AN OBJECT DETECTION METHOD FOR DESCRIBING SOCCER GAMES FROM VIDEO

Okihisa UTSUMI†,*, *Koichi MIURA*††, *Ichiro IDE*‡, *Shuichi SAKAI*††, *Hidehiko TANAKA*††

†Graduate School of Engineering, The University of Tokyo
††Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{utsumi, miura, sakai, tanaka}@mtl.t.u-tokyo.ac.jp

‡National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
ide@nii.ac.jp

ABSTRACT

We propose a novel object detection and tracking method in order to detect and track objects necessary to describe contents of a soccer game. On the contrary to intensity oriented conventional object detection methods, the proposed method refers to color rarity and local edge property, and integrally evaluate them by a fuzzy function to achieve better detection quality. These image features were chosen considering the characteristics of soccer video images, that most non-object regions are roughly single colored (green) and most objects tend to have locally strong edges. We also propose a simple object tracking method, that could track objects with occlusion with other objects using a color based template matching. The result of an evaluation experiment applied to actual soccer video showed very high detection rate in detecting player regions without occlusion, and promising ability for regions with occlusion.

1. INTRODUCTION

Following the increase of broadcast video data, it is becoming important to index and store them considering their retrieval and recycling. In order to enable detailed indexing, understanding the semantic contents of the video is inevitable. In this paper, we focus on soccer (football) game broadcast video, which is a popular television program, and various queries that require contents understanding is expected. There are highlight event detection methods for sports videos that refer to camera operations [1], cheers (audio volume) [2], keyword spotting [3], and so on, but these employ only surface clues, which do not enable thorough understanding of what is actually happening in the video.

Concretely speaking, contents understanding of soccer videos could be considered as understanding the process and strategy of the games. In order to realize such understanding of games from videos, detection and tracking of objects (players, ball, and lines) is required. The detection and tracking reveal the movement of the players and the ball on the field, which could be used to retrieve certain plays ('pass', 'shoot', etc.) or to understand the overall trend and strategy of the game.

Although detection and tracking of objects have been a popular topic in computer vision and image processing fields, most of them assume special conditions such as fixed camera or single moving object. Thus detecting and tracking of objects is difficult in soccer videos, where cameras are not fixed and numerous objects move in various directions.

Fig. 1 shows the process of player detection and tracking. In this paper, we will mainly focus on introducing and evaluating a novel object detection and tracking method that refers to color and local edge properties, to transcend the ability of conventional methods that simply refer to intensity. First, gallery and other non-field regions are excluded to extract the field region as described in **2.** Next, object detection is performed applying the method described in **3.** Finally, player regions are detected and tracked as described in **4.**

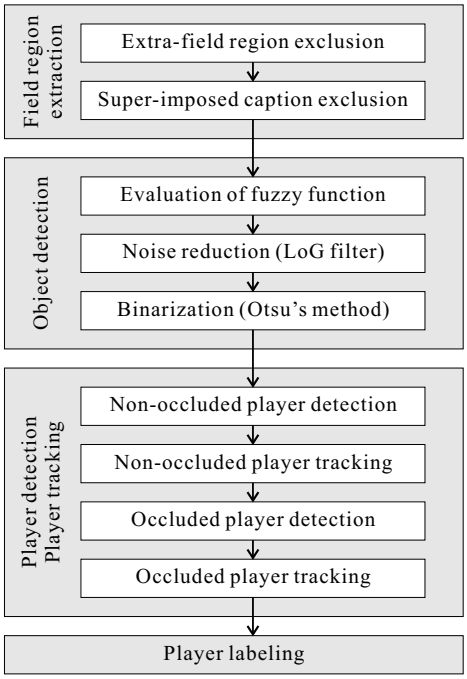


Figure 1: Player detection and tracking process.

2. FIELD REGION EXTRACTION

In order to concentrate on detection and tracking of objects that are necessary for contents understanding of the soccer game, field

*Currently at Matsushita Electric Corp.

region should be extracted, and objects that exist outside the field should be excluded. Thus extra-field regions and super-imposed captions are excluded from the image, and only objects inside the extracted field region are considered in subsequent processes.

2.1. Extra-field region exclusion

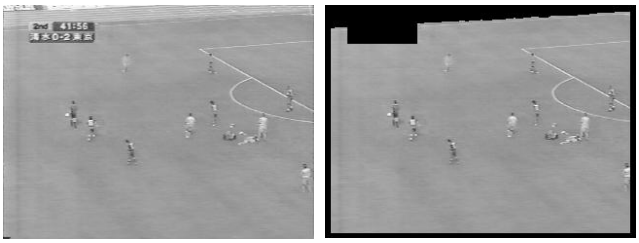
Extra-field regions such as the gallery should be excluded. In soccer video, field region is roughly green colored. Thus first, green regions in the image are detected. We referred to the hue H defined in Smith's hexagonal cone model, and defined a certain range ($\frac{1}{3}\pi \leq H \leq \frac{5}{6}\pi$) to detect green colored pixels. A continuous region with the largest area is considered as the field region, since there are occasionally small green colored regions outside the field.

2.2. Super-imposed caption exclusion

Even after the previous extra-field region exclusion, there are still super-imposed captions (ex. score, time) inside the field region that should be excluded. The following procedure is taken to detect caption regions, taking advantage of the characteristics that captions tend to 1) appear in the same position, 2) have high and stable intensity, and 3) appear horizontally.

1. Detect edges by applying gradient filter to the intensity.
2. Compare the edges of five serial frames and eliminate the uncommon ones.
3. Divide the image by 10×10 pixel blocks, and if the edge in the block is dense, consider it as a caption region candidate.
4. If more than 4 horizontally adjoining blocks are candidates, consider the concatenated block as a caption region. One interval block is allowed for a space between characters.

Fig. 2 shows an example of field region extraction.



(a) Original image (b) After field region extraction

Figure 2: Field region extraction.

3. OBJECT DETECTION REFERRING TO COLOR RARITY AND LOCAL EDGE PROPERTY

Most object detection methods are intensity oriented. This incurs the problem that some parts of object regions are occasionally absorbed into the background region, since color information is neglected when calculating the intensity. This problem degrades the detection ability before applying any kind of methods. In addition, regions with weak edges tend to be oversighted affected by stronger edges in the global image. For example, in soccer images, player regions are well detected due to relatively strong edges, but lines are frequently oversighted due to relatively weak edges.

To compensate for these problems, we propose to refer to color rarity and local edge property when performing object detection. Referring to color rarity enables to extract regions that were not wholly extracted referring simply to intensity, based on the premise that objects on the field are differently colored (i.e. less frequent) compared to the roughly single colored (green) field. Referring to local edge property enables to detect globally weak but locally strong edges. Color rarity is determined using a RGB color histogram, and local edge property is evaluated within a 3×3 pixel window surrounding the pixel in concern. The two features are integrally evaluated by a fuzzy function, and object regions are extracted by binarization using the so called Otsu's method.

3.1. Evaluation function

We define an evaluation function that evaluates whether each pixel belongs to an object region or not, based on fuzzy operation that integrates the two features; color rarity and local edge property. Fuzzy operation was adopted here, since it evaluates the tendency of various properties, which is suitable to handle such heterogeneous features integrally.

The evaluation function O at pixel $p(i,j)$ is defined as:

$$O(i,j) = \max(R(i,j), \min(E_R(i,j), E_G(i,j), E_B(i,j))) \tag{1}$$

where R and E_R, E_G, E_B represent the evaluation of color rarity and local edge property (in each of R,G,B color space), respectively. R,E,O have the value within $[0,1]$, and the higher the value is, the more the pixel is considered to belong to an object region, and vice versa. Evaluating the minimum of E_R, E_G, E_B is based on the assumption that edge properties are strong in all of R,G,B color space, but on the contrary, noises tend to appear only in certain color spaces. Next, evaluating the maximum of R and $\min(E_R, E_G, E_B)$ enables to detect both significant color rarity and/or edges.

Evaluation of color rarity R at pixel $p(i,j)$ is defined as:

$$R(i,j) = \begin{cases} 0 & (H(c(i,j)) > 2\overline{H}) \\ 1 - \frac{H(c(i,j))}{2\overline{H}} & (otherwise) \end{cases} \tag{2}$$

where $c, H, \overline{H}$ represent the color at pixel $p(i,j)$, the global color histogram of the entire image, and the average of the distribution H, respectively.

Next, evaluation of local edge property E at pixel $p(i,j)$ is defined as:

$$E_K(i,j) = \frac{1}{8}\sum_{n=1}^{8} e_K^n(i,j) \tag{3}$$

where $K = \{R,G,B\}$ and e_K^n defined as:

$$d_K^n(i,j) = |c_K^n(i,j) - c_K(i,j)| \quad (n = 1,2,...,8) \tag{4}$$

$$e_K^n(i,j) = \begin{cases} 1 & (d_K^n(i,j) > 2\overline{d_K}) \\ \frac{d_K^n}{2\overline{d_K}} & (otherwise) \end{cases} \tag{5}$$

where $c_K, c_K^n, \overline{d_K}$ represent the RGB color value at pixel $p(i,j)$, at eight surrounding pixels $p^n(i,j)$, and the global average of $d_K^n(i,j)$, respectively.

3.2. Object detection

After evaluating O at every pixel, LoG (Laplacian of Gradient) filter is applied to reduce noises. This combined filter is employed since simply applying Gradient filter eliminates not only noises but also edges. Next, Otsu's method is applied to determine the threshold whether each pixel belongs to an object or not. Fig. 3(a)

shows an example of the result of object detection applying the proposed method. For comparison, Fig. 3(b) shows the result obtained by simply applying Sobel filter to the intensity of each pixel. In Fig. 3(b), some players are not wholly detected as one region, since the player's body's color was not referred and the intensity was absorbed into the background. In addition, lines are chopped into pieces, since they did not have strong edges compared to other objects. Fig. 3(a) shows that the proposed method has overcome these problems.

(a) Applied proposed method (b) Applied Sobel filter to intensity

Figure 3: Result of object detection.

4. PLAYER DETECTION AND TRACKING

After object detection, each object should be classified and tracked. Here we concentrate on detecting and tracking player regions.

Major moving object recognition methods could be categorized to a) boundary-based and b) region-based. As an example for a), there is the active contour model that applies SNAKES in time space [4]. Such methods are frequently used to track moving objects, but requires precise initialization, which is not realistic when applying to broadcast sports video. As an example for b), there is the region segmentation and unification method [5]. There are various works on region segmentation, but the common problems are over-segmentation and sensitivity to initial cluster shape.

We detect player regions in a heuristics way, since the above-mentioned methods are too time consuming. The proposed method combines results from several adjacent frames, in order to enhance the detection ability.

4.1. Non-occluded player detection

Among object regions, player region is defined as:

$$15 \leq h \leq 50, 6 \leq w \leq 45, w \leq h \tag{6}$$

where h and w represent the height and the width of the object region in concern, respectively. After detecting player region candidates that fulfill these criteria, average area size of the regions is calculated in the upper and the lower half of the image separately. A player region's area size should be smaller than 1.5 times of the average size of the half it exists in. This additional criterion is employed under the estimation that all players' sizes should be approximately the same, and that players closer to the camera (lower half) are larger than further ones (upper half). Fig. 4 shows an example of the player detection result. Players without occlusion are well detected.

4.2. Occluded player detection and player tracking

Player tracking is done by the following procedure:

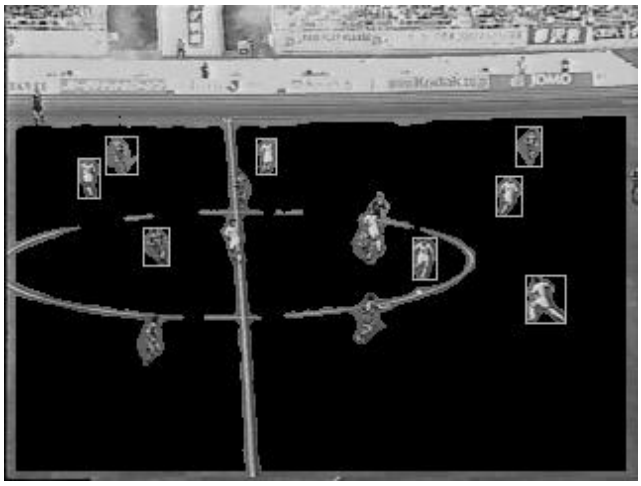


Figure 4: Result of non-occluded player detection.

1. Detect player-player occlusion
2. Track players without occlusion
3. Detect player-non player occlusion
4. Track players with occlusion

4.2.1. Player-player occlusion detection

When several players almost completely occlude each other, the detection criteria in **4.1.** mis-detects the occluded region as a single player. To exclude such a mistake, each player region is temporally tracked to detect merging and division of players regions. The detection is performed by comparing the center of gravity of each player region between adjoining frames. When two players' centers of gravity are detected within a player region of the next frame, merging is detected, and if vice versa, division is detected.

4.2.2. Tracking players without occlusion

After player-player occlusion detection, players without occlusion are tracked. Two adjoining frames are compared, and the player regions with the largest overlap between the two frames are considered as the same player. To prevent mis-tracking in case of fast camera motion, the overlap should be larger than a certain area size (set to 30 pixels).

4.2.3. Player-non player occlusion detection

Even after the tracking in *4.2.2.*, there might be player regions that are not tracked. This is because a player region that occludes with non player objects (usually lines) are oversighted in the player detection process in **4.1.**, due to the shape and size of the integrated region. These oversighted regions are considered as player-non player occlusion.

4.2.4. Tracking players with occlusion

In order to track player regions with occlusion as detected in *4.2.1.* and *4.2.3.*, color based template matching is performed within 3 pixels surrounding the player in concern. To evaluate the correlation C to the template, the following function [6] is defined:

$$C = \frac{\sum_{c=1}^{1000} \min(H_T(c), H_i(c))}{\sum_{c=1}^{1000} H_T(c)} \tag{7}$$

where H_T, H_i are the color histograms of the template and the region in concern, respectively. Here the color histogram consists of 1,000 bins (10 bins each for R, G, B) in order to cope with small player regions with few pixels. The region with the largest C is considered as the tracked player region, if C is larger than a threshold (set to 0.4).

5. EXPERIMENT

We applied the proposed method to two actual soccer videos in order to evaluate the player detection and tracking ability.

5.1. Condition

Tab. 1 shows the property of each of the two video used for the experiment. Note that these were taken from different games, and were completely continuous without any shot boundaries.

Table 1: Property of the videos.

Duration	30 seconds
Format	Motion JPEG (transformed into 24bit color bitmap images)
Resolution	320×240 pixels
Frame rate	30 frames/second

5.2. Result

Tab. 2 shows the detection rate of players without occlusion. The detection rate represents the percentage of correctly detected players among all existing players in all frames (4,680 player regions in total). This detection does not require any tracking, so it is the result of simply applying the criterion defined in **4.1.** To show the significance of the proposed method, the result is compared with the result derived from applying conventional intensity based object detection (Applied Sobel filter to intensity as exemplified in Fig. 3(b)). The proposed method showed high detection rate especially in recall. Mis-detection was due to the mistake in extra-field region exclusion, and oversight was due to the blur caused by fast camera motion. The poor ability of the conventional method was mainly due to absorption of the player into the field region. This shows the dominance of referring to color information to detect objects in the proposed method.

Next, Tab. 3 shows the detection rate of all players including those with occlusion (7,016 player regions in total). This requires tracking as described in **4.2.** The decrease in detection ability is mainly due to mis-tracking of occluded players. Mis-tracking could be categorized to a) mis-detection of other objects with similar color property during template matching, and b) oversight of players. a) is due to loose criteria for similarity evaluation to maintain robustness, and b) is due to fast camera motion and rapid change in a player's pose.

6. CONCLUSION

We have introduced a novel object detection and tracking method adopted to soccer video, which showed high detection rates compared to conventional method. Although the overall player detection rate is still not sufficient to label players and precisely describe a game, we consider the result promising.

Table 2: Detection rate (Players without occlusion).

	Conventional method		Proposed method	
	Recall	Precision	Recall	Precision
Video 1	79.3%	51.1%	97.1%	81.2%
Video 2	80.1%	51.1%	91.1%	76.1%
Overall	79.7%	51.1%	94.1%	78.6%

Table 3: Detection rate (All players).

	Recall	Precision
Video 1	72.7%	62.3%
Video 2	60.1%	76.3%
Overall	64.9%	66.9%

Considering object detection, introducing knowledge on object shapes and colors should improve the ability, which is important to ensure sufficient tracking ability. Detection and tracking of players with occlusion still requires further improvement. In order to cope with rapid change of a player's pose, simple template matching is not sufficient, thus requires more robust matching and combination with motion estimation. The former should be realized by reducing noises in the template, and the latter is employed in [7, 8], and should be considered in the future.

We will further investigate on labeling players and understanding the trend and strategy of the game for thorough understanding of sports video contents. In addition, videos taken under various situations should also be applied to confirm the robustness of the proposed method.

7. REFERENCES

[1] Y. Iwai, J. Maruo, M. Yachida, T. Echigo, and S. Iisaku, "A framework of visual event extraction from soccer games," *Proc. 4th Asian Conf. on Computer Vision*, vol. 1, pp. 222–227, 2000.

[2] Y. L. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated image and speech analysis for content-based video indexing," *Proc. Intl. Conf. on Multimedia Computing and Systems*, pp. 306–313, 1996.

[3] N. Nitta, N. Babaguchi, and T. Kitahashi, "Extracting actors, actions and events from sports video –a fundamental approach to story tracking–," *Proc. 15th Intl. Conf. on Pattern Recognition*, pp. 718–721, 2000.

[4] G. Kuhne, S. Richter, and M. Beier, "Motion-based segmentation and contour-based classification of video objects," *Proc. 9th ACM Intl. Conf. on Multimedia*, pp. 41–50, 2001.

[5] T. Echigo, R. J. Radke, P. J. Ramadge, H. Miyamori, and S. Iisaku, "Ghost error elimination and superimposition of moving objects in video mosaicing," *Proc. IEEE Intl. Conf. on Image Processing '99*, 1999.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *Intl. J. on Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[7] T. Misu, M. Naemura, S. Sakaida, W. Zheng, and Y. Kanatsugu, "Robust tracking method of soccer players based on data fusion of heterogeneous information (in Japanese)," *Tech. Rep. of IEICE, PRMU2001-67*, 2001.

[8] K. Matsui, M. Iwase, M. Agata, T. Tanaka, and N. Onishi, "Soccer image sequence computed by a virtual camera," *Proc. Conf. on Computer Vision and Pattern Recognition '98*, pp. 860–865, 1998.