AUTOMATIC EXTRACTION OF SEMANTIC COLORS IN SPORTS VIDEO

Lei Wang[†]*, Boyi Zeng*[†]*, Steve Lin*[‡]*, Guangyou Xu*[†]*, Heung-Yeung Shum*[‡]

[†]: Computer Department, Tsinghua University, Beijing, 100084, P.R.China
[‡]:Microsoft Research Asia, No. 49, Zhichun Road, Haidian, Beijing, 100080, P.R.China.

ABSTRACT

Color has been widely used in sports video analysis. Previous techniques, however, require color models from prior information or user interaction, and do not address the problem of how to automatically form color models from a video in an arbitrary sports setting. In this paper, we propose an automatic technique for extracting color models of the playing surface and the team uniforms, which can be used in higher-level processes such as tracking and recognition. Unlike most previous methods, our approach is capable of handling multi-colored patterns like striped uniforms and playing fields. Multiple forms of color processing are used to analyze video frame content, which are then used iteratively to refine the color models. The results of our color modeling technique have been applied to shot classification, and experiments on videos of different sports have verified our approach.

1. INTRODUCTION

Semantic colors refer to meaningful colors that could be used to identify objects in video. Although color-based object recognition may not be practical in many video scenarios, it is suitable for sports, where colors are purposely used to differentiate players and clearly defined rules constrain the action. A result of this is that the colors of the playing surface, the player uniforms and the referee outfit usually consist of one or two (striped) dominant colors. These semantic colors are quite useful for more detailed analysis of sports video.

1.1. Relevant Work

Much work on processing sports video has focused on color information. Color histogram differences between successive video frames has commonly been used for shot segmentation [1]. Xu et al. [2] used grass-area-ratio for play/pause detection in soccer video. Xie et al . [3] extended this work to incorporate HMM models. Babaguchi et al . [4] used the field color ratio to detect zoom-out views in American football video. Li et al. [5] similarly used the green color ratio to detect close-up views based on normalized R and G components. [6, 7] used color for ball tracking in soccer and baseball respectively. [8] used color for player tracking to generate a virtual sequence for a soccer game.

This research is supported by Chinese NSF grant No. 60273005.

Although these systems have been successful in their respective applications, they generally utilize predefined color models specific to their sports setting. As a result, new color models often must be provided to process a different sporting event. Ekin and Tekalp [9] used mean values and thresholds to automatically model the dominant color. However, only a single dominant color is modeled, which excludes striped playing fields (see Figure 1 for examples). Further more, their team uniform is modeled by a color histogram, but the histogram is formed from user-specified examples of the uniform in the video sequence. This necessitates manual intervention for each match.

1.2. Our Approach

We approach the color modeling problem in a fully automatic way, and without any prior knowledge of the sport. The Gaussian mixture model [10] was employed to estimate the dominant color, taking into account possible stripe patterns. A voting scheme is then taken to filter the extracted dominant colors. With the estimated semantic color, an iterative scheme is proposed to refine the color model into one with high accuracy.

The formulation of our approach is based on basic principles of robust estimation. For the problem of semantic color estimation in video, there is no lack of color samples, which are often available on the order of a million pixels per second. The challenge is how to form accurate color models from often unreliable samples with noise. By strictly filtering the pixels for estimation, we are able to robustly extract the semantic color.

The major contributions of our work include the following:

1. a novel algorithm to determine the playing field color including striped fields in a possibility framework;
2. an automatic method to extract player uniform colors including striped uniforms;
3. a robust new approach to classify shot types using both playing field color and uniform colors.

In this paper, we focus on the extraction of playground color and player uniform colors, which provide the most critical information for analyzing sports videos. Some semantic colors, such as the uniform color of the referees, are not taken into account in this work because of their relative unimportance for further video analysis, but our color

Fig. 1. Example of a striped playing field and player uniform.

modeling framework could easily be augmented to include additional components.

2. MODELING PLAYING FIELD COLOR

A number of different color models have been explored in the literature. In this paper, we work in the CIE LUV color space, since it is a perceptually uniform derivation of the standard CIE XYZ space, meaning that two colors of equal cartesian distance in the color space are also equally distant perceptually.

Rather than processing every frame, our technique analyzes only one out of every N ($N = 30$ for all experiments) frames for determining the field color. Dense frame sampling leads to an unnecessary computational burden, since there is typically little difference among consecutive frames. Moreover, dense frame sampling over a short time interval may mistakenly emphasize a close-up or out-of-field view in which the playground color is not the dominant color. All the pixels in the sampled frames are projected into an LUV space where each dimension is discretized into 64 bins, and then a Gaussian mixture model consisting of two Gaussians $G(\mu_1, \Sigma_1), G(\mu_2, \Sigma_2)$ is used to estimate the color distribution with the EM algorithm [10]. To facilitate convergence of the EM algorithm, we initialize the two Gaussian mean values μ_1, μ_2 according to two separate peaks in the color distribution. The μ_1 is set to the color c_0 of the overall histogram peak, and μ_2 is initialized to the second peak color c that maximizes

$$D_c \times F_c,$$

where D_c is the cartesian distance from color c to c_0, and F_c is the frequency of the color c in the sampled frames. This dependence on distance allows two distinct peaks to be found, and avoids both Gaussians modeling the same color. Essentially, the second peak color is the most frequent color that is not too close from the first peak color, as illustrated in Figure 2.

Once the Gaussian mixture model is estimated, we determine whether the playing surface is of a single color or of two striped colors. If the peaks of the estimated two Gaussians are close to each other ($||\mu_1 - \mu_2|| <$ threshold), or one of the Gaussians is significantly weaker than the other ($F_{\mu_2} << F_{\mu_1}$), the playground is considered to be of a single color and is modeled by a single Gaussian instead.

For each Gaussian model, we can express the conditional

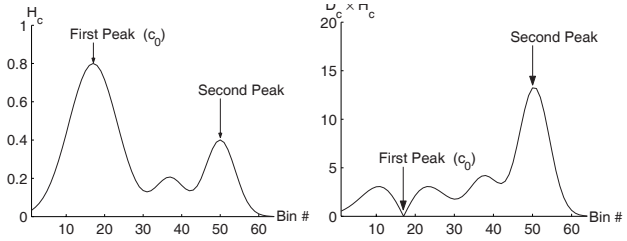

Fig. 2. Two-dimensional illustration of Gaussian center initialization from a color distribution. (Left) The mean value of one Gaussian is initialized to the color of the global distribution peak. (Right) The mean value of the second Gaussian is determined from a measure dependent on peak height and distance from the global peak.

probability p of a given pixel x sharing its color as

$$p(x|\mu_k, \Sigma_k) = \frac{exp[-(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)/2]}{2\pi|\Sigma|^{1/2}}$$

for $k = 1, 2$. A pixel x is then considered a playground color pixel when $p(x|\mu_1, \Sigma_1) > t$ or $p(x|\mu_2, \Sigma_2) > t$ for some threshold t.

To refine the playground model, the frame data is filtered using the current color model, and the filtered data is used to re-estimate the Gaussian mixture model. In the filtering process, we first discard the frames in which the playground color is not the dominant color, e.g., when less than half of the frame pixels have the playground color. Second, within each frame, the largest contiguous region having the playground color is computed, and only those pixels within such regions are included in the estimation of the next Gaussian model. This procedure is iterated until the change in the estimated Gaussian centers μ_1, μ_2 falls below a threshold. As shown in Section 5, by filtering out the outliers in the data, which can perturb the Gaussian centers, a more accurate color model can be obtained.

3. MODELING PLAYER UNIFORM COLORS

Automatic color modeling of player uniforms is a relatively difficult problem mainly due to three reasons:

1. Unlike the playground color, the player uniform color usually is not the dominant color of a sports video clip or even of a frame;
2. Uniform colors of TWO teams need to be determined;
3. Player uniform colors must be clearly distinguished from the uniform colors of the referee, goal keeper, coach, etc.

To deal with the first obstacle, we utilize some contextual clues to aid in locating player uniforms. Specifically, our technique utilizes robust face detection and the physiological structure of the human body to constrain the location of uniforms. We deal with the last two obstacles using a

voting scheme based on the empirical observation that most close-up shots in sports videos are of players.

The robust face detector presented in [11] is employed to detect faces in the video clip (1 frame every N frames). Note that we need not detect every face in the video, we just need several reliable ones to extract the color model, so a non-perfect face detector is fine for our approach. Additionally, we incorporate a voting scheme for robustness. For each detected face, a box indicating the human body area is predicted by the size and position of the face according to a general model of an upright human body, as shown in Figure 3. Only the non-playground pixel colors within the box are taken to estimate the Gaussian mixture model of the player uniform, in a manner similar to that described in the previous section. Note that both two-color and single-color uniform models can be learned properly.

The Gaussian color model of each detected uniform is recorded in a candidate list. If a new model is similar to a current list entry, as determined by the squared distance of the Gaussian centers, then the vote count of the list entry is incremented and the original Gaussian color model is updated with these additional new samples. Otherwise, a new list entry is added. The first two color models that reach a threshold vote level V ($V = 5$ for all of our experiments) are taken as the uniform color models for two teams respectively.

Fig. 3. Prediction of body region by face detection.

4. SHOT CLASSIFICATION BASED ON EXTRACTED COLORS

Color models substantially facilitate a deeper analysis of sports video. As has been explored in number of works, the extracted semantic colors can be used for many purposes such as shot segmentation, play-pause detection, structure analysis and tracking. Due to page limitations, in this paper we describe the application of our extracted color models only to shot classification.

We categorize shots according to the definition given in [9]. However, since we have obtained color models of player uniforms, we are additionally able to distinguish close-up shots from out-of-field shots, which were considered a single class in [9]. Furthermore, our technique can classify the in-field medium shot more accurately with an explicit analysis of players.

Certain parameters of each frame are used in our method to classify sports video shots. R_g is the ratio of playground color pixels, N_p is the number of players. N_p is estimated by counting connected components of player color regions whose bounding box satisfies a size and aspect ratio constraint. This type of bounding box processing has been widely used to reduce noise effects [9]. And R_p is the ratio of player uniform color pixels. Note that only the player uniform color pixels within a valid bounding box are taken into account for R_p. This constraint reduces errors caused by a background with a similar color to the player uniform colors. Besides these parameters, two thresholds T_g and T_p are set by off-line training, and are kept constant for all our experiments.

With these quantities, the shots are categorized as follows.

Global: $R_g > T_g$ and $R_p \leq T_p$ (a large playground color ratio, and a small player uniform color ratio)

Out-of-field: $R_g \leq T_g$ and $R_p \leq T_p$ (a small playground color ratio, and a small player uniform color ratio)

Close-up: $R_p > T_p$ and $N_p \leq 2$ (a large player uniform color ratio and few players)

In-field medium: $R_p > T_p$ and $N_p > 2$ (a large player uniform color ratio and many players).

Once the system is setup, it works automatically for video clips regardless of the sport, which have significantly different playground color and player uniform colors.

5. EXPERIMENTS

Our system was tested on ten sports video clips (six soccer clips from World Cup 2002, two soccer clips from the English Premier League 2003, and two basketball clips from the NBA 2002) that add up to about 350 minutes. An analysis of the experimental results is presented in this section.

5.1. Extracting Playground Color

Figure 4 shows results for using the computed playground color model to extract the playing surface from video frames. Our approach works well for both single colored and double colored playing surfaces, as well as for different sports. The improvement of the color models from the iterative refinement algorithm is evident. The post area (in purple) of the basketball clip is not included in playing field model, since its Gaussian model is relatively weak compared to the brown wood surface. For such cases, higher level processing based on motion analysis could be used to identify such areas, and this possible extension would benefit from the current color models.

5.2. Extracting Player Uniform Colors

Three examples of using the estimated color model to extract player uniforms are shown in Figure 5. The color models were correctly computed for all the ten video clips and twenty team uniforms. However, when the much of the audience is dressed in the same color as a team and forms the video background, even a correct color model may fail to generate a correct player bounding box and player uniform color ratio. Note that the results in this work for playground color extraction and player uniform color extraction were obtained with no prior knowledge or human interaction.

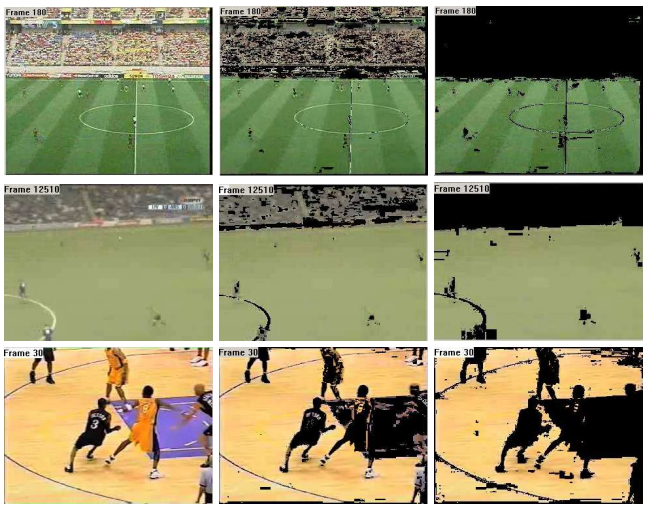

Fig. 4. Results for extracting the playground color. From top to bottom: results for World Cup video A1, results for English Premier League video B2, results for NBA video C1; From left to right: original frame, playground color pixels without refinement, playground color pixels with refinement.

Fig. 5. Results for extracting player uniforms using the recovered color models. Top row: original frame; Bottom row: player uniform color pixels with bounding boxes showing a valid player region; From left to right: results for England Premier League video B1, results for World Cup video A5, results for NBA video C2.

5.3. Shot Classification Based on Extracted Colors

A 10-minute video clip from World Cup A1 was used to learn the thresholds T_g and T_p. These threshold values were then kept constant for all the test video clips. Since the video presentation style of soccer and basketball are quite different, we considered only soccer videos for this application. The shot classification results are shown in Table 1. Compared to the results reported in [2, 9], there is an obvious improvement.

Sequence	WC A1	WC A2	WC A3	WC A4
Accuracy (%)	94.3	93.2	92.2	91.1
Sequence	WC A5	WC A6	PL B1	PL B2
Accuracy (%)	92.9	93.7	90.5	91.3

Table 1. Results for our shot classification algorithm.

6. CONCLUSIONS

An automatic system for extracting playground color and player uniform color models is presented in this paper. The system is general so that different kinds of sports videos can be processed. With the presented Gaussian mixture model, our technique can handle the case of multiple-color patterns like striped turf and uniforms. To facilitate uniform color modeling, face detection is employed to predict the body region. The color modeling results are refined by voting and filtering to reduce noisy samples as much as possible. One application, shot classification, demonstrates the effectiveness of the extracted color models. In future work, we intend to perform experimentation on a broader range of sports and to use the results for more detailed sports video analysis.

7. REFERENCES

[1] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *CSVT*, vol. 12, no. 2, pp. 90–105, February 2002.

[2] P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and systems for segmentation and structure analysis in soccer video," in *ICME*, Japan, 2001.

[3] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," in *ICASSP*, Orlando, Florida, May 2002, vol. 4, pp. 4096–4099.

[4] N. Babaguchi, Y. Kawai, Y. Yasugi, and T.Kitahashi, "Linking live and replay scenes in broadcasted sports video," in *MIR*, Los Angeles, USA, November 2000, pp. 205–208.

[5] B.X. Li, H. Pan, and M.I. Sezan, "A general framework for sports video summarization with its application to soccer," in *ICASSP*, Hong Kong, April 2003, vol. 3, pp. 169–172.

[6] D. Yow, B.L.Yeo, M. Yeung, and G. Liu, "Analysis and presentation of soccer highlights from digital video," in *ACCV*, Singapore, December 1995, pp. 499–503.

[7] A. Guéziec, "Tracking pitches for broadcast television," *IEEE Computer*, vol. 35, no. 3, pp. 38–43, March 2002.

[8] K. Matsui, M. Iwase, M. Agata, T. Tanaka, and N. Ohnishi, "Soccer image sequence computed by a virtual camera," in *CVPR*, Santa Barbara, CA, June 1998, pp. 860–865.

[9] A. Ekin and A.M. Tekalp, "Shot type classification by dominant color for sports video segmentation and summarization," in *ICASSP*, April 2003, vol. 3, pp. 173–176.

[10] J.A. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," TR-97-021, Berkeley, April 1998.

[11] S.Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *ECCV*, Copenhagen, Denmark, May 2002.