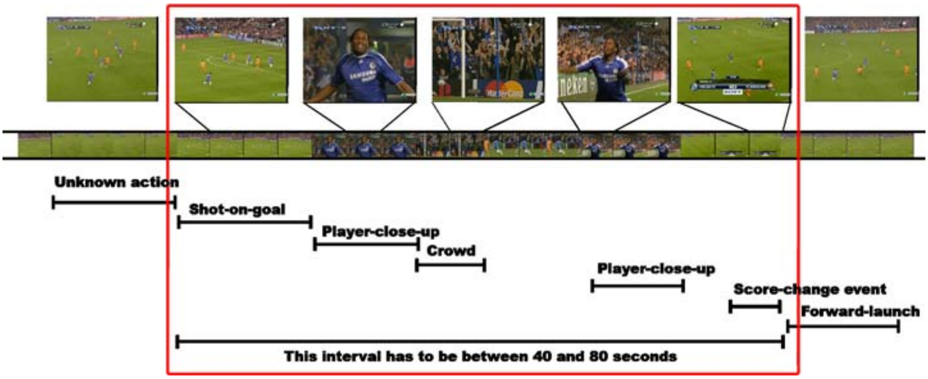

Fig. 5 Example of scored-goal sequence extracted using the scored-goal pattern

Rule 1 Scored-goal pattern in SWRL

Shot_on_Goal(?g) ^	instance of shot-on-goal
has_Valid_Period(?g, ?Vpg) ^	time interval of the shot-on-goal
temporal:hasStartTime (?Vpg, ?Stg) ^	start time of the shot-on-goal
Crowd(?c) ^	instance of crowd
has_Valid_Period(?c, ?Vpc) ^	time interval of crowd
Player_Close_up(?p1) ^	instance of player-close-up *p1*
Player_Close_up(?p2) ^	instance of player-close-up *p2*
differentFrom(?p1, ?p2) ^	the two player-close-ups have to be different
has_Valid_Period(?p1, ?Vpp1) ^	time interval of player-close-up *p1*
has_Valid_Period(?p2, ?Vpp2) ^	time interval of player-close-up *p2*
Score_Change_Event(?s) ^	instance of score-change event
has_Valid_Period(?s, ?Vps) ^	time interval of score-change event
temporal:hasFinishTime (?Vps, ?Fts) ^	finish time of the score-change event
temporal:before(?Vpg, ?Vpc) ^	time interval of shot-on-goal has to be before the time interval of crowd
temporal:before(?Vpg, ?Vpp1) ^	time interval of shot-on-goal has to be before the time interval of player-close-up *p1*
temporal:before(?Vpg, ?Vpp2) ^	time interval of shot-on-goal has to be before the time interval of player-close-up *p2*
temporal:after(?Vps, ?Vpc) ^	time interval of score-change event has to be after the time interval of crowd
temporal:after(?Vps, ?Vpp1) ^	time interval of shot-on-goal has to be after the time interval of player-close-up *p1*
temporal:after(?Vps, ?Vpp2) ^	time interval of shot-on-goal has to be after the time interval of player-close-up *p2*
temporal:duration(?diff, ?Stg, ?Fts, temporal:Seconds) ^	duration between start time of shot-on-goal and finish time of score-change event
swrlb:greaterThan(?diff, 40) ^	the duration has to be greater than 40 seconds
swrlb:lessThan(?diff, 80) ^	the duration has to be less than 80 seconds
swrlx:createOWLThing(?x, ?g)	creation of an instance x
->	then
Scored_Goal(?x) ^	the instance x is scored-goal
temporal:hasStartTime (?x,?Stg) ^	start time of the scored-goal is the start time of shot-on-goal
temporal:hasFinishTime (?x,?Fts)	finish time of the scored-goal is the finish time of score-change event

challenging. Higher values of threshold provide higher precision (more SIFT points are requested to match, between the observed frame and the ontology prototypes) but lower recall (many misses are probable, because of the hard conditions in which billboards are observed). On the other hand, with lower thresholds recall is improved at the expense of precision (many falses can be observed if few SIFT points are requested to match). Reasoning about the proximity of the playground was used to keep thresholding sufficiently high, and have at the same time both misses and falses at reasonable rates.

6 Experimental results

We verified the performance of the annotation framework over a large test base of soccer video. In particular, we analysed 75 hours of video (50 national and international matches played in years 2001, 2005 and 2006 by a few top Italian and European teams, so as to reduce the variability of the patterns displayed). We extracted the following clips for play actions: 210 *shot-on-goal* (70 per year approx), 240 *placed-kick* (80 per year approx), 60 *foul*; and the following clips for break actions: 120 *close-up*, 40 *players-group*, 70 *crowd*, 50 *player medium-view*. Video data was acquired from DV tapes and DVDs, encoded in MPEG-2, at full PAL resolution (720 × 576) and frame rate (25 fps). Depending on their content, clips were of variable length, ranging from a maximum of $2'15''$ to a minimum of $3''$, and might include several shots. We discuss experiments and results in the following paragraphs. The computational cost in terms of processing time per frame (using an Intel Core 2 Duo, 2.2 GHz), yields an average of about 0.1 s/frame, mostly due to the extraction of MPEG-7 and playground elements; the cost of ontology reasoning is almost negligible.

6.1 Video annotation using visual prototypes

In the following we discuss results of semantic annotation of concepts with large changes in motion patterns, using visual prototypes. Measures have been taken with the ontology trained according a 4-fold cross-validation. In this way, the original dataset is partitioned in a training set of about 600 clips and a test set of about 200 clips.

To evaluate the performance of the proposed framework, we use the common *precision* and *recall* measures. In particular, because of our definition of the *unknown-entity* concept, unknown have been considered as misses when computing the recall figure.

6.1.1 Break action annotation

Results for the annotation of break events are reported in Table 2 for the break action clips of the test set.

The *unknown* column reports the percentage of shots classified as unknown-entity for break actions. Typically the number of instances in this class decreases as new instances are presented to the ontology and associated with some of the break-action concepts. We observe that *players-group* and *players-medium-view* have worse figures than *player-close-up* and *crowd*. *Crowd* and *player-close-up* are easily distinguished from each other, because of the edge histogram descriptor. The low value of recall of *medium-view* is mostly due to the fact that the shots used have a very small number of frontal faces, most of them of small size, so that a

Table 2 Performance figures of annotation for break actions

Break action	Precision	Recall	False	Miss	Unknown
Crowd	0.88	0.81	0.12	0.19	0
Players-group	0.69	0.68	0.31	0.23	0.10
Players-medium-view	0.73	0.60	0.27	0.26	0.14
Player-close-up	0.87	0.75	0.13	0.25	0

Table 3 Performance figures of annotation for play actions

Play action	Precision	Recall	False	Miss	Unknown
Shot-on-goal	0.75	0.70	0.25	0.09	0.21
Placed-kick	0.85	0.46	0.15	0.20	0.34
Foul	0.87	0.22	0.13	0.10	0.68

high number of face missed detections is measured. *Players-group* have the smallest precision because in many clips they have similar descriptors (edge histogram and face descriptors) to *crowd* clips with the faces of supporters into evidence.

6.1.2 Play action annotation

Results for annotation of play actions are reported in Table 3 for the play action clips of the test set.

The low recall figure of *placed-kick* is in direct relation with the changes in the way in which this highlight is filmed from 2001 to 2006. In fact, placed-kick clips have changed their temporal evolution (and therefore their descriptors) due to the fact that the initial preparation of the *placed-kick* has been replaced in the years by other break sequences. This also impacts false detection of *shot-on-goal* that become, in some cases, quite similar to the recent appearance of *placed-kick* clips. The large changes in the descriptor patterns of both *shot-on-goal* and *placed-kick* highlights also determine that the *unknown* entity cluster is densely populated. *Fouls* have very high values for the unknown figure because of the extremely large number of patterns in which they may appear. We will show in the following section how reasoning on *unknown* entity instances can be used to improve the classification.

6.2 Video annotation by spatial/temporal reasoning over concepts and concept instances

The use of the ontology-based framework, as described in Section 3, allows to exploit spatial/temporal reasoning to improve shot classification of concepts and derive new semantic annotations. Experimental results are discussed in the following. The SWRL language and Jess reasoning engine were used to define rules and to perform spatial/temporal reasoning, respectively.

6.2.1 Temporal reasoning

Results for the annotation of *shot-on-goal*, *placed-kick* and *foul* with the application of reasoning (VP + SWRL) in our test are reported in Table 4. We can observed

Table 4 Performance figures of annotation for play action, with visual prototypes and reasoning (VP + SWRL), and reasoning only (SWRL)

Play action	Precision	Recall	False	Miss	Unknown
Shot-on-goal (VP)	0.75	0.70	0.25	0.09	0.21
Shot-on-goal (VP + SWRL)	0.78	0.76	0.22	0.11	0.13
Placed-kick (VP)	0.85	0.46	0.15	0.20	0.34
Placed-kick (VP + SWRL)	0.90	0.70	0.10	0.21	0.09
Foul (VP)	0.87	0.22	0.13	0.10	0.68
Foul (VP + SWRL)	0.90	0.66	0.10	0.10	0.24
Scored goal (SWRL)	0.84	0.66	0.10	0.37	–

Table 5 Performance figures of annotation for play action, with visual prototypes and reasoning (VP + SWRL) and SVM classifiers (SVM)

Play action	Precision	Recall	False	Miss	Unknown
Shot-on-goal (VP + SWRL)	0.78	0.76	0.22	0.11	0.13
Shot-on-goal (SVM)	0.73	0.84	0.27	0.16	–
Placed-kick (VP + SWRL)	0.90	0.70	0.10	0.21	0.09
Placed-kick (SVM)	0.81	0.75	0.19	0.25	–
Foul (VP + SWRL)	0.90	0.66	0.10	0.10	0.24
Foul (SVM)	0.51	0.38	0.49	0.62	–

that the main improvement of the reasoning process is shown in terms recall figure especially of *placed-kick* and *fouls* highlights in which the *unknown* entities is densely populated. Less strict rules can of course improve the rates of precision, unknown and miss. Performance in the detection of the *scored-goal* event is also reported in the same table.

In Table 5 we compare the performance obtained through the use of visual prototypes and SWRL rules to the traditionally employed SVM classification, for each play actions considered. In order to have a fair comparison, SVM classifiers (with RBF kernel) have been trained over the same training set of the proposed framework. Video clips have been represented with the same vector descriptors used in the ontology, with a fixed number of samples (5) per clip so as to have feature vectors of the same length. The improvement of the performance of our ontology framework is essentially due to the fact that, differently from SVM, SWRL rules permit to include some contextual information that allows to disambiguate situations (classified as *unknown-entity*) that can not be classified using only the value of visual features.

6.2.2 Spatial reasoning

Spatial reasoning is exploited to refine the annotation of the face and trademark billboard entities. In order to check the capability of reclassification of unknown face sequences, we considered 38 annotated face sequences of five distinct players (each player has between four to ten sequences) and 40 non annotated face sequences (twelve with jersey numbers and six with superimposed text captions). The test was therefore performed over the 78 face sequences, according to a leave-one-out scheme. Table 6 reports the performance observed with SIFT matching (VP) and with reasoning on jersey numbers and superimposed captions information (VP + SWRL).

The improvement of advertising billboards annotation by spatial reasoning was tested with videos of two soccer games (180′). In particular we use four trademarks

Table 6 Performance figures of annotation for face, with SIFT matching to visual prototypes (VP) and reasoning (VP + SWRL)

Face	Precision	Recall	False	Miss
Face only (VP)	0.70	0.46	0.30	0.54
Face + jersey number (VP + SWRL)	0.75	0.64	0.25	0.36
Face + caption (VP + SWRL)	0.76	0.61	0.24	0.39
Face + jersey + caption (VP + SWRL)	0.79	0.79	0.21	0.21

Table 7 Performance figures of trademark billboards, with SIFT matching to visual prototypes (VP) and reasoning (VP + SWRL)

Trademark billboard (threshold τ)	Precision	Recall	False	Miss
0.07 (VP)	0.22	0.78	0.78	0.22
0.10 (VP)	0.68	0.56	0.32	0.44
0.13 (VP)	0.82	0.34	0.18	0.66
0.13 (VP + SWRL)	0.82	0.52	0.18	0.48

where each of them was represented using three visual prototypes. Table 7 shows the average performance observed with SIFT matching at different threshold values (VP) and when applying spatial reasoning with the highest threshold (VP + SWRL). It can be noticed that a good improvement of recall, still keeping the same precision, is obtained with reasoning.

7 Conclusions

In this paper we have presented a complete framework for semantic annotation of soccer video clips based on Dynamic Pictorially Enriched Ontologies, where visual instances are included in the ontology as visual prototypes of clusters of similar visual or motion patterns. In this way, we can represent effectively concepts that present large changes in their appearance or motion pattern within one shot, and can update their representation whenever new observations are presented to the ontology for annotation, thus providing some form of evolution of the ontology knowledge. Since the ontology has been defined in OWL, SWRL has been used to reason over both concepts and concept instance values to improve the performance of automatic semantic annotation. The experimental results in the soccer video domain have shown evidence of the capability of this ontology model and framework to support effectively semantic annotation of entities or events with large changes in their appearance or motion patterns. Future work will address automatic learning of SWRL rules to improve the reasoning capabilities of the ontology.

Acknowledgements The authors are grateful to Sport System Europe s.r.l., Bologna, Italy, for having provided the large set of sports video used for the experimental validation of this research. This work was partially supported by the Information Society Technologies (IST) Program of the European Commission DELOS Network of Excellence on Digital Libraries (Contract G038-507618) and by the VidiVideo Project (Contract FP6-045547).

References

1. Assfalg J, Bertini M, Del Bimbo A, Nunziati W, Pala P (2002) Soccer highlights detection and recognition using HMMs. In: Proc of IEEE int'l conference on multimedia & expo (ICME)
2. Assfalg J, Bertini M, Colombo C, Del Bimbo A, Nunziati W (2003) Semantic annotation of soccer videos: automatic highlights identification. Comput Vis Image Underst 92(2–3):285–305
3. Bagdanov AD, Ballan L, Bertini M, Del Bimbo A (2007) Trademark matching and retrieval in sports video databases. In: Proc of ACM int'l workshop on multimedia information retrieval (MIR), Augsburg
4. Bagdanov AD, Del Bimbo A, Dini F, Nunziati W (2007) Improving the robustness of particle filter-based visual trackers using online parameter adaptation. In: Proc of IEEE int'l conference on AVSS, London, pp 218–223

5. Bai L, Lao S, Jones G, Smeaton AF (2007) Video semantic content analysis based on ontology. In: Proc of int'l machine vision and image processing conference, Maynooth, pp 117–124
6. Bai L, Lao S, Zhang W, Jones G, Smeaton A (2007) A semantic event detection approach for soccer video based on perception concepts and finite state machines. In: Proc intl'l workshop on image analysis for multimedia interactive services (WIAMIS)
7. Ballan L, Bertini M, Del Bimbo A, Nunziati W (2007) Soccer players identification based on visual local features. In: Proc of ACM int'l conference on image and video retrieval (CIVR), Amsterdam
8. Bertini M, Cucchiara R, Del Bimbo A, Torniai C (2005) Video annotation with pictorially enriched ontologies. In: Proc of IEEE int'l conference on multimedia & expo (ICME), Amsterdam, pp 1428–1431
9. Bertini M, Del Bimbo A, Nunziati W (2006) Automatic detection of player's identity in soccer videos using faces and text cues. In: Proc of ACM multimedia, Santa Barbara, pp 663–666
10. Bloehdorn S, Simou N, Tzouvaras V, Petridis K, Handschuh S, Avrithis Y, Kompatsiaris I, Staab S, Strintzis MG (2004) Knowledge representation for semantic multimedia content analysis and reasoning. In: Proc of EWIMT, London
11. Buitelaar P, Cimiano P, Racioppa S (2006) Ontology-based information extraction with soba. In: Proc of international conference on language resources and evaluation
12. Dasiopoulou S, Mezaris V, Kompatsiaris I, Papastathis VK, Strintzis MG (2005) Knowledge-assisted semantic video object detection. IEEE Trans Circuits Syst Video Technol 15(10): 1210–1224
13. Dublin Core Metadata Initiative (2009) Dublin Core Metadata Initiative homepage. http://dublincore.org/
14. Ekin A, Tekalp AM, Mehrotra R (2003) Automatic soccer video analysis and summarization. IEEE Trans Image Process 12(7):796–807
15. Espinosa S, Kaya A, Melzer S, Moller R, Wessel M (2007) Towards a media interpretation framework for the semantic web. In: Proc ICWI, pp 374–380
16. Fellbaum C (ed) (1998) Wordnet. An electronic lexical database. MIT, Cambridge
17. FIFA (2006) 2006 FIFA world cup broadcast wider, longer and farther than ever before. http://www.fifa.com/aboutfifa/marketingtv/news/newsid=111247.html
18. Francois A, Nevatia R, Hobbs J, Bolles R, Smith J (2005) VERL: an ontology framework for representing and annotating video events. IEEE Multimed 12(4):76–86
19. Grana C, Cucchiara R (2007) Linear transition detection as a unified shot detection approach. IEEE Trans Circuits Syst Video Technol 17(4):483–489
20. Haubold A, Naphade M (2007) Classification of video events using 4-dimensional time-compressed motion features. In: Proc of ACM int'l conference on image and video retrieval (CIVR), Amsterdam, pp 178–185
21. Hauptmann A, Chen M-Y, Christel M, Lin WH, Yang J (2007) A hybrid approach to improving semantic extraction of news video. In: Proc of IEEE int'l conference on semantic computing (ICSC), Irvine, pp 79–86
22. Huang CL, Shih HC, Chao CY (2006) Semantic analysis of soccer video using dynamic Bayesian network. IEEE Trans Multimed 8(4):749–760
23. Kasutani E, Yamada A (2001) The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In: Proc IEEE int'l conference on image processing (ICIP), Thessaloniki
24. Kokaram A, Rea N, Dahyot R, Tekalp AM, Bouthemy P, Gros P, Sezan I (2006) Browsing sports video: trends in sports-related indexing and retrieval work. IEEE Signal Process Mag 23(2):47–58
25. Lenat D, Guha R (1990) Building large knowledge-based systems: representation and inference in the cyc project. Addison-Wesley, Reading
26. Leonardi R, Migliorati P (2002) Semantic indexing of multimedia documents. IEEE Multimed 9(2):44–51
27. Leslie L, Chua T, Ramesh J (2007) Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In: Proc of ACM multimedia, Augsburg, pp 443–452
28. Liu J, Tong X, Li W, Wang T, Zhang Y, Wang H (2009) Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recogn Lett 30:103–113

29. Luo M, Ma YF, Zhang HJ (2003) Pyramidwise structuring for soccer highlight extraction. In: Proc of ICICS-PCM
30. Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A, Schneider L (2002) The wonderweb library of foundational ontologies. Tech rep, WonderWeb Deliverable D17. http://www.loa-cnr.it/DOLCE.html
31. Mei T, Hua XS (2008) Structure and event mining in sports video with efficient mosaic. Multimedia Tools and Applications 40:89–110
32. Naphade M, Smith J, Tesic J, Chang SF, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91
33. Neumann B, Moeller R (2006) On scene interpretation with description logics. In: Cognitive vision systems: sampling the spectrum of approaches, LNCS. Springer, New York, pp 247–278
34. Qasemizadeh B, Haghi H, Kangavari M (2006) A framework for temporal content modeling of video data using an ontological infrastructure. In: Proc of semantics, knowledge and grid, Guilin
35. Sadlier D, O'Connor N (2005) Event detection in field sports video using audio-visual features and a support vector machine. IEEE Trans Circuits Syst Video Technol 15(10):1225–1233
36. Sedgewick R (1983) Algorithms. Addison Wesley, Reading
37. Shyu ML, Xie Z, Chen M, Chen SC (2008) Video semantic event/concept detection using a subspace-based multimedia data mining framework. IEEE Trans Multimedia 10(2):252–259
38. Simou N, Saathoff C, Dasiopoulou S, Spyrou E, Voisine N, Tzouvaras V, Kompatsiaris I, Avrithis Y, Staab S (2005) An ontology infrastructure for multimedia reasoning. In: Proc of VLBV, Italy, pp 51–60
39. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proc of ACM int'l workshop on multimedia information retrieval (MIR), Santa Barbara, pp 321–330
40. Snoek C, Worring M (2005) Multimedia event-based video indexing multimedia event-based video indexing using time intervals. IEEE Trans Multimed 7(4):638–647
41. Snoek C, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications 25(1):5–35
42. Snoek C, Huurnink B, Hollink L, de Rijke M, Schreiber G, Worring M (2007) Adding semantics to detectors for video retrieval. IEEE Trans Multimedia 9(5):975–986
43. Tsinaraki C, Polydoros P, Kazasis F, Christodoulakis S (2005) Ontology-based semantic indexing for MPEG-7 and TV-Anytime audiovisual content. Multimedia Tools and Applications (26): 299–325
44. Utsumi O, Miura K, Ide I, Sakai S, Tanaka H (2002) An object detection method for describing soccer games from video. In: Proc of IEEE int'l conference on multimedia & expo (ICME)
45. Watve A, Sural S (2008) Soccer video processing for the detection of advertisement billboards. Pattern Recogn Lett (29):994–1006
46. Wei X, Ngo CW (2007) Ontology-enriched semantic space for video search. In: Proc of ACM multimedia
47. Wu Y, Tseng B, Smith J (2004) Ontology-based multi-classification learning for video concept detection. In: Proc of IEEE int'l conference on multimedia & expo (ICME)
48. Xie L, Xu P, Chang SF, Divakaran A, Sun H (2004) Structure analysis of soccer video with domain knowledge and hidden Markov models. Pattern Recogn Lett 25(7):767–775
49. Xu C, Wang J, Lu H, Zhang Y (2008) A novel framework for semantic annotation and personalized retrieval of sports video. IEEE Trans Multimed 10(3):421–436
50. Xu P, Xie L, Chang SF, Divakaran A, Vetro A, Sun H (2001) Algorithms and system for segmentation and structure analysis in soccer video. In: Proc of IEEE int'l conference on multimedia & expo (ICME)
51. Yang Y, Lin S, Zhang Y, Tang S (2008) A statisticall framework for replay detection in soccer video. In: Proc of IEEE international symposium on circuits and systems
52. Ye Q, Huang Q, Jang S (2005) Jersey number detection in sports video for athlete identification. In: Proc of visual communications & image processing (VCIP)
53. Yu X, Farin D (2005) Current and emerging topics in sports video processing. In: Proc IEEE ICME
54. Zha ZJ, Mei T, Wang Z, Hua XS (2007) Building a comprehensive ontology to refine video concept detection. In: Proc of ACM int'l workshop on multimedia information retrieval (MIR), Augsburg, pp 227–236

Lamberto Ballan received the MS degree in computer engineering in 2006 from the University of Florence, Italy, where he is currently a PhD student at the Visual Information and Media Lab at Media Integration and Communication Center. His main research interests focus on Multimedia Information Retrieval, Computer Vision and related fields such as Pattern Recognition and Machine Learning.

Marco Bertini is Assistant Professor at the Department of Systems and Informatics at the University of Florence, Italy. He received a MS in electronic engineering from the University of Florence in 1999, and PhD in 2004 from the same University. His main research interest is content-based indexing and retrieval of videos and Semantic Web technologies.

Alberto Del Bimbo is Full Professor of Computer Engineering at the University of Florence, Italy. He is also the Director of the Master in Multimedia, and the President of the Foundation for Research and Innovation at the same university. His scientific interests are Pattern Recognition, Image Databases, Human Computer Interaction and Multimedia applications. Prof. Del Bimbo is the author of over 230 publications in the most distinguished international journals and conference proceedings. He is the Associate Editor of Pattern Recognition, Journal of Visual Languages and Computing, Multimedia Tools and Applications, Pattern Analysis and Applications, and International Journal of Image and Video Processing, and was the Associate Editor of IEEE Transactions on Multimedia, and IEEE Transactions on Pattern Analysis and Machine Intelligence.

Giuseppe Serra received the laurea degree in computer engineering from the University of Florence in 2006. He is a PhD student at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence. His research interests focus on Video Understanding based on Statistical Pattern Recognition and Ontologies, Multiple View Geometry, Self-calibration and 3D Reconstruction.