Accurate Silhouette Extraction of Multiple Moving Objects for Free Viewpoint Sports Video Synthesis

Qiang Yao, Hiroshi Sankoh, Houari Sabirin, Sei Naito

KDDI R&D Laboratories, Inc.
Ohara 2-1-15, Fujimino, Saitama, Japan
{qi-yao, sankoh, ho-sabirin, sei}@kddilabs.jp

Abstract—In this paper, we propose a new method of automatic silhouette extraction of multiple moving objects with high accuracy for free viewpoint stadium sports video synthesis. The proposed method is basically composed of three parts, including a global extraction based on temporal background subtraction, a classification step based on the constraints of extracted candidates of objects, and local refinement based on the statistical information of the chrominance component of each extracted object. Experimental results show that the proposal outperforms the temporal background subtraction model and Gaussian Mixture Model (GMM) in terms of both objective and subjective evaluations. In addition, the quality of the synthesized free viewpoint sports video is also enhanced by adopting more accurate silhouettes of objects that are extracted by our proposed method. Furthermore, as there is no manual operation in the proposed method, the automatic multiple silhouettes extraction has also been fully realized.

I. INTRODUCTION

In recent years, the appearance of Free Viewpoint Video (FVV) has gained increased popularity since it can provide a beyond-3D experience, where the virtual viewpoints can be selected freely and moved around, back and forth as well as up and down. [1] The system brings an immersive and ultra realistic feeling to the audience and this experience is called “walk-through” and “fly-through”. [2]

Regarding the application of the FVV system, the scenes of dynamic sports games, such as soccer, tennis, and baseball, are very attractive and suitable because the system enables people to select different viewpoints actively to enhance the interest and pleasure in watching games. Basically, there are three types of methods, depth-based, image-based, and 3D model-based, in generating the free viewpoint video. Both the depth-based and image-based method require dense arrangement of camera array, and the viewpoint selection is limited. In contrast, for the 3D model-based method, there is no restriction in the selection of virtual viewpoints in a 3D space.

In the previous work [3], the 3D model based method was adopted and developed for generation of the FVV system. Furthermore, a simplified 3D model, also named the billboard model [4] has also been adopted in [5]. Typically the billboard model is constructed by the shape of silhouettes of objects that are extracted from camera images. For instance, each object, such as a player in a sports game, is represented as a billboard, and the visual texture acquired from multi-view cameras is mapped to the virtual viewpoint selected by the viewers. It is noted that the accuracy of the billboard model has a great impact on the quality of video generation, and inaccurate silhouette of the object, such as the extracted silhouette attached with a shadow area, will ruin the synthesis quality of free viewpoint video. Therefore, the accurate extraction of objects is very significant in the generation of free viewpoint sports video with high quality. In addition, for the practical real-time application, an automatic extraction method is also necessary.

II. RELATED WORKS

Object extraction is a traditional research topic in the field of computer vision, and it is of great significance in numerous applications of computer vision, such as traffic detection, human tracking, pattern recognition, and object segmentation. In this paper, especially for the application of synthesizing free viewpoint sports video, we focus on the accurate silhouette extraction of multiple moving objects.

Basically, to extract moving objects in a single camera with static background, the background subtraction method [6] is simple and effective. One typical method is to take several consecutive frames of one video as one Gaussian Model, and to set a threshold value to extract moving objects. This method is called temporal background subtraction. However, this method relies greatly on the learned background model and the threshold value. Because the threshold value is very sensitive and shadows are also moving with objects, many undesired areas, such as shadows, are also extracted with objects. Therefore, simple temporal background subtraction cannot obtain clear silhouettes of the moving objects. In addition, the Gaussian Mixture Model (GMM), proposed in [7] and modified in [8] is also proved to be effective in object extraction. The method is also based on a learning process, and each pixel in the image gets one Gaussian mixture model. However, as the pixels of an object are similar to the pixels of background, those pixels might also be regarded as background and discarded in object extraction, resulting in many missing areas in the extracted objects. In recent years, other methods have also been proposed for object extraction. The method in [9] tries to detect a shadow area for obtaining a precise extracted object. However, this work only serves
for a single object. As there are multiple objects, the method does not work well in local areas. In addition, the method in [10] integrates several saliences together for automatic object extraction. However, the method does not consider shadow removal. Moreover, the authors in [5] proposed one method based on 3D model projections with the likelihood of removing the shadow areas. However, this method only works in a densely-arranged multiple camera system, but is not suitable for the sparsely-arranged multiple camera system nor single camera system. The method in [11] is able to extract a precise object based on the optimization procedure. However, this method mainly serves for foreground extraction in a still image, and it requires manual operation to assign initial points to indicate foreground and background before extraction. Thus, a full automatic object extraction cannot be realized.

Therefore, according to the related works mentioned above, how to simultaneously and automatically remove the shadow areas attached with silhouettes and restore the missing areas inside each silhouette after extraction is the core issue for accurate multiple silhouettes extraction. In this paper, based on this core issue, we propose a new method to extract the accurate silhouettes of multiple objects automatically by considering various features of shadows and local statistical chrominance information. The remaining part of this paper is organized as follows. In section three, the proposed method is described in detail. The experimental results are presented in the fourth section to illustrate the effectiveness of the proposal. Finally, a brief conclusion is provided in the last part.

III. PROPOSED METHOD

The flowchart of the proposal is illustrated in Figure 1. Roughly speaking, the procedure is composed of a global extraction part, a classification part and a local refinement part. First of all, in order to reduce the missing areas inside the extracted objects, a low threshold value is set in temporal background subtraction. After the background subtraction, different from other methods for shadow detection and removal, we do not directly detect the shadow area, but we find that there are two types of shadows, independent shadow that is departed from the extracted objects and a dependent one that is attached to the extracted objects. Based on this observation, a classification is proposed to remove the independent shadows by considering the constraints of the shadows’ size, shape and luminance consistency. Finally, a refinement is conducted in the local area to remove the dependent shadows. We observe that the shadow of an object is caused by flood-lights in the stadium and shares similar chrominance with the background. Besides, it is assumed that the color difference between objects (e.g. player’s uniform) and background is at least recognizable by human eyes. Thus, a threshold is calculated based on the histogram of chrominance information in each local bounding box of the object to separate the object and the background. Therefore, the shadows can be removed and the missing area inside the object can be restored. In the following subsections, after roughly describing the temporal background subtraction, we present the proposed method in detail.

A. Global background subtraction for rough extraction

In the global background subtraction, a background model is firstly learned at the pixel level based on several consecutive frames in a video sequence. \( I(t) \) is defined as each frame image at the moment \( t \), and \( T \) is the total number of frames in learning the background model. In addition, \( M \) and \( N \) are defined as the width and height of each frame \( I(t) \). For a color image \( I(t) \), it is further decomposed into three color components, such as RGB or YUV, and we use the YUV format in this paper for decomposing the image \( I(t) \) without losing generality, written as \( I(t) = \{ I^y(t), I^u(t), I^v(t) \} \). Next, each pixel \( I^y_{i,j} \) along the period of \( T \), \( \{ I^y_{i,j}(1), I^y_{i,j}(2), ..., I^y_{i,j}(T) \} \) is approximated by one Gaussian model, and the pixels in one frame are independent of each other.  

There is one estimated mean \( \mu^y_{i,j} \) and one estimated standard deviation \( \sigma^y_{i,j} \) along the learning period \( T \) for each pixel \( I^y_{i,j} \) in the background model, written as

\[
\mu^y_{i,j} = \frac{1}{T} \sum_{t=1}^{T} I^y_{i,j}(t) \tag{1}
\]

and

\[
\sigma^y_{i,j} = \left\{ \frac{1}{T} \sum_{t=1}^{T} (I^y_{i,j}(t) - \mu^y_{i,j}) \right\}^{1/2}, \tag{2}
\]

where \( 1 \leq i \leq M, 1 \leq j \leq N \).

\[1\] Please note that the assumption and procedure to process the remaining two image components, \( U \) and \( V \), are similar to those used to process the \( Y \) component.
Based on the learned background model, the residual of each frame after background subtraction is represented as \( R(t) = | I(t) - \mu | \). In addition, there is a threshold \( th \) for enhancing the robustness against the change of luminance. If the absolute residual value \( R_{o,i,j}(t) \) of the residual image is no larger than \( \sigma_{o,i,j} + th \), written as \( R_{o,i,j}(t) \leq \sigma_{o,i,j} + th \), this pixel is regarded as background. (Please note that the method is applied in the YUV space in this paper, and there are \( \mu = \{\mu^y, \mu^u, \mu^v\} \), \( \sigma = \{\sigma^y, \sigma^u, \sigma^v\} \), and \( th = \{th^y, th^u, th^v\} \). However, this method is also suitable for the image in other color spaces or gray scale image.) After the background subtraction, a median filter is adopted for denoising. However, this method is also suitable for the image in other color spaces or gray scale image.) After the background subtraction, a median filter is adopted for denoising. (There will be further refinement inside each bounding box, which will be discussed in the next subsection.) After the assignment of the bounding box, all the extracted candidates are grouped into one set \( \Omega = \{o^k\}, k \in \{1, 2, \ldots, K\} \), where \( K \) is the total number in the set \( \Omega \), as shown in Figure 2.

In the next part, the size, shape, and luminance consistency information of the silhouette candidate are utilized to distinguish the silhouette candidates \( \Omega_{bg} \) and background candidates \( \Omega_{bg} \) in the set \( \Omega \). The classification is composed of three main steps as shown in Figure 3. In the first step, the size information is utilized. All the \( o^k \) in the \( \Omega \) are sorted according to the size of \( o^k \), which is measured by the total number of pixels in each \( o^k \), and then the largest \( \alpha K \), \( 0 < \alpha < 1 \), candidates are preserved in \( \Omega_{obj} \) because the silhouette candidates are not supposed to be too small, such as “obj 1” and “obj 2” as shown in Figure 2. Moreover, the remaining candidates are classified into \( \Omega_{bg} \) as background. (For example, the “obj 5” in Figure 2 will be classified into \( \Omega_{bg} \).)

In the second step, the shape information is utilized. Because our target object is a player and the shadow is always projected on the ground from one or two sides of each object, the aspect ratio \( \beta \) between the width \( w^k \) and the height \( h^k \) of the \( o^k \) is supposed to be in a reasonable range. Thus, if the aspect ratio \( \beta = w^k/h^k \) is in the range of \( [\tau_1, \tau_2] \), \( \tau_1 < \beta < \tau_2 \), the \( o^k \) is classified into \( \Omega_{obj} \); otherwise it is classified into \( \Omega_{bg} \). (For example, the “obj 3” in Figure 2 will be classified into \( \Omega_{bg} \).) In the third step, the information of luminance consistency is considered and calculated in terms of pixel variance in each \( o^k \). It is believed that if \( o^k \) only contains background area, the pixel variance in \( o^k \) is small, because the background is assumed to be smooth and textureless. Thus, if \( Var^y(o^k) > \gamma \), the \( o^k \) is classified into \( \Omega_{obj} \); otherwise it is classified into \( \Omega_{bg} \). (For example, the “obj 4” in Figure 2 will be classified into \( \Omega_{bg} \).)

C. Refinement to extract precise silhouette

After the false silhouette candidates are rejected out, the histogram-based thresholding method is adopted to refine each silhouette candidate in \( \Omega_{obj} \); because the local statistical information can provide more robust cues in segmentation and extraction. However, the histogram is not calculated from the luminance component but from the chrominance component, because the shadow and the background share similar chrominance information, as shown in Figure 4, and the histogram of chrominance can clearly distinguish the silhouette of objects from the background in local region \( P^k \) that is obtained from the original image by the position information of each bounding box. One example of a histogram is illustrated in Figure 5. Obviously there is a valley between two peaks in Figure 5 (b), where the larger peak indicates the background.

---

2 The ball might be also discarded in this step. However, we only focus on accurate extraction of player’s silhouette in this paper, and the ball can be extracted by other methods and synthesized in final image.
while the smaller one indicates the silhouette in $P_k$. In contrast, there is not a clear valley in Figure 5 (a) because the luminance variation is not large in each $P_k$. Next, based on the local chrominance histogram, a thresholding value is calculated by adopting the Otsu method [12] to maximize the between-class variance of the silhouette pixels and the background pixels of $P_k$. As the threshold $\rho$ is found, one binary mask $B$ is employed to identify whether the local pixel $p_i$ belongs to the background or silhouette, which is written as

$$
\begin{align*}
    b_i &= 1, \quad \text{if} \quad p_i \geq \rho \\
    b_i &= 0, \quad \text{if} \quad p_i < \rho,
\end{align*}
$$

where $b_i \in B$. Finally, as all the binary masks are collected, the multiple silhouettes of moving objects can be extracted automatically with high accuracy.

### IV. Experimental Results

In this part, the experimental results are illustrated to show the effectiveness of our proposed method. The configuration of experiments is set as follows. Our system is a static dual-camera system, where two cameras are sparsely fixed in two different positions in the baseball court, respectively, as shown in Figure 6. We adopt 4K cameras with the frame rate of 30 fps and the resolution of $3840 \times 2160$. For each camera, the beginning 600 frames are adopted to learn the background model, and the other 300 frames are employed to test the accuracy of silhouette extraction. The threshold ($th^b$, $th^v$, and $th^\gamma$) in temporal background subtraction is set as 5, 5, 5 and 20, 5, 5 for two videos, respectively. In our proposal, the additional parameters are set as $\alpha = 0.8$, $\tau_1 = 1$, $\tau_2 = 2.1$, and $\gamma = 8$ for both sequences in the classification part. For the GMM based method, we directly implement it by using OpenCV with default parameter values.

#### A. The quantitative evaluation of silhouette extraction

First of all, the quantitative evaluation for the accuracy of silhouette extraction is presented. Three metrics, including recall, precision, and F-measure, are adopted for the evaluation, and the metrics are defined [13] as

$$
\begin{align*}
    \text{Recall} &= \frac{TP}{TP + FN}, \\
    \text{Precision} &= \frac{TP}{TP + FP}, \\
    \text{F-Measure} &= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}},
\end{align*}
$$

where $TP$ is the total number of true positive pixels, $FN$ is the total number of false negative pixels, and $FP$ is the total number of false positive pixels. Without losing generality, we regard the shadow areas as false positive ($FP$) and missing areas in extracted silhouettes as false negative ($FN$). There are three methods in the comparison. One is the temporal background subtraction method. One is the GMM based object extraction. The other one is the proposed method. In addition, we also extract the silhouettes of objects by manual operation, and take this as the ground truth.

The quantitative comparison result of the three methods is illustrated in Table I, and the result is calculated in pixel-level. Generally speaking, a higher Recall value corresponds to fewer missing regions, and a higher Precision value reflects fewer undesired regions (shadows), and a higher F-Measure value shows more robust extraction result. The temporal background subtraction method gets the highest Recall value because we set a low thresholding value in background subtraction. However, the Precision of the temporal background subtraction is quite low because the large parts of shadows are also extracted with objects due to the low threshold. As for the GMM method, the Precision value is still not high because the color of the court line is quite different from the background.
TABLE I
QUANTITATIVE EVALUATION OF THREE METHODS FOR SILHOUETTES
EXTRACTION

<table>
<thead>
<tr>
<th>Method</th>
<th>Temporal background subtraction</th>
<th>GMM [8]</th>
<th>Proposal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.9750</td>
<td>0.9385</td>
<td>0.9115</td>
</tr>
<tr>
<td>Precision</td>
<td>0.6074</td>
<td>0.7156</td>
<td>0.9416</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.7484</td>
<td>0.8113</td>
<td>0.9262</td>
</tr>
</tbody>
</table>

Finally, we conduct synthesis to generate free viewpoint sports video to check the improvement of the synthesized view by using a more accurate silhouette generated by our proposal. Due to the copyright of images in a baseball game, please refer to the supplementary material and demo video in the visualization comparison. In the comparison, it is confirmed that the quality of the synthesized view is enhanced by adopting a more accurate silhouette.

V. CONCLUSIONS
In this paper, we proposed a new method of automatic silhouette extraction of multiple moving objects with high accuracy for free viewpoint stadium sports video synthesis. After the background subtraction, the proposed method removed the independent shadow areas by classification and removed the dependent shadow areas by conducting local refinement for each extracted silhouette. Experimental results showed that the proposed method was quite effective in shadow removal and missing area restoration. More significantly, the visual quality of the synthesized image from a virtual viewpoint was also enhanced by the proposed method. In addition, there was no manual operation in the proposed method, so that the automatic multiple silhouettes extraction was fully realized.

REFERENCES

and the court line is also extracted as an object. Regarding our proposal, the Precision value is the highest one, which proves that the shadow areas can be greatly removed. In addition, the proposal achieves the highest F-Measure value, illustrating that the proposed method is the most robust one. Although the proposed method does not achieve the highest Recall value, the missing parts in the extracted objects almost exist along the boundary, which has a minor effect on the final free viewpoint video synthesis. Please also refer to the subjective evaluation part and the supplementary material.

B. The subjective evaluation of silhouette extraction
In the following part, the extracted silhouettes by different methods are compared in Figure 7, 8. (Please refer to the supplementary material for more experimental results to show the robustness of our proposal.) In a detailed comparison (shadow removal) of (d)-(f) in Figure 7, 8, the shadow is attached with the extracted silhouette in (d) that is obtained by the temporal background subtraction method, while the shadow can be eliminated by the GMM based method and the proposed method. However, the GMM method considers global information in model establishment, thus the court line whose color is different from the background color is also regarded as an object. For the proposed method, since the refinement is conducted in a local area, the global information for some undesired part (such as court line) will not be considered and the high accuracy of extraction can be achieved.

Moreover, in the comparison (missing area restoration) of (g)-(i) in Figure 7, 8, there are many missing areas in the extracted silhouettes by the temporal background subtraction method, because this method only considers temporal information and the threshold in global extraction is very sensitive. As an object remains stationary (or minor movement) among several consecutive frames, parts of the object have a high probability of being regarded as background and not being extracted, resulting in the undesired missing areas. Concerning the GMM based method, the faces and arms of players are quite easily regarded as background because the color of faces and arms is similar to the color of background as shown in Figure 4. For the proposed method, based on the more robust local statistical information, the extraction is conducted inside of each bounding box of the silhouette in the original image during the refinement process. Therefore, most parts of the missing areas caused by temporal background subtraction can also be restored.

³Due to the image copyright of baseball games, we have converted the color images to binary images. Please refer to the supplementary material for better visualization.
Fig. 7. The comparison among three methods for silhouettes extraction, where (a)-(c) are extracted masks by three different methods and (d)-(i) are corresponding closeup results for detailed comparison (Central camera).

Fig. 8. The comparison among three methods for silhouettes extraction, where (a)-(c) are extracted masks by three different methods and (d)-(i) are corresponding closeup results for detailed comparison (Side camera).