3D Model Hypotheses for Player Segmentation and Rendering in Free-Viewpoint Soccer Video

Haopeng Li and Markus Flierl
School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm
{*haopeng, mflierl*}*@kth.se*

***Abstract*—This paper presents a player segmentation approach based on 3D model hypotheses for soccer games. We use a hyperplane model for player modeling and a collection of piecewise geometric models for background modeling. To determine the assignment of each pixel in the image plane, we test it with two model hypotheses. We construct a cost function that measures the fitness of model hypotheses for each pixel. To fully utilize the perspective diversity of the multiview imagery, we propose a three-step strategy to choose the best model for each pixel. The experimental results show that our segmentation approach based on 3D model hypotheses outperforms conventional temporal median and graph cut methods for both subjective and objective evaluation.**

***Keywords*- Soccer video; segmentation; 3D geometric model.**

I. INTRODUCTION

The development of user-defined soccer video has raised the interest in content-adaptive coding and rendering techniques [1][2], which allow users to access content items freely and efficiently [3][4]. Therefore, a fast and reliable player segmentation method is desirable which distinguishes the players from the background for feasible coding, rendering and tracking purposes [5][6].

Conventional background substraction methods mainly utilize spectral (color or texture) [7][8], spatial or temporal [9][10] features to separate dynamic contents from background content. To incorporate one or several features, several statistical models have been developed, such as Gaussian mixture models [11][12] and Bayesian frameworks [13]. However, due to the high complexity of the statistical models, the computational complexity increases rapidly.

We assume that future soccer events are captured by an array of fixed high-definition cameras which provide multiview image sequences for a free-viewpoint experience. The multiview scenario allows us to efficiently exploit the interview correlation for our player segmentation purpose. Therefore, 3D models are desirable that represents the players. Existing methods for player modeling can be classified into two classes. The first class uses 3D grids of voxels to implement a full 3D representation of the player, referred to as *voxel-based* modeling [14]. The rendering quality of this method is usually good. However, the computational complexity of this method is very high and challenging for real-time applications. The second class uses a billboard (planar) model to approximate the player, referred to as *planar-based* modeling [5]. This method can be implemented in real-time and is robust to noise, however, with lower rendering quality. But as the players are relatively small in each view, the artifacts caused by the planar model are negligible.

In this paper, we propose a player segmentation approach based on 3D model hypotheses. In particular, we use a hyperplane model for player modeling and a collection of piecewise geometric models for the background modeling. For an arbitrary pixel in an image plane, we assign it to different 3D model hypotheses. Moreover, we construct a cost function to measure the fitness of each hypothesis. By minimizing the cost function, conditional model hypotheses are determined, given individual references. Due to our multiview system, we obtain a consistent result from multiple references, leading to the final selection of the best 3D model hypothesis. With the resulting silhouetted hyperplane model, the player segmentation problem is solved.

II. ALGORITHM

A. Definition of Model Hypotheses

In this subsection, we define 3D model hypotheses for player segmentation in the multiview scenario. The soccer video scenes are divided into static and dynamic content items. Model hypotheses are defined by a 3D model which characterizes sufficiently the geometry of each content item.

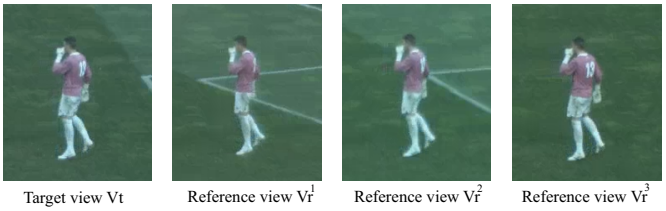

Figure 1. Target view and reference views as defined by rectangular regions.

1) Multiview Image Data: With the multiview player tracking information [15], we are able to extract a rect-

angular region for each player in different views, where each region defines a dynamic content item from a different perspective. A four-view example is depicted in Fig. 1. For our algorithm, we define V_t as the target view to be segmented and $V_r^k, \quad k = 1, 2, \ldots, N$, as the reference views.

2) Hyperplane Model for Dynamic Content: Since the dynamic part (i.e. the players) are relatively small in each view, we assume a rigid body for each player. Therefore, we model the player as a hyperplane P in 3D space which is perpendicular to the plane at $Z = 0$

$$P \subset \mathbb{R}^3, P \perp \{(X, Y, Z) | Z = 0\}. \tag{1}$$

In our earlier work [15], we utilize 3D SIFT features [16] to track the players in multiview scenarios. Now, to estimate the geometric parameters of the hyperplane, we reuse efficiently the accurate 3D features. We define a set Q for the 3D SIFT features of a player. As we assume that the model plane is perpendicular to the plane at $Z = 0$, the hyperplane model is specified by fitting the $[X, Y]$ components of the 3D features with the hyperplane function

$$\min \sum_{X(i), Y(i) \in Q} \|Y(i) - (A \cdot X(i) + B)\|_2, \tag{2}$$

where A, B are parameters of the model. They are determined by the least square error solution and will specify the model of the dynamic part. An example of this model is shown in Fig. 2.

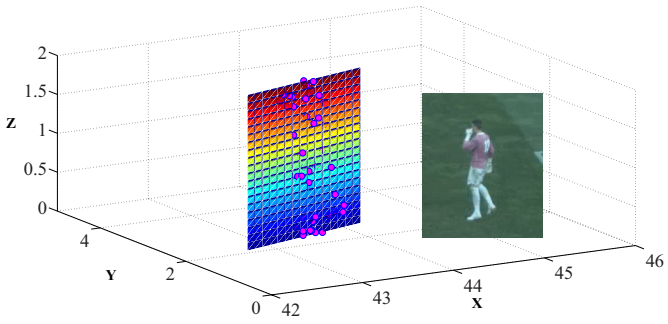

Figure 2. Hyperplane model of the dynamic part; red circles indicate 3D SIFT features, which are located on both sides of the hyperplane.

3) Geometric Model for Background Content: The background content captured by an array of static cameras in a soccer stadium, comprising mostly of areas depicting the field and the background objects, is varying slowly over time. Therefore, piecewise geometric models are often used to model such man-made urban objects by approximating them with piecewise structures.

In our earlier work [17], we model the background contents as a collection of piecewise geometric models. More precisely, the goal is characterized by an assembly of 3D planar models; the grandstand is modeled by a non-planar model and the soccer field is modeled by a cylindrical surface. By projecting the 3D geometric model onto the camera plane, the depth image of the background content is obtained, as depicted in Fig. 3.

Figure 3. Depth image of background content from geometric model.

4) Model Hypotheses: To address the segmentation problem in the target view V_t, we let the set of pixels in the dynamic part be denoted by $\mathcal{D}$, the set of pixels of the background part by $\mathcal{B}$, with $\mathcal{D} \cup \mathcal{B} = V_t$ and $\mathcal{D} \cap \mathcal{B} = \emptyset$.

For each pixel $p_t(x, y)$ with the image coordinate (x, y) in the view V_t, we denote its corresponding 3D point by $P_t(X, Y, Z)$. Therefore, the segmentation of the player in the image plane V_t translates essentially to the calculation of the set of the 3D points $P_t(X, Y, Z)$ in $\mathbb{R}^3$ space. In particular, for $\forall p_t(x, y) \in \mathcal{D}$, the corresponding 3D points $P_t(X, Y, Z)$ will result from the hyperplane model of the dynamic content. On the other hand, for $\forall p_t(x, y) \in \mathcal{B}$, the corresponding 3D points $P_t(X, Y, Z)$ will result from the geometric model of the background content.

Therefore, for each $p_t(x, y)$ in V_t, we define two hypotheses which are denoted by H_0 and H_1:

$$H_0 \quad : \quad p_t(x, y) \in \mathcal{D}, \tag{3}$$

$$H_1 \quad : \quad p_t(x, y) \in \mathcal{B}. \tag{4}$$

Note, this two-hypothesis scenario can be efficiently extended to a multi-hypothesis scenario.

B. Measurement of Model Hypotheses

In this subsection, we construct a cost function that measures the fitness of model hypotheses for each pixel. In addition, the potential occlusion problem related to 3D projection is addressed. Moreover, we propose a method to refine the results of projection localization by considering possible artifacts caused by the geometric models.

1) Cost function: With each model hypothesis and the corresponding geometric information, the pixel $p_t(x, y)$ is mapped to its 3D position $P_t(X, Y, Z)$ in world coordinates. This 3D point is then projected onto the reference image plane with the image coordinate $(\tilde{x}, \tilde{y})$. Generally speaking, by assigning a correct model hypothesis, the difference of intensities between $V_t(x, y)$ and $V_r(\tilde{x}, \tilde{y})$ is relatively small when compared to assigning a wrong model hypothesis.

Considering the noise and illumination conditions in target and reference views, we use a m-by-m Gaussian window W_m for $V_t(x, y)$ and $V_r(\tilde{x}, \tilde{y})$ when calculating the difference. Therefore, we construct a cost function J which

measures the difference between target and hypothesis-dependent reference as a Gaussian weighted sum of squared differences

$$J^k \quad (h,c) = \sum_{(i,j)\in W_m} G(i,j) \cdot \|V_t(i+x, j+y, c) - \quad V_r(i+\tilde{x}_h^k, j+\tilde{y}_h^k, c)\|^2, h \in \{H_0, H_1\}, \tag{5}$$

where G is the Gaussian coefficient window, k is the index of the reference, h indicates the assigned hypothesis, and c indicates the color space. We define both target and reference images with four color components $\mathcal{C} = \{R, G, B, Y\}$, where the Y component is the illumination.

There are two main advantages of choosing $\mathcal{C} = \{R, G, B, Y\}$ spaces. First, the measurements produced by different components are more robust to noise. Second, since the Y component affects the objective quality (Y-PSNR) and the RGB components affect the subjective quality (color experience), the joint measurement will allow a balance between objective and subjective results.

2) Projection Localization: To warp the pixel $p_t(x, y)$ from the target view V_t to the reference view V_r^k, we project the 3D model (dynamic or background) onto the image plane of V_t and obtain the depth image I_t. Knowing the camera calibration parameters, we are able to project $p_t(x, y)$ into 3D world coordinates according to its depth value as specified by the associated depth image I_t, resulting in a 3D point $P_t(X, Y, Z)$. Then, this 3D point is projected onto the image plane of the reference view. In particular, let the warped pixel in V_r^k be denoted by $p_r(x_0^k, y_0^k)$ for H_0 and $p_r(x_1^k, y_1^k)$ for H_1.

3) Background Occlusion Detection: Before calculating the cost function for each hypothesis, we should detect potential occlusion problems due to the interference between background and dynamic model. In other words, since the dynamic part always overrides the background part, it is possible that the background hypothesis $p_r(x_1^k, y_1^k)$ is occluded by the dynamic part.

One example is shown in Fig. 4. Here, we warp the pixel $p_t(x, y)$ to $p_r(x_1^k, y_1^k)$ by assuming background hypothesis H_1. Obviously, H_1 is the correct hypothesis for $p_t(x, y)$ since it is a point in the background. However, the background pixel $p_r(x_1^k, y_1^k)$ in V_r^k is covered by the dynamic part due to occlusion. Note, this scenario might occur for both dynamic and background pixels.

To solve this problem, we propose a recursive approach to determine if the pixel $p_r(x_1^k, y_1^k)$ has been occluded. Initially, we assign the background hypothesis H_1 and obtain the pixel $p_r(x_1^k, y_1^k)$. Contrary to the initial assignment, we assume that the pixel $p_r(x_1^k, y_1^k)$ belongs to the dynamic part. To find the corresponding depth value, we project the hyperplane model P onto the image plane of V_r^k and obtain the depth image I_r^k. With the associated depth value $I_r^k(x_1^k, y_1^k)$, we are able to warp the pixel $p_r(x_1^k, y_1^k)$ to the image plane

Target view Vt Reference view Vr

Figure 4. Background hypothesis occluded by dynamic part. The red point indicates the target pixel $p_t(x, y)$. The blue point indicates the warped point $p_r(x_1^k, y_1^k)$ in the reference view when assigning the background hypothesis H_1.

of the target view V_t, as denoted by $p_t(x_1^{'k}, y_1^{'k})$. There are three possible outcomes:

- The pixel $p_t(x_1^{'k}, y_1^{'k})$ in the target view has been determined to be a pixel of the dynamic part $\mathcal{D}$. Therefore, the assumption that $p_r(x_1^k, y_1^k)$ is a pixel of the dynamic part is correct. This implies the occlusion of the background hypothesis H_1, as shown in Fig. 4.
- The pixel $p_t(x_1^{'k}, y_1^{'k})$ in the target view has been determined to be a pixel of the background part $\mathcal{B}$. Therefore, the assumption is wrong. In practice, if the projection result shows that $(x_1^{'k}, y_1^{'k}) \notin V_t$, the assumed depth value is wrong. In other words, the pixel $p_r(x_1^k, y_1^k)$ is not occluded in this reference view and, hence, the assigned background hypothesis H_1 is possible.
- The pixel $p_t(x_1^{'k}, y_1^{'k})$ in the target view has not been determined yet. We will skip pixel $p_t(x, y)$ until $p_t(x_1^{'k}, y_1^{'k})$ has been determined in a future iteration.

Hence, our algorithm needs several iterations until all pixels have been determined. The reference views, in which the projection $p_r(x_1^k, y_1^k)$ is occluded, will simply be discarded (as the case in Fig. 4).

4) Accurate Projection Localization: As the planar and non-planar models are used to characterize the geometry of dynamic and background content items, we will allow some local adjustments to the depth image in order to make the projection localization more accurate.

To implement a local adjustment, we allow a small depth adjustment Δd for each $d_t(x, y)$, leading to slightly different projection results $(\tilde{x^k}, \tilde{y^k})$ for k-th reference. We find the optimal adjustment of Δd that satisfies

$$\min_{\Delta d} \sum_{k=1}^{N} \sum_{c\subset\mathcal{C}} \|V_t(x, y, c) - V_r(\tilde{x^k}(\Delta d), \tilde{y^k}(\Delta d), c)\|^2. \tag{6}$$

This results in the final projection localization $(\tilde{x^k}, \tilde{y^k})$ which is more accurate and, hence, is used in (5) to compute

the cost function.

C. Hypothesis Selection

In this subsection, we propose a three-step strategy to choose between hypothesis H_0 and H_1 for each pixel. First, for each individual reference, make a decision based on the cost function for each pixel. Second, compute the acceptance rate for multiple references. Finally, aggregate the acceptance rate to obtain the best model.

1) Hypothesis Decision Rule: To make the decision for each hypothesis, we exploit the cost function (5) in Section II-B1. We compute the cost value $J^k(H_0, c)$ and $J^k(H_1, c)$ for each pixel with $c \in \{R, G, B, Y\}$. Moreover, due to the fact that the color of the background is very similar, it is very likely that once $p_t(x, y)$ is a background pixel, both of its projections $p_r(x_0^k, y_0^k)$ and $p_r(x_1^k, y_1^k)$ are localized in the background. In this case, the cost value of $J^k(H_0, c)$ and $J^k(H_1, c)$ will be very close. Hence, we also consider the absolute difference between two hypotheses

$$T_h = |J^k(H_0, c) - J^k(H_1, c)|. \tag{7}$$

Based on the above three measures $J^k(H_0, c)$, $J^k(H_1, c)$ and T_h, we propose the following rule

$$p_t(x, y) \in \begin{cases} \mathcal{D}, & \text{if } T_h > \delta_d \text{ and } J^k(H_0, c) < J^k(H_1, c) \\ \mathcal{B}, & \text{if } T_h > \delta_d \text{ and } J^k(H_0, c) > J^k(H_1, c) \\ \mathcal{B}, & \text{if } T_h < \delta_d, \end{cases} \tag{8}$$

where δ_d is a small threshold to measure the significant difference between two hypotheses. Note that we investigate all $\{R, G, B, Y\}$ components and produce four decisions for each pixel $p_t(x, y)$ in the reference k.

2) Computation of Acceptance Rate: Due to the large baseline of the multiview soccer video, the dynamic content is usually very noisy, in particular, the contour of the players. On the other hand, by utilizing the diversity of multiple references, we are able to compute the acceptance rate for each hypothesis among multiple references. Therefore, the decision for each hypothesis is represented by a probability-like acceptance rate. This is more favorable for noisy environments.

Assume that we have k reliable references (after discarding occluded references). The total number of decisions for pixel $p_t(x, y)$ will be 4 times k (considering $\{R, G, B, Y\}$ components). Let the set of accepted dynamic hypotheses be denoted by $\mathcal{S}_d$, and the set of accepted background hypotheses by $\mathcal{S}_b$. The acceptance rate P_d for dynamic hypotheses and P_b for background hypotheses are defined by

$$P_d = \frac{|\mathcal{S}_d|}{4k}, \tag{9}$$

$$P_b = \frac{|\mathcal{S}_b|}{4k}, \tag{10}$$

where $|P_d| + |P_b| = 1$. A distribution of the hypothesis acceptance rate is shown in Fig. 5. We can observe that hypothesis selection is still challenging in certain areas.

Dynamic hypothesis **Background hypothesis**

Figure 5. Distribution of hypothesis acceptance rate. Bright areas indicate high acceptance rate.

3) Aggregation of Acceptance Rate: To improve the robustness of hypothesis selection, we should not only consider the acceptance rate of the current pixel, but also that of the neighboring pixels. In particular, for the feet of the player, it is difficult to choose the hypothesis since the depth values of dynamic and background hypotheses are very similar.

Therefore, we aggregate the hypothesis acceptance rate over a Gaussian window and compute the likelihood of H_0 over H_1

$$L(x, y) = \frac{\sum_{(i,j) \in W_m} G(i, j) P_d(i + x, j + y)}{\sum_{(i,j) \in W_m} G(i, j) P_b(i + x, j + y)}, \tag{11}$$

where G is the same Gaussian window as used before in (5) and (6). We choose the dynamic hypothesis for $L > 1$. Otherwise we choose the background hypothesis. Fig. 6 depicts the result of model selection by aggregating the acceptance rate.

Target view Vt **Segmentation**

Figure 6. Segmentation of target view.

III. Experimental Results

We evaluate our method using 3D model hypotheses for the soccer test video set *Barca-St. Andreu*, which is provided by the MEDIAPRO group. The videos are captured by four fixed broadcast cameras. The resolution of the videos is 1920×1080 at 25 fps. With the multiview player tracking information [15], we extract sub-sequence for each player. The resolution of the sub-sequences is 180×200.

We compare our algorithm with two conventional methods which utilize temporal and spectral features to separate dynamic from background content. For our algorithm, we use camera 3 as the target view, while camera 1, 2 and 4 are used as reference views. The first reference algorithm is a temporal median method [18][9] which adaptively computes the temporal median frame from previous frames and compares the difference between current frame and temporal median frame. Once the difference is larger than a threshold, the current pixel is marked as dynamic. Some morphological closing options are implemented to eliminate small regions. The threshold for the first reference algorithm is appropriately adjusted from empirical trials. The second reference algorithm is a texture-based method [8] which minimizes the segmentation energy by using graph cuts.

A. Comparison of Segmentation Results

Fig. 7 shows a comparison of segmentation results for different algorithms. For the temporal median method, the segmentation is sensitive to noise as it uses just previous frames as references. In addition, it also has the risk that dynamic parts that have stopped moving for some time will be wrongly classified as background content. For graph cuts, the white lines on the ground can be easily classified as dynamic parts, as the method considers only texture information. Our method offers better visual quality than the reference algorithms. Using multiview references and hyperplane models, the consistency and smoothness of the segmentation is preserved.

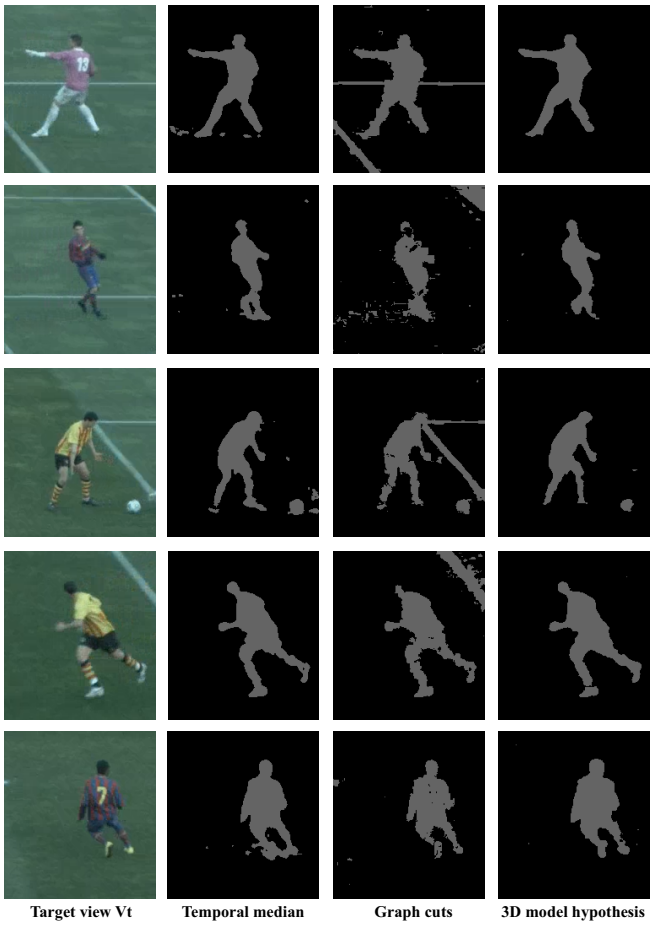

Figure 7. Comparison of segmentation between reference algorithms and proposed algorithm. The first column shows the target image; the second column shows the segmentation results for the temporal median; the third column shows the segmentation results for graph cuts; the fourth column shows the segmentation results of our 3D hypothesis algorithm.

B. Comparison of Rendering Quality

Table I
Comparison of rendering quality.

Sequence index	Temporal median (dB)	Graph cuts (dB)	3D model hypotheses (dB)
01	28.7	27.0	32.9
02	30.3	28.7	34.1
03	26.9	25.8	32.0
04	26.6	24.4	30.7
05	30.7	28.4	34.6

The quality of virtual view rendering is also an important aspect. Therefore, we measure the rendering quality as luminance PSNR (Y-PSNR) for a given camera viewpoint. The texture for the given viewpoint is synthesized by conventional depth image based rendering (DIBR) [19]. In particular, we use camera 3 as the given camera viewpoint while camera views 1, 2 and 4 are used as references. Note, since our implementation uses the Mean-Square-Error (MSE) criterion for optimization, the PSNR is more suitable to measure the objective quality.

The depth images are generated by our 3D models with segmentation mask. For our 3D model hypothesis method, as stated in Section II-B4, the depth image has been slightly adjusted to make the projection localization more accurate. For the reference algorithms, the depth images are generated by the same geometric model, but with different segmentation masks. A total of 5 sequences with 5 different players are used to evaluate the performance of the algorithms. We calculate the average PSNR value over 64 frames for each sequence. We set the size of the Guassian window to 5, the variance to 1, and the threshold to $\delta_d = 2$ for (8).

As shown in Table I, our 3D model hypothesis method outperforms the conventional temporal median and the graph

cuts for all sequences. Since our algorithm optimizes classification by using the cost function (5), our average PSNR is much higher than that of the two reference methods. Note, the camera baseline is 10m for this data set. Hence, the varying lighting conditions are challenging and affect the average Y-PSNR.

Fig. 8 shows a subjective comparison of the rendering quality. Since our algorithm produces more consistent and smoother segmentation masks, the subjective rendering quality is improved significantly.

IV. Conclusions

We discussed a player segmentation and rendering approach based on 3D model hypotheses for soccer video. We use a hyperplane model for player modeling and a collection of piecewise geometric models for background modeling. To determine the classification of each pixel in the image plane, we define two model hypotheses. A cost function is introduced to measure the fitness of model hypotheses for each pixel. Further, we propose a three-step strategy to choose the best hypothesis for each pixel, given multiple reference views. The experimental results show that our 3D model hypothesis method outperforms conventional temporal median and graph cut methods for both subjective and objective evaluation.

V. Acknowledgments

This work was supported in part by the European Commission in the context of the project ICT-FP7-248020 "FINE – Free-Viewpoint Immersive Networked Experience".

References

[1] P. Ndjiki-Nya, T. Hinz, A. Smolic, and T. Wiegand, "A generic and automatic content-based approach for improved H.264/MPEG4-AVC video coding," in *Proc. of the IEEE International Conference on Image Processing*, Sept. 2005.

[2] H. Li and M. Flierl, "Rate-distortion-optimized content-adaptive coding for immersive networked experience of sports events," in *Proc. of the IEEE International Conference on Image Processing*, Sept. 2011.

[3] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image Processing*, vol. 12, no. 7, pp. 796 – 807, July 2003.

[4] V. Tovinkere and R. Qian, "Detecting semantic events in soccer games: towards a complete solution," in *Proc. of the IEEE International Conference on Multimedia & Expo*, Aug. 2001.

[5] T. Koyama, I. Kitahara, and Y. Ohta, "Live mixed-reality 3D video in soccer stadium," in *Proc. of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2003.

[6] D. Delannay, N. Danhier, and C. Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *Distributed Smart Cameras. Third ACM/IEEE International Conference on*, Sept. 2009.

[7] O. Javed, K. Shafique, and M. Shah, "A hierarchical approach to robust background subtraction using color and gradient information," in *Proc. of the Workshop on Motion and Video Computing*, Dec. 2002.

[8] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124 –1137, Sept. 2004.

[9] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1997, pp. 2657–2660.

[10] P. Spagnolo, T. Orazio, M. Leo, and A. Distante, "Moving object segmentation by background subtraction and temporal analysis," *Image and Vision Computing*, vol. 24, no. 5, pp. 411 – 423, 2006.

[11] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proc. of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.

[12] M. Harville, G. Gordon, and J. Woodfill, "Foreground segmentation using adaptive mixture models in color and depth," in *Proc. of the IEEE Workshop on Detection and Recognition of Events in Video*, July 2001.

[13] L. Li, W. Huang, I. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. on Image Processing*, vol. 13, no. 11, pp. 1459 –1472, nov. 2004.

[14] S. Seitz and C. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 1997.

[15] H. Li and M. Flierl, "SIFT-based multi-view cooperative tracking for soccer video," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2012.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.

[17] H. Li and M. Flierl, "SIFT-based modeling and coding of background scenes for multiview soccer video," in *Proc. of the IEEE International Conference on Image Processing*, Sept. 2012.

[18] T. Aach, A. Kaup, and R. Mester, "Statistical model-based change detection in moving video," *Signal Processing*, vol. 31, no. 2, pp. 165–180, 1993.

[19] Y. Mori, N. Fukushima, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 229–232, May 2008.

Figure 8. Comparison of rendering quality between reference algorithms and proposed algorithm. The first column shows the actual image; the second column shows the rendered image by using the mask of temporal median; the third column shows the rendered image by using the mask of graph cuts; the fourth column shows the rendered image as generated by our 3D hypothesis algorithm.