MATCHING FACES WITH TEXTUAL CUES IN SOCCER VIDEOS

Marco Bertini, Alberto Del Bimbo and Walter Nunziati
Dipartimento di Sistemi e Informatica – Università degli Studi di Firenze
{bertini,delbimbo,nunziati}@dsi.unifi.it

ABSTRACT

In soccer videos, most significant actions are usually followed by close–up shots of players that take part in the action itself. Automatically annotating the identity of the players present in these shots would be considerably valuable for indexing and retrieval applications. Due to high variations in pose and illumination across shots however, current face recognition methods are not suitable for this task. We show how the inherent multiple media structure of soccer videos can be exploited to understand the players' identity without relying on direct face recognition. The proposed method is based on a combination of interest point detector to "read" textual cues that allow to label a player with its name, such as the number depicted on its jersey, or the superimposed text caption showing its name. Players not identified by this process are then assigned to one of the labeled faces by means of a face similarity measure, again based on the appearance of local salient patches. We present results obtained from soccer videos taken from various recent games between national teams.

1 Introduction

To provide effective archiving and retrieval of video material, video streams must be annotated with respect to their semantic content, producing metadata that is attached to the video data and stored in databases. This will permit, for example, to produce special video summaries for a sport program such those that recollect the best actions occurred during a typical soccer turn, or those where there are notable actions of a certain player.

Most of the existing works addressing annotation of sports videos have been devoted to parse the video stream at semantic level, in order to detect various events typical of the sport under consideration (like goals 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. A recent comprehensive review can be found in [9].

People identification from video sequences is a widely studied topic as well. Recent works are [3] and [6]. Both of them studied the problem of detecting faces from broadcasted video sequences, and how to find instances of the same person among all the detected faces. For the vast literature on face detection and recognition, the reader is also referred to the surveys presented in [8] and [10].

Among the works on video text recognition, we can cite [11] where a multiresolution approach that uses edge directions and neural networks is used to detect superimposed text. In [12] captions and scene text regions are recognized using mean, second- and third order central moments in the wavelet domain to train a neural network. Regarding the detection of jersey's numbers we can cite [13], where each image is segmented in color homogeneous regions, and Zernike moments are extracted from pipe-like regions to recognize numbers.

Associating interpreted textual content to faces has been investigated in the context of news video in [5] and [1]. They use several information sources typically available for this type of videos, such as transcripts and video captions, as well as faces automatically detected using color analysis to recognize skin tones.

Our goal is to provide automatic annotation of player's identity in close-up shots of soccer videos. The method performs its analysis in two steps: first, faces are detected and tracked in close–up shots (Sect. 2). When a face is detected, the frame is also searched for the player's number depicted on its jersey, or for superimposed text caption showing its name (Sect. 3). Textual cues of this type easily lead to an immediate identification of the player, simply checking team standings or official soccer organization websites. In the second step, players that have been detected, but not identified in the above described process, are assigned to one of the labeled faces using a face similarity measure, based on the appearance of local face patches (Sect. 4).

We bring contribution in two different areas. First, we exploit the presence of a very peculiar media of soccer videos, which is the "stream" of jersey's numbers, to perform recognition. Second, we improve the work presented in [2], introducing a novel method to identify and "read" textual cues from a generic video stream. The method relies on a combination of interest point detectors, and does not requires an unreasonable amount of manually annotated examples to work.

Results have been obtained from videos of recent soccer games between national teams. About 80% of accuracy has been achieved for the face and number detectors, while superimposed text caption detector achieved almost 95%. As can be expected, less accurate, but still promising, results have been obtained on the face matching task.

2 Extracting face sequences

The annotation process begins when a frontal face is detected in a frame. To avoid detection in non close–up shots, only faces bigger than a minimum size trigger the subsequent processing. To detect faces, we use a slightly modified version of the adaboost face detector of [7]. The basic algorithm was tuned to the specific case of the soccer domain by using negative ex-

amples taken from actual soccer videos. Moreover, we split the training phase in two steps: in the first step, seven stages of the adaboost training were carried out. The obtained face detector was then run on a selection of soccer videos, and false detection were extracted from these videos and added to the original training set, then the training phase is continued for the remaining stages.

Once a face is detected, it is tracked throughout the entire shot (back and forth), to collect a face sequence of the same individual. Given the position in the current frame, the tracker uses dense optical flow to estimate the face's position and scale in the next frame. Then these estimates are refined searching the highest correlation of the face region. To this end, the face's appearance is modeled with a 32–bins histogram in the HS space. The searching process is carried out out in a window of predefined size around the estimated position. Fig. 1 shows an example of multiple faces tracked in the same shot. Note the scale changes of the face indicated by the red box.

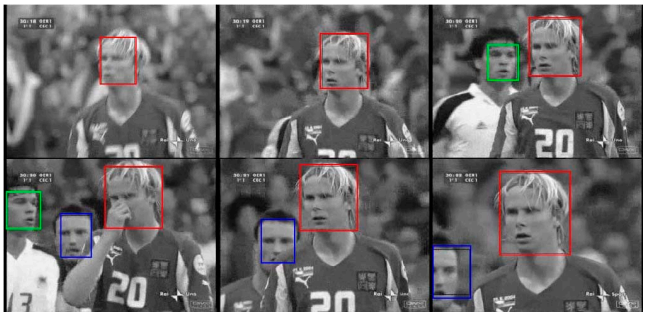

Fig. 1. Keyframes with multiple detections and tracking through scale.

3 Text detection and recognition

Text detection and recognition in videos still presents several issues that are to be solved, in particular when dealing with scene text. In general, the main problems that affect text in videos, are *i)* low resolution, *ii)* unknown text color, size, position, orientation and layout, *iii)* low contrast, color aliasing and unconstrained background. One of the typical techniques that is used to cope with some of these problems is the exploitation of temporal redundancy: in fact a superimposed caption is maintained in the image enough time, to let viewers to read it. In the case of scene text this technique can not be exploited since the appearance of scene text can hardly be controlled by the camera operators or by the program producers; moreover text orientation and rigidity may vary greatly.

We have developed an unified approach for detection of scene and superimposed text that does not require any knowledge or training on super-imposed captions or scene text features, and does not use temporal redundancy for the text detection and extraction.

The first processing step is the extraction and selection of image corners, using the Harris detector, as locally salient points. This feature allows to work with very small font sizes, typically used in caption. For each corner we analyze the surrounding area, to check if there are some neighbours, to filter out outliers that may be caused by high contrast objects in the background. Then an unsupervised clustering process is performed on the remaining corners, and the bounding box of each cluster in checked counting the percentage of pixels belonging to the surroundings of corners w.r.t. the bounding box area. This strategy reduces the noise due to high contrast background during static scenes, that typically produce small scattered zones of corners that can not be eliminated by the previous analysis.

In the second processing step the Maximally Stable Extremal Regions ([14]) are extracted. These regions allow to further refine the selection of text area candidates, and to segment the image in order to process it with a commercially available OCR system. The pixels resulting from intersection of pixels belonging to the corners' neighborhoods and the MSERs pixels are used as the base for the final image processing step that prepares the image for the OCR, using a K-Fill filtering that eliminates salt and pepper noise, and a color uniformity analysis that eliminates blobs like those of the team logos.

When applying this processing framework to superimposed captions and jersey's numbers we classify the resulting blobs according to their size, to separate the small numbers belonging to captions (e.g. the match score) and the large numbers of the jerseys. Fig. 2 shows the detected captions, Fig. 3 shows the jersey's number blobs.

4 Matching faces

To assign every non identified face to one of the player classes we exploit the fact that players are a fixed and somewhat limited population. More in detail, we considered each annotated example as an individual, avoiding to merge clusters relative to the same player.

An example of this situation is given in Fig. 4, where the unlabeled face in the center must be assigned to one of the labeled faces, which are (a subset of) faces annotated by means of number or text caption. The faces on the first row of the left side, and of the first and third row of the right side represent the same player. Considering each row as a distinct individual, we built a compact representation based on local facial features. This has the effects of increasing inter-class distances in our classification task, but at the cost of having an increased number of classes. Hence, for the unlabeled example there are three possible correct pairings. In practice, to

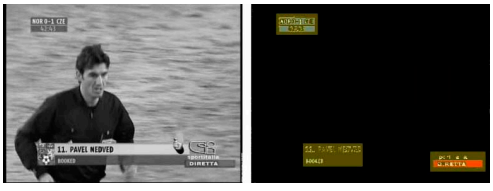

Fig. 2. Original frame and detected captions (highlighted).

Fig. 3. Original frame and detected jersey number blobs.

label an unknown face, we require to find a face of the same player with a similar pose and expression.

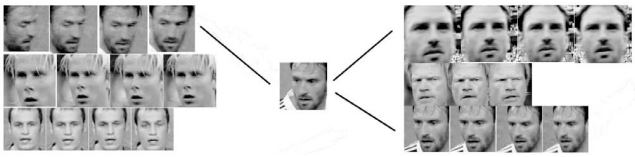

Fig. 4. Left - examples labeled by means of text or number. Right - an unlabeled example to be assigned to one of the labels. Lines represent possible correct pairings.

To cope with the large variation of poses and expressions we followed a part-based approach to recognition, similarly to [6], using the SIFT descriptor [4]. We experimented with several part–based representation schemes, and obtained the most satisfying results using three SIFT descriptors centered on the two eyes (20×20 pixel, with the face size normalized to be 80 pixels wide), and on the midpoint of the eyes (15×30 pixels). This choice is motivated by the facts that *a)* these are the most robust facial features to detect and track throughout the shots and *b)* the lower part of the face is often characterized by appearance changes due to variation in expression, that exceed those due to identity changes. The basic SIFT descriptor has been modified to avoid to include in the descriptor non-face part of the image. In particular, we rely on skin–maps to adaptively compute the weights of the components of the SIFT descriptor. For each pixel of the patch, its weight in the descriptor is cut to zero as the pixel falls off the region defined by the skin–map.

Matching is carried out between pairs of face sequences. Similarity is computed using the minimum distance between the two sequences. If U is a face track corresponding to a non–labeled player, and L is a labeled face track, their distance is computed as follows:

$$d(U, L) = \min_{i,j} \|U_i - L_j\|,$$

where U_i and L_j are two 384–length vector, and their distance is measured using the l_1 norm.

5 Results

5.1 Face and text detection results

On average, the system selected about 6000 frames for each game, providing approximately four minutes of close-up shots with name–face association. The average number of players identified is 12 for game, without repetition.

Table 1 reports performance of the number, face and text detectors. Number and text detection has been performed only when a face is detected, hence the "Present" column in this case is referred to the 94 shots obtained from the face detector, and the candidate text blobs have been recognized using OmniPage Pro OCR; analysis of jersey number blobs has been forced to detect numbers only. The adaboost–based face detector was trained with a C++ program, running on a Pentium 4 with 1.5GB of Ram. A cascade of 14 stages have been used. The procedure took about 16 hours for the face detector.

Table 1. Face, number and text detector performances.

Detector	Present	Detected	Correct	False	Missed
Face	118	104	95	9	23
Numbers	40	32	27	5	13
Text	16	16	15	1	1

Errors in recognition of numbers arise either if a number detector signals that a number is present when there is none, or when the recognized number is wrong. The first situation is very unlikely, since we require that recognition is stable for a minimum number of frames, while we experienced the second type of errors for certain numbers in particular: it happened for instance that the number 9 was detected instead of number 8. These errors are likely to occur either when the player is moving, the number is highly skewed with respect to the camera, or the jersey folded so that part of the number is hidden. Misses are due to images that are not sharp enough to let the detection of enough image corners (e.g. the player is not precisely focused). Caption detection and recognition shows good results because of the high contrast used and text stability. Name recognition is also improved using approximate string matching with a players' name database. Table 2 reports results of player identification on a sample game. Not surprisingly, detection of "face and caption" shots is more reliable than detection of "face and number" shots. This is mainly due to misdetections performed by the face detector, while the text detector correctly detected nearly all the shots where a caption or a number was present. Moreover, the number of close-up shots detected was fairly low if compared with the total number of close-up shots, where identification is not performed because neither jersey's number nor text caption was present.

Table 2. Summarized results of the annotation of a sample game for the player identification task.

	Present	Detected	Correct
Face and jersey's number	40	32	27
Face and caption	16	16	15

5.2 Face matching results

The face matching method was tested on a dataset taken from four games. Five players were selected, and for each player there were from five to ten face–sequences in the dataset, for a total of 38 sequences, manually annotated with the player's

identity. For each player, there were sequences obtained in the same shot, in different shots of the same game, and in different games (Fig. 5, top). This is the query set. Other 40 face–sequences of other players were added to increase the variability, for a total of 78 sequences in the dataset. Fig. 5, bottom shows examples taken from the query sequences. To obtain the representation described in Sect.4, eyes were manually marked in the first frame, and then tracked using a simple correlation based tracker. Fig.6 reports precision–recall curves for each of the players in the query set (thin curves). Curves have been obtained adopting the "leave–one–out" scheme. The thick curve shows the global PR curve, obtained averaging all the results.

Fig. 7 shows a typical result, obtained using the top–left face as a query. As can be expected, best results are obtained matching sequences taken from the same shot, while performances decrease for sequences taken for different shots or games. However, a high recall is usually achieved for an average precision of 70%, which is a reasonably encouraging result. We found that face pose affects consistently our similarity measure: often, two faces of different individuals in the same pose are found more similar then two faces of the same person in different poses.

Fig. 5. **Top:** keyframes of the same player in the test set. From left to right: two images of the same player from the same shot, the same player in a different shot, and the same player in a different game. **Bottom:** sample faces from the query set.

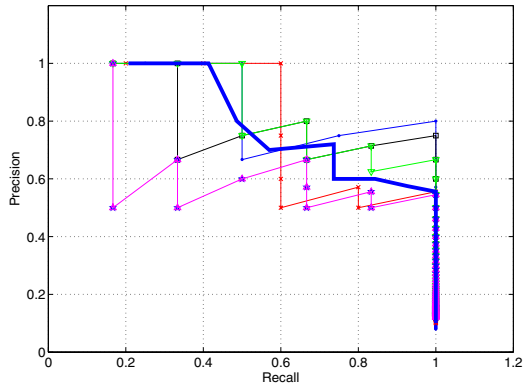

Fig. 6. Precision recall curves. The thin lines are averages of queries using different sequences of the same player, the thick line represent the global average.

Fig. 7. Results of a sample query. First keyframe represent the query track, red boxes indicate correct matches. Only the first 12 results are shown, out of 78.

6 Conclusions and future work

A method to understand the identity of players from close–up images of soccer videos has been presented. Starting from the detection of a frontal face, the system tries to label the face using textual cues, automatically read from the player's jersey or from text caption. This is performed with a novel method based on a combination of interest region detectors. If no textual cue is found, a face similarity measure is adopted to link unknown faces with some of the labeled faces. Although the system is not yet completely automatic, preliminary experimental results are certainly encouraging. In particular, very good results have been obtained with the textual cues recognizer, while the face representation scheme needs a more consistent effort to be improved.

7 References

[1] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, D. A. Forsyth. *"Names and Faces in the News."*. In Proc. of CVPR, 2004.

[2] M. Bertini, A. Del Bimbo, and W. Nunziati, *"Player identification in soccer videos."* In Proc. of MIR, 2005.

[3] M. Everingham, and A. Zisserman. *"Automated Person Identification in Video."* In Proc. of CIVR, 2004.

[4] D. Lowe, *"Distinctive image features from scale–invariant keypoints."* International Journal of Computer Vision, 60, 2, 2004.

[5] S. Satoh, Y. Nakamura, and T. Kanade. *"Name-It: Naming and Detecting Faces in News Videos."* IEEE MultiMedia, January-March 1999.

[6] J. Sivic, M. Everingham, and A. Zissermann. *"Person spotting: video shot retrieval for face sets."* In Proc. of CIVR, July 2005.

[7] P. Viola and M. Jones. *"Rapid object detection using a boosted cascade of simple features."* In Proc. of CVPR, 2001.

[8] M.H. Yang, D.J. Kriegman, and N. Ahuja. *"Detecting faces in images: a survey."* IEEE TPAMI, Jan. 2002.

[9] X. Yu and D. Farin. *"Current and Emerging Topics in Sports Video processing."* In Proc. of ICME, 2005

[10] W.Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. *"Face recognition: A literature survey."* ACM Computing Surveys (CSUR), December 2003.

[11] R. Lienhart and A. Wernicke. *"Localizing and Segmenting Text in Images and Videos."* IEEE TCSVT, vol. 12, no. 14, April 2002.

[12] H. Li, D. Doermann and O. Kia. *"Automatic text detection and tracking in digital video."* IEEE TIP, vol. 9, no. 1, pp.147-156, April 2000.

[13] Q. Ye, Q. Huang and S. Jang. *"Jersey Number Detection in Sports Video for Athlete Identification."* In Proc. of VCIP, July 2005.

[14] J. Matas, O. Chum, M. Urban and T. Pajdla. *"Robust Wide Baseline Stereo from Maximally Stable Extremal Regions."* In Proc. of BMVC, vol. 1, pp. 384-393, London, 2002.