Sports Video Annotation Using Enhanced HSV Histograms in Multimedia Ontologies

M. Bertini, A. Del Bimbo, C. Torniai
Università di Firenze - Italy
Via S. Marta, 3 - 50139 Firenze
bertini,delbimbo,torniai@dsi.unifi.it

C. Grana, R. Vezzani, R. Cucchiara
Università di Modena e Reggio Emilia - Italy
Via Vignolese, 905/b - 41100 Modena
cgrana,rvezzani,cucchiara@unimore.it

Abstract

This paper presents multimedia ontologies, where multimedia data and traditional textual ontologies are merged. A solution for their implementation for the soccer video domain and a method to perform automatic soccer video annotation using these extended ontologies is shown. HSV is a widely adopted space in image and video retrieval, but its quantization for histogram generation can create misleading errors in classification of achromatic and low saturated colors. In this paper we propose an Enhanced HSV Histogram with achromatic point detection based on a single Hue and Saturation parameter that can correct this limitation. The more general concepts of the sport domain (e.g. play/break, crowd, etc.) are put in correspondence with the more general visual features of the video like color and texture, while the more specific concepts of the soccer domain (e.g. highlights such as attack actions) are put in correspondence with domain specific visual feature like the soccer playfield and the players. Experimental results for annotation of soccer videos using generic concepts are presented.

1. Introduction

The relevance of audio-visual data and more in general of media other from text in modern digital libraries has grown in the last few years. Modern digital libraries are expected to include digital content of heterogeneous media, with the perspective that visual data will have the major role in many specialized contexts very soon.

In this scenario, digital video has probably the highest relevance. There are, in fact, huge amounts of digital video daily produced for news, sport, entertainment, and education by broadcasters, media companies and various institutions. In particular the sport domain is receiving a growing attention due to the interests of broadcasters, sponsors and the audience.

A growing need is the ability to search for videos based on their content instead of relying on manually provided metadata. The diffusion of such systems has been strongly limited by the difficulty to generalize results of visual and audio automated processing techniques obtained on tuned test data sets. On the other hand, internet users are very demanding on search results, so the media search technologies for the mass have to be very intuitive and easy to use such as text retrieval is [10].

Organization of concepts into appropriate knowledge structures, such as ontologies, development of tools for the construction of such knowledge, and the usage of these structures and tools to support effective access to video content, are also the subject of recent and very active research. Moreover in recent years ontologies have been effectively used to perform semantic annotation and retrieval of multimedia content. In the case of video annotation the terms of the ontologies are associated to the individual elements of the video (e.g. blobs, shots, etc.) either manually or automatically, exploiting the results of the advancements in pattern recognition and image/video analysis. In this latter case spatio-temporal combinations of multimedia features are extracted from the media and linked to the terms of the ontology.

Although linguistic terms, commonly used to define concepts in ontologies, are appropriate to distinguish event and object categories, they are inadequate when they must describe specific patterns of events or video entities. Consider for example the many different ways in which an attack action can occur in soccer. We can easily distinguish several different patterns that differ each other by the playfield zone, the number of players involved, the player's motion direction, the speed, etc. Each of these patterns represents a specific type of attack action that could be expressed in linguistic terms only with a complex sentence, explaining the way in which the event has developed. To support effective retrieval of video data, capturing the different patterns of a concept, we need extended ontologies, that allow to inte-

grate the low-level expressive power of visual data with the structured high level semantic knowledge expressed in textual form. The basic idea behind these extended ontologies (that we will refer to as "multimedia ontologies") is that the concepts and categories defined in a traditional ontology are not rich enough to fully describe or distinguish the diversity of the possible visual events and cannot support video annotation up to the level of detail of pattern specification.

2. Related work

The possibility of extending linguistic ontologies with multimedia ontologies, has been suggested in [12] to support video understanding. Differently from our approach, the authors suggest to use *modal keywords*, i.e. keywords that represent perceptual concepts in several categories, such as visual, aural, etc. A method is presented to automatically classify keywords from speech recognition, queries or related text into these categories. In [14], a hierarchy of ontologies has been defined for the representation of the results of video segmentation. Concepts are expressed in keywords and are mapped in an *object ontology*, a *shot ontology* and a *semantic ontology*. In [21] the limitations in describing the semantics of the highly structured domain of sports videos using MPEG-7 have been overcome proposing a Sport Event ontology, that reuses only the structural and media information concepts of MPEG-7. A Protégé plug-in that allows manual annotation of soccer videos has been presented. Multimedia ontologies are constructed manually in [11]: text information available in videos and visual features are extracted and manually assigned to concepts, properties, or relationships in the ontology. In [2] new methods for extracting semantic knowledge from annotated images is presented. Perceptual knowledge is discovered grouping images into clusters based on their visual and text features and semantic knowledge is extracted by disambiguating the senses of words in annotations using WordNet and image clusters. In [18] a Visual Descriptors Ontology and a Multimedia Structure Ontology, based on MPEG-7 Visual Descriptors and MPEG-7 MDS respectively, are used together with domain ontology in order to support content annotation. Visual prototypes instances are manually linked to the domain ontology. An approach to semantic video object detection is presented in [5]. Semantic concepts for a given domain are defined in an RDF(S) ontology together with qualitative attributes (e.g. color homogeneity), low-level features (e.g. model components distribution), object spatial relations and multimedia processing methods (e.g. color clustering); rules in F-logic are then used for detection on video objects. In [7] a taxonomy defined with MPEG-7 is used to annotate videos using unsupervised clustering. In the proposed system it is not possible to add domain specific visual descriptors, nor high level relationships between the concepts defined in the taxonomy. In [16] an ontology infrastructure composed by three ontologies (a domain specific, a visual descriptor and an upper ontology) has been proposed to interpret a scene. An initial automatic segmentation of images is performed, extracting a set of low-level visual descriptors for each region. A training set of manually labeled regions is used to label new regions, and a reasoning engine is used to reduce the number of segmented regions using the labels and the adjacency relations.

3. Ontology languages and visual descriptors

The aim of using ontologies to describe multimedia documents is to provide ways to define well structured concepts and their relations that may ease the tasks of annotation and retrieval. As noted in [9] ontologies make it possible to improve both automatic annotation and retrieval in presence of imperfect annotation; in fact, an automatic annotator that uses concepts instead of labels maintains consistency of annotations in two ways: *i)* avoiding unlikely combinations of annotations; *ii)* using generic concepts (instead of the more specific) in case of uncertainty. Retrieval is improved using the ontology to drive the query rewriting, e.g. to select more general concepts, or concepts that are somehow related to the query.

Ontologies may be expressed in many different languages. In general it is possible to create an ontology using a proprietary language, or using open standard languages such as XML, RDF(S) or OWL. MPEG-7 is a standard, based on XML, that has been built to define entities and their properties w.r.t. the specific domain of multimedia content while RDFS and OWL are languages that can define an ontology in terms of concepts and their relationships regardless of the domain of interest. It has to be noted that the possibility to translate MPEG-7 into an ontology language [17] such as RDF and OWL has been exploited to overcome the lack of formal semantics of the MPEG-7 standard that could extend the traditional text descriptions into machine understandable ones. There are several advantages in using ontologies specific standards as RDF and OWL instead of MPEG-7: *i)* it is easier to add mid-level audio-visual descriptors that are related to the domain that is being annotated; *ii)* it is possible to express easily complex semantic entities and their relations; *iii)* it is possible to use reasoning tools to add high-level annotation or to perform queries related to semantic content, rather than low level audio-visual descriptors. In our approach the extended ontology is expressed using OWL and both domain and visual concepts are included in a single ontology. Visual concepts are defined by mean of both generic and domain specific visual descriptors.

The generic descriptors used in the ontology capture dif-

ferent aspects of the video content: *i)* global color; *ii)* layout of colored areas and *iii)* texture. These features are represented using the corresponding MPEG-7 descriptors: scalable color descriptor (SCD), color layout descriptor (CLD) and edge histogram descriptor (EHD).

The domain specific descriptors are typically more computationally expensive, but their extraction can be postponed since they are needed to recognize concepts that may be specializations of the more generic concepts. For instance the generic descriptors are used to detect if a video clip contains a play action, and only when there is need to specify what kind of highlight is contained, domain specific descriptors are used. In the case of the soccer the domain the specific descriptors used are *i)* the playfield area; *ii)* the number of players in the upper part of the playfield; *iii)* the number of players in the lower part of the playfield; *iv)* the motion intensity; *v)* the motion direction; *vi)* the motion acceleration. Their extraction and use has been presented in [1] and [4].

3.1. HSV Color Space

While the importance of color feature is straightforward, the selection of the best color features and color space is still an open issue. In particular, the MPEG-7 Scalable Color Descriptor feature has been tested and proved to be misleading for achromatic and dark colors, since it is based on the HSV color space. To this aim, we propose in this paper an enhanced HSV color histogram able to explicitly handle achromatic colors.

The HSV color space is widely adopted in image and video retrieval, since it allows to separate the chromatic contribution of the image colors. Kotoulas *et al.* [13] for example use HSV color histograms for content-based image retrieval proposing hardware implementations of the algorithms. Despite that, the use of HSV color space has three well known drawbacks: (1) hue is meaningless when the intensity is very low; (2) hue is unstable when the saturation is very low; and (3) saturation is meaningless when the intensity is very low. The same consideration is reported in [20] but applied to the IHS color space, that is conceptually similar to the HSV. In other words, dark colors are insensitive to saturation and hue changes and, similarly, the hue value is negligible for achromatic colors (low saturation). Thus, the MPEG-7 Scalable Color Descriptor suffers of this defect: dark or low saturated colors can be assigned to different bins even if they are visually very similar (see Fig. 1).

To solve this problem, in [20] two regions are defined and separately treated, one for the chromatic and one for the achromatic colors. These areas are obtained with a complicated set of thresholds in the IHS color space. Similarly, in [15] a fuzzy technique has been proposed in order to

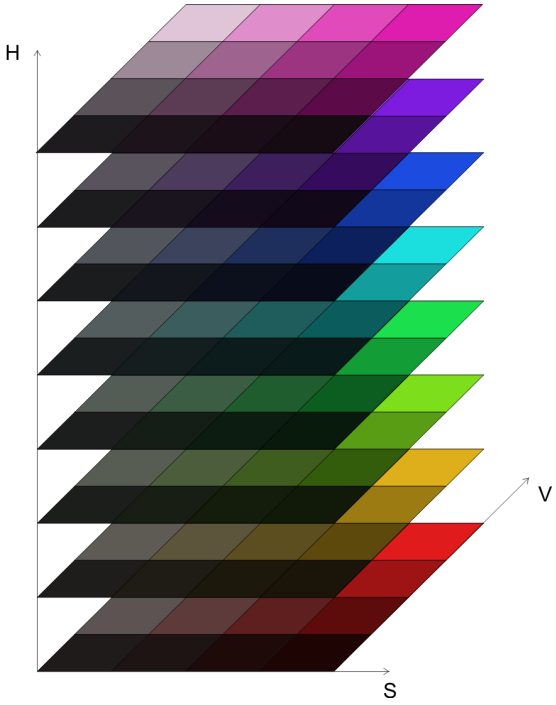

Figure 1. Quantized view of the HSV color space. S and V are quantized to 4 values, while H to 8 values. The bins with lowest V and S present barely noticeable visual differences.

distinguish among chromatic, achromatic, and black colors. Sural *et al.* [19] propose an histogram modification that takes into account the above mentioned regions. In particular, they identify the achromatic region by thresholding the saturation coordinate with a linear function of the intensity value and based on the outcome chose to represent the color with its hue or its value only. In [6] a detailed comparison of the MPEG-7 color descriptors can be found, proving that the Scalable Color Descriptor is not suitable for monochromatic images. In the following section we propose an enhancement to the traditional HSV histograms that takes into account the achromatic regions overcoming the MPEG-7 Scalable Color Descriptor drawbacks.

3.2. Enhanced HSV Histograms

The Scalable Color Descriptor requires a quantization of the HSV color space, with 16 values in H and 4 values in each S and V (256 bins in total). Supposing every color channel in the range [0,1), the bin index may be obtained as:

$$\begin{aligned} \text{bin} &= f(H,S,V) = \\ &\lfloor n_H H \rfloor \, n_S n_V + \lfloor n_S S \rfloor \, n_V + \lfloor n_V V \rfloor \end{aligned} \tag{1}$$

Figure 2. Example of the HSV problem. The image on the right, shows in different colors pixels of the original image assigned to different histogram bins. Dark pixels which are visually very similar, fall in different bins when the color space is quantized.

where n_H, n_S, n_V are the quantization levels devoted to every color channel. Usually these are chosen to be powers of 2 for ease of representation. Adopting a linear quantization of each coordinate leads to have, for example, 64 different bins for the darkest colors characterized by the lowest values of V. Thus, a visually uniform background can be split on different bins (see Fig. 2).

We propose to add n_A bins to the HSV histogram that contains all the achromatic and dark colors. These n_A bins correspond to gray levels, from black to white; for convenience, we choose to set $n_A = n_V =4$, as the number of levels assigned to the V axis in the MPEG-7 standard. The dark and achromatic colors are selected by imposing a unique threshold λ on the S and V coordinates respectively, as reported in Table 1. In the third column, the index computation for the new bins is reported. The value of λ has been empirically set to 0.2, since it proved to be a good tradeoff between color loss and matching of similar dark or achromatic colors. In Fig. 3 the obtained results over four sample images are reported. In the first column the input images are shown, while in the second column the image segmentation obtained with the quantization of the MPEG-7 Scalable Color Descriptor is drawn assigning a different random color to each bin. The background color of the first and of the third row images, the sea in the second row, and the hat in the last row are some examples of dark or achromatic uniform areas that are split into different bins. Chromatic areas are marked in red on the rightmost column images, leaving original colors in the dark or achromatic areas. In the third column, the results of the Enhanced HSV Histogram are reported, where the same random colors as before are used for the chromatic area of the histogram, while gray levels are employed for the achromatic area.

Moving some of the colors from the original bins to the n_A achromatic ones makes these original bins less used

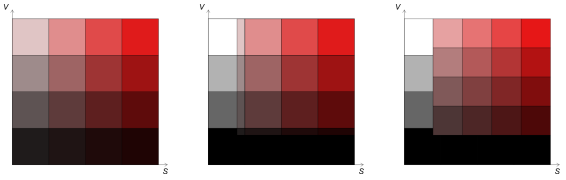

Figure 4. a) Original quantization of the SV plane with H=0 (red); b) Achromatic area detection (λ=0.2); c) linear requantization of the chromatic area.

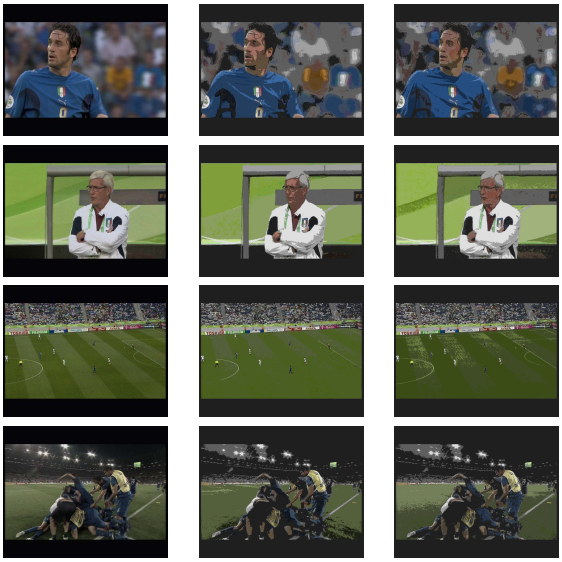

Figure 5. Reconstruction of the images in the left column, after the quantization with the functions f (center) and f' (right).

with respect to the others. In fact, it does not make sense anymore to uniformly subdivide the S and V channels, if part of it is then discarded. A better solution is to fully employ the chromatic bins to describe only the effective chromatic area. To this aim we linearly quantize the remaining HSV space, by simply redefining the f function of Eq. 1:

$$f'(H,S,V) = f\left(H, \frac{S-\lambda}{1-\lambda}, \frac{V-\lambda}{1-\lambda}\right) \qquad (2)$$

For the achromatic area the original f function is still used, since the whole range of V has to be quantized. A graphical visualization of the effect of Eq. 2 is given in Fig. 4. In Fig. 5 the reconstruction of the images after the quantization with the two different functions is reported. Clearly the second approach gives a better match with the original chromatic area.

Figure 3. Enhanced HSV Histogram results over four sample images are reported. The columns show respectively: input images, segmentation obtained with the HSV quantization, gray levels assignment for the achromatic area, chromatic area masking ($\lambda = 0.2$).

A different approach could be used to reduce the number of bins in the histogram, without affecting the chromatic area. The threshold λ can be set to $1/n_V$, thus making the achromatic area exactly match the first set of bins for S and V. This forces these bins to 0, thus allowing their removal. This indeed induces a compression with respect to the color representation, but it is selectively applied to the least significant colors. For example with reference to the aforementioned 16,4,4 subdivision, this would lead to $16 \times 3 \times 3 = 144$ bins, plus 4 bins for the achromatic area.

4. Multimedia ontology creation and usage

The creation process of the multimedia ontology is performed by selecting a representative set of sequences containing concepts described in the linguistic ontology, extracting the visual descriptors and performing an unsupervised clustering. The clustering process, based on visual features, generates clusters of sequences representing specific patterns of the same concept that are regarded as specialization of the concept. Visual prototypes for each concept specialization are automatically obtained as the centers of these clusters. These prototypes are used in the annotation process and can be exploited to perform queries (e.g. searching videos whose content is similar to some visual prototypes).

The first step of the multimedia ontology creation is to define for each clip a feature vector V containing all the distinct components of the visual descriptors used to define the concepts. Each component is a vector U that contains the sequence of values of each visual descriptor. The length of feature vectors U may be different in different clips, according to their duration and to the content of the clips. Then the clustering process groups the clips of the representative set according to their visual descriptors. We have employed the fuzzy c-means clustering algorithm to take into account the fact that a clip could belong to a cluster, still being similar to clips of different clusters. The distance between two different clips has been computed as the sum of all the normalized Needleman-Wunch distances between the U com-

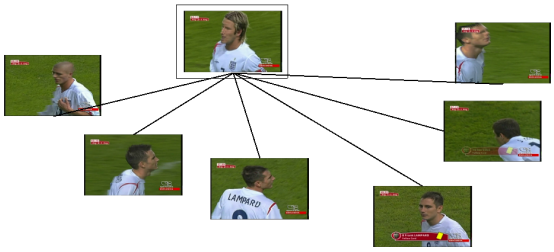

Figure 6. Example of a cluster for players close-up concept, created using generic visual descriptors. The visual prototype is highlighted.

ponents of the feature vector V of the clips, to take into account the differences in the duration and the temporal changes of the descriptors values. This allows to take into account the dynamic aspects of videos and the video editing operations typically used (e.g. trimming, insertion, filtering). The distance is a generalization of the Levenshtein edit distance and has been used since the cost of character substitutions is an arbitrary distance function. In our case the cost is used to weight differently some high level descriptors, while the generic descriptors use the same cost. The normalization is used in order to better discriminate differences between short and long sequences and is performed dividing the Needleman-Wunch distance by the length of the shorter sequence. The edit distance can be applied also to the MPEG-7 based generic descriptors since, even if the size of their "alphabet" is extremely large, in order to perform the calculation of the distance it is only required to determine if two descriptors are the same or not, using the appropriate metric described in the previous section. An example of cluster used to identify a visual prototype for the players close-up concept is shown in Fig. 6. Performance evaluation of the generation of multimedia ontology for high level concepts has been analyzed in our previous work [3].

In Fig. 7 it is shown a simplified schema of the events and actions that can be detected using a multimedia ontology defined for the soccer domain, along with some examples of the visual concepts. The more generic concepts, that are represented in the upper part of the figure, are related to the more generic visual descriptors. The concepts that are more domain specific, as the different soccer highlights, can be recognized using the more specific visual descriptors. Using the ontology it is possible to perform a refinement of concepts analyzing only the video clips which may contain the more specific highlights. For example the leftmost example of "Play detected action" of Fig.7 can be further analyzed and then associated with the "Shot on goal" concept, or a "Players close-up" concept can be refined performing face recognition, and then associating the video clips of the cluster to the corresponding players concepts.

Once created, the multimedia ontology can be effectively used to perform automatic annotation of unknown sequences. This is made by checking the similarity of the visual descriptors of clips that have to be annotated with the descriptors of the visual prototypes and of their cluster members. The annotation algorithm (described in detail in [4]) aims at obtaining a high precision at the expense of recall. This is due to the fact that while annotating the videos the ontology is growing at the same time, adding and changing the visual prototypes. Therefore, if a clip is erroneously associated to the wrong concept (in particular if it becomes a prototype of that concept) the error would propagate, attracting other clips to the wrong concept. This is the reason why we have chosen to associate clips to an "Unknown events" special class if there is some uncertainty in their annotation. These clips can be annotated at a later stage automatically, when the ontology contains more knowledge about the concepts that they contain. In fact the annotation algorithm tries to re-classify the "Unknown events" clips when a certain number of new visual prototypes are added to the ontology.

Once the videos have been classified and annotated within the multimedia ontology, using only the visual features, it is possible to refine the annotation exploiting the relations between the concepts, for example defining domain specific patterns of events and objects. Since the ontology is expressed in OWL, a reasoner can be used to add to the ontology the knowledge inferred from these relations. In a similar way it is possible to define complex queries based on visual prototypes and their relationships.

For instance, a temporal sequence of "Shot on goal" (obtained as specialization of the generic play concept, through analysis of domain specific features), followed by some "Players close-up" and "Crowd" can be defined as possible "Scored goal" event; a "Play" followed by a some "Players close-up" and "Players medium views" can be defined as possible "Foul" event. These concepts could not be detected solely based on visual features, but can be detected through reasoning on the ontology. For query purposes these temporal sequences can be refined requiring some concepts visually similar to some particularly interesting visual prototypes (e.g. a "Scored goal" that contains a "Shot on goal" similar to a given sequence).

5. Experimental results

The videos used in our experiments were digitally recorded and acquired at full PAL frame size and rate (720×576 pixels, 25 fps) from 3 different matches (one from UEFA Champions League and two from World Championship 2006), for a total length of about 270 minutes.

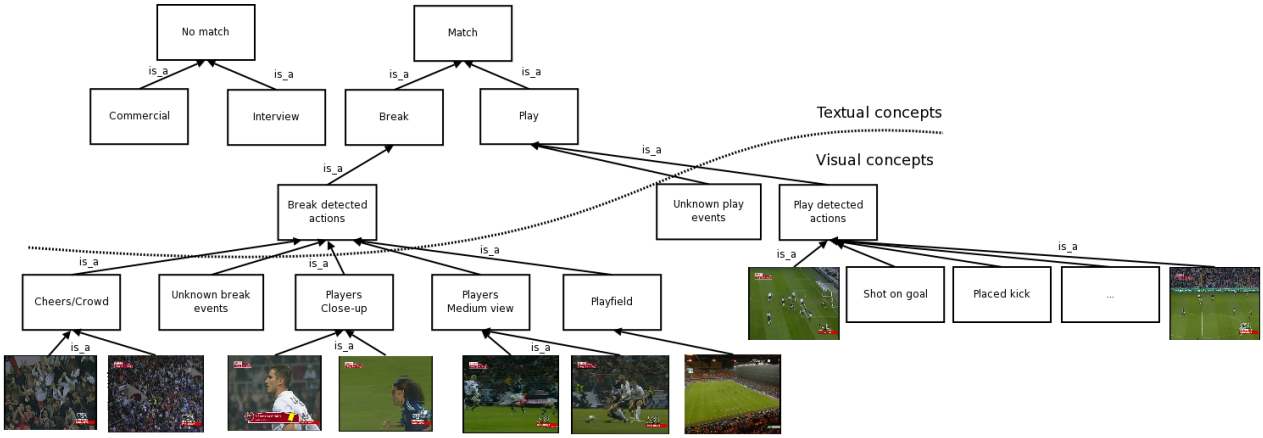

Figure 7. A simplified schema for the detected events and actions of the soccer ontology

Table 1. Precision and recall of generic soccer concepts classification

Concept	Miss	Unknown	False	Precision	Recall
Play	1%	0%	12%	90%	99%
Crowd	10%	0%	20%	82%	90%
Close-up	10%	43%	1%	98%	48%
Medium view	16%	18%	16%	79%	65%

Videos were automatically segmented using a Linear Transition Detector[8], a system based on color histograms and motion features to detect and characterize scene changes, and features vectors were extracted from each shot. Manual ground truth of generic soccer concepts of these shots has been created resulting in 711 play actions, 62 crowd actions, 905 players close-up and 492 players medium views. A set of sequences representative of the above concepts was selected as a training set for the multimedia ontology creation. In particular 50 different actions representative of the play concept, 10 crowd events, 150 players close-up and 80 players medium views, were manually annotated. The high number of close-up and medium views used to train and build the ontology was required to cope with the great diversity of the visual appearance of these concepts (e.g. the different players faces and team shirts). The ontology creation process performed using this training set led to the selection of 8 visual concepts for the play, 3 for the crowd, 48 for the players close-up and 20 for the players medium view. This ontology has been then used to automatically annotate the remaining video sequences. Results are reported in table 1.

In Table 1 we have reported the percentage of clips annotated in the special "Unknown events" class instead of reporting them within the "Miss" column since, as explained in Sect.4, these error can be automatically corrected at a later stage, when more clips are fed to the system. However these figures have been taken into account to compute the recall figure. Analysis of the results shows that close-up and medium view are more critical than play and crowd, that show good results for both precision and recall. The relatively low recall figure of close-ups is partly due to the fact that the majority of missed detections is wrongly classified as medium view; this is because a close-up sequence is often zoomed out becoming a medium view. Moreover, the high figure of close-up classified as "Unknown" is due to the high variability in visual appearance related to all the different players that are framed; however this problem may be solved adding more videos to the ontology, thus learning more prototypes of players close-ups. The figure of recall for the medium views is due to the fact that some of these sequences are framed in such a way that the players are completely surrounded by the playfield, thus resulting in a scene that is very similar to a play action scene. Also in this case the figure of medium views classified as "Unknown" is due to the high variability of this type of concept, and thus can be solved as in the close-up case.

6. Conclusions

In this paper we presented an approach for HSV color space analysis and enhanced histogram generation. The proposed enhancement allows to better describe the chromatic area and to avoid meaningless assignment by hue for low saturated and dark colors. In particular an achromatic area detection has been defined. We have shown the advantages of using multimedia ontologies to perform video annotation, and motivations to express them in a knowledge representation language such as OWL. Multimedia ontologies can include both generic concepts, represented by generic visual descriptors such as MPEG-7 descriptors, and

high-level concepts represented by domain specific descriptors. Experiments on automatic video annotation of soccer videos using the generic concepts have been presented. This approach allows to leverage the knowledge contained in the ontology to perform more refined annotation that otherwise would not be possible by means of visual data analysis only.

7. Acknowledgments

This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the DELOS Network of Excellence on Digital Libraries (Contract G038-507618).

References

[1] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, 92(2-3):285–305, November-December 2003.

[2] A. Benitez and S.-F. Chang. Automatic multimedia knowledge discovery, summarization and evaluation. *IEEE Transactions on Multimedia, Submitted*, 2003.

[3] M. Bertini, R. Cucchiara, A. Del Bimbo, and C. Torniai. Video annotation with pictorially enriched ontologies. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2005.

[4] M. Bertini, A. Del Bimbo, and C. Torniai. Ontologies for video digital libraries. In *Proceedings of Multimedia Data Mining (MDM/KDD)*, August 2006.

[5] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1210–1224, Oct. 2005.

[6] H. Eidenberger. Statistical analysis of mpeg-7 image descriptions. *ACM Multimedia Systems Journal*, 10(2):84–97, 2004.

[7] C. Grana, D. Bulgarelli, and R. Cucchiara. Video clip clustering for assisted creation of MPEG-7 pictorially enriched ontologies. In *Proc. Second International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Marrakech, Morocco, March 2006.

[8] C. Grana and R. Cucchiara. Linear transition detection as a unified shot detection approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 2007. (in press).

[9] J. S. Hare, P. Sinclair, P. Lewis, K. Martinez, P. Enser, and C. J. Sandom. Bridging the semantic gap in multimedia information retrieval - top-down and bottom-up approaches. In *Proc. $3rd$ European Semantic Web Conference*, Budva (Montenegro), June 2006.

[10] A. Jaimes, M. Christel, S. Gilles, R. Sarukkai, and W.-Y. Ma. Multimedia information retrieval: what is it, and why isn't anyone using it? In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 3–8, New York, NY, USA, 2005. ACM Press.

[11] A. Jaimes and J. Smith. Semi-automatic, data-driven construction of multimedia ontologies. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2003.

[12] A. Jaimes, B. Tseng, and J. Smith. Modal keywords, ontologies, and reasoning for video understanding. In *Int'l Conference on Image and Video Retrieval (CIVR 2003)*, July 2003.

[13] L. Kotoulas and I. Andreadis. Color histogram content based image retrieval and hardware implementation. *IEE Proc. Circuits Devices & Systems*, 150(5):387–93, 2003.

[14] V. Mezaris, I. Kompatsiaris, N. Boulgouris, and M. Strintzis. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):606–621, 2004.

[15] M. Seaborn, L. Hepplewhite, and J. Stonham. Color segmentation using perceptual attributes. In *Proceedings of the British Machine Vision Conference, BMVC99*, 1999.

[16] N. Simou, C. Saathoff, S. Dasiopoulou, E. Spyrou, N. Voisine, V. Tzouvaras, I. Kompatsiaris, Y. Avrithis, and S. Staab. An ontology infrastructure for multimedia reasoning. In *Proc. International Workshop VLBV 2005*, Sardinia (ITaly), September 2005.

[17] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias. A visual descriptor ontology for multimedia reasoning. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, Montreux (Switzerland), April 2005.

[18] J. Strintzis, S. Bloehdorn, S. Handschuh, S. Staab, N. Simou, V. Tzouvaras, K. Petridis, I. Kompatsiaris, and Y. Avrithis. Knowledge representation for semantic multimedia content analysis and reasoning. In *European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, Nov. 2004.

[19] S. Sural, G. Qian, and S. Pramanik. Segmentation and histogram generation using the hsv color space for image retrieval. In *ICIP (2)*, pages 589–592, 2002.

[20] D. C. Tseng and C. Chang. Color segmentation using perceptual attributes. In *International Conference on Pattern Recognition*, volume 3, pages 228–231, 1992.

[21] S. Vembu, M. Kiesel, M. Sintek, and S. Bauman. Towards bridging the semantic gap in multimedia annotation and retrieval. In *Proc. First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, Edinburgh (Scotland), May 2006.