SIFT-BASED MULTI-VIEW COOPERATIVE TRACKING FOR 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 SIFT-based multi-view cooperative tracking scheme for multiple player tracking in soccer games. We assume that future sports events will be captured by an array of fixed high-definition cameras which provide multi-view video sequences. The imagery will then be used to provide a free-viewpoint networked experience. In this work, SIFT features are used to extract the inter-view and inter-frame correlation among related views. Hence, accurate 3D information of each player can be efficiently utilized for real time multiple player tracking. By sharing the 3D information with all cameras and exploiting the perspective diversity of the multi-camera system, occlusion problems can be solved effectively. The extracted 3D information improves the average reliability of tracking by more than 10% when compared to SIFT-based 2D tracking.

Index Terms— Multiple object tracking, multi-view, SIFT.

1. INTRODUCTION

The rise of high-definition television (HDTV) and the desire for user-defined soccer video have raised the interest in content-adaptive coding techniques [1], which allow users to access content items freely and semantically meaningfully. Thus, an accurate and reliable multiple player tracking system is desirable which allows us to extract a rectangular region for each player, where each region defines a dynamic content item and results in an individual image sequence.

Prevalent particle filter work well for isolated object tracking by estimating the non-Gaussian and non-linear posterior probability distribution of the object [2]. However, it is ambiguous to distinguish objects when tracking is applied to multiple objects with similar behaviors or in situations of congestion and confusion. The mean-shift method is another efficient approach to find the optimal matching region between frames [3]. However, manually labeled regions of interest are required for initialization or entering objects. Additionally, due to region-based properties, tracking is burdened by occlusions.

In this paper, we propose a Scale Invariant Feature Transform (SIFT) based multi-view cooperative tracking scheme for real time multiple player tracking in soccer video. SIFT has been widely used in object recognition, robotic mapping, video tracking, and match moving [4]. The primary advantage of SIFT is its invariance under rotation, scale change and affine transformation. It performs better than conventional feature matching techniques like ordinary correlation and Harris corners which are not invariant under rotation and changes in scale.

Moreover, we assume that future soccer events are captured by an array of fixed high-definition cameras which provide multi-view image sequences for a free-viewpoint experience in a home environment. We fully exploit the inter-view and temporal correlation of multi-view video by matching SIFT features in both view imagery and temporal frames. Different from the conventional 2D SIFT-based tracking approach [5], the uniquely defined 3D position of each feature is extracted by utilizing multi-view geometry constraints. In other words, the 3D coordinates of moving points are obtained, hence, accurate 3D positions of players are used for tracking.

By knowing the 3D position information of players, player-oriented tracking in local regions can be realized. Compared to the joint-state-oriented multi-view tracking scheme [6] whose computational complexity increases exponentially in terms of the number of players and cameras, the computational complexity of our proposed scheme increases linearly with the number of players and cameras. These benefits ensure that our scheme can be parallelized and implemented in real-time.

2. MULTI-VIEW TRACKING SCHEME

To facilitate real time multiple player tracking for soccer events, we discuss a SIFT-based multi-view cooperative tracking scheme. To match the properties of multi-view video captured by an array of static cameras in a soccer stadium, we exploit the inter-view and temporal correlation among related views. The inter-view correlation between adjacent views, defining the same object captured by different views instantaneously, can be exploited efficiently by inter-view geometric constraints. On the other hand, the temporal correlation between successive frames, defining the motion of features captured by the same view over time, can be exploited efficiently by inter-frame feature matching.

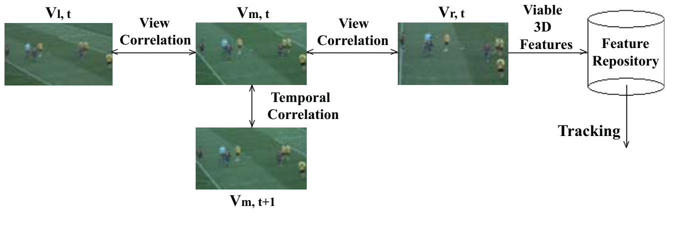

Fig. 1. SIFT-based tracking scheme.

The SIFT-based multi-view cooperative tracking scheme comprises the SIFT feature matching between spatial and temporal views, the feature repository, and a tracking unit that realizes the real time implementation. Fig. 1 depicts the discussed SIFT-based multi-view cooperative tracking scheme.

2.1. SIFT Feature Matching and Refinement

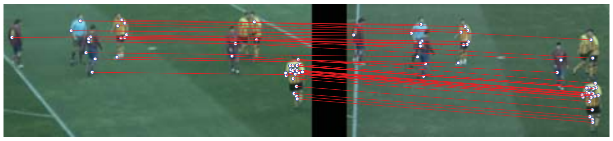

Fig. 2. SIFT features in left and right reference images.

First, we define a view group which includes both spatially correlated views $V_{l,t}$, $V_{r,t}$ and temporally correlated views $V_{m,t}$, $V_{m,t+1}$ with respect to time t, as depicted in Fig. 1. We extract SIFT features in views $V_{m,t}$ and $V_{m,t+1}$, and find correct correspondences. Let $p_i^{m,t} \leftrightarrow p_j^{m,t+1}$ be a temporal feature correspondence, where $p_i^{m,t}$ denotes the i-th feature point with the image coordinate $(x_i^{m,t}, y_i^{m,t})$ in the view $V_{m,t}$ and $p_j^{m,t+1}$ the j-th feature point with the image coordinate $(x_j^{m,t+1}, y_j^{m,t+1})$ in the view $V_{m,t+1}$. As we assume that the sports events are captured by an array of fixed high-definition cameras, the static scene can be generated by traditional temporal median methods [7]. Thus, feature correspondences which belong to the static scene can be easily filtered. In that scene, the remaining correspondences are related to the moving items between two successive frames, hence, they can be used to obtain tracking information. We define two sets of moving features for views $V_{m,t}$ and $V_{m,t+1}$

$$\begin{aligned} F^{m,t} &= \{p_i^{m,t} | p_i^{m,t} \leftrightarrow p_j^{m,t+1}\} , \\ F^{m,t+1} &= \{p_j^{m,t+1} | p_i^{m,t} \leftrightarrow p_j^{m,t+1}\} . \end{aligned} \quad (1)$$

Next, we extract SIFT features in views $V_{l,t}$ and $V_{r,t}$, and find correct correspondences of $p^{m,t} \leftrightarrow p^{r,t}$ and $p^{m,t} \leftrightarrow p^{l,t}$. Correct correspondences in adjacent views relate to the same 3D point in the scene and, hence, can be used to obtain reliable 3D information. An example is depicted in Fig. 2.

Let $p_i^r \leftrightarrow p_j^m$ be a feature correspondence between views $V_{r,t}$ and $V_{m,t}$. Knowing the camera calibration parameters, we are able to check the correctness of the feature match based on a geometric constraint. If it is a correct correspondence, p_i^r and p_j^m are originally projected from the same 3D world coordinate. Thus, by using projection relations for corresponding points, the following relations hold:

$$R_r^{-1} \cdot A_r^{-1} \cdot [x_i^r, y_i^r, 1]^T \cdot \lambda_r + \begin{pmatrix} C_r^x \\ C_r^y \\ C_r^z \end{pmatrix} = \begin{pmatrix} X_i^r \\ Y_i^r \\ Z_i^r \end{pmatrix}, \quad (2)$$

$$R_m^{-1} \cdot A_m^{-1} \cdot [x_j^m, y_j^m, 1]^T \cdot \lambda_m + \begin{pmatrix} C_m^x \\ C_m^y \\ C_m^z \end{pmatrix} = \begin{pmatrix} X_j^m \\ Y_j^m \\ Z_j^m \end{pmatrix}, \quad (3)$$

where $[X, Y, Z]^T$ is the 3D world coordinate, and where R, A and C are the camera calibration parameters which depend on the camera position. The factors λ_r and λ_m in (2) and (3) define the position of the 3D point on the rays. To determine the scaling factors, let the third row of the 3×3 matrix $R_r^{-1} \cdot A_r^{-1}$ be $[\alpha_r, \beta_r, \gamma_r]$, and let the third row of the 3×3 matrix $R_m^{-1} \cdot A_m^{-1}$ be $[\alpha_m, \beta_m, \gamma_m]$. Thus, the factor λ_r is given by

$$\lambda_r(Z_i^r) = \frac{Z_i^r - C_r^z}{\alpha_r x_i^r + \beta_r y_i^r + \gamma_r}. \quad (4)$$

Similarly for the factor λ_m. Therefore, the factors λ_r and λ_m are a function of the world coordinate Z. Note, if we assume that the players move on the plane Z=0, the world coordinate Z relates to the height of players.

With the scaling factors λ_r and λ_m, (2) and (3) need to be equal for corresponding points $p_i^r \leftrightarrow p_j^m$. As we assume to know the true camera calibration parameters, the resulting expression is over-determined. For our practical application, we determine the least square error solution of Z^* according to

$$\begin{aligned} Z^* = \arg\min_Z \| & R_r^{-1} \cdot A_r^{-1} \cdot \begin{pmatrix} x_i^r \\ y_i^r \\ 1 \end{pmatrix} \cdot \lambda_r(Z) + \begin{pmatrix} C_r^x \\ C_r^y \\ C_r^z \end{pmatrix} \\ & - R_m^{-1} \cdot A_m^{-1} \cdot \begin{pmatrix} x_j^m \\ y_j^m \\ 1 \end{pmatrix} \cdot \lambda_m(Z) - \begin{pmatrix} C_m^x \\ C_m^y \\ C_m^z \end{pmatrix} \|_2 . \end{aligned} \quad (5)$$

The two resulting 3D world coordinates $[X, Y, Z]_i^{rT}$ and $[X, Y, Z]_j^{mT}$ are obtained by the least square error solution (5) with respect to p_i^r and p_j^m. However, some small misalignment caused by calibration parameters should also be considered. Thus, we use an additional criterion. If $\|[X, Y, Z]_i^{rT} - [X, Y, Z]_j^{mT}\|_2 < \delta_d$, where δ_d is a small threshold for the Euclidean distance in 3D space, the correctness of the correspondence $p_i^r \leftrightarrow p_j^m$ is sufficiently reliable. Thus, we define a set of inter-view features for view $V_{r,t}$

$$F^{r,t} = \{p_i^{r,t} | p_i^{r,t} \leftrightarrow p_j^{m,t}\} . \quad (6)$$

Similarly, for the feature correspondence $p_i^l \leftrightarrow p_j^m$ between views $V_{l,t}$ and $V_{m,t}$, we define a set of inter-view features for view $V_{l,t}$

$$F^{l,t} = \{p_i^{l,t} | p_i^{l,t} \leftrightarrow p_j^{m,t}\} . \quad (7)$$

To obtain reliably the 3D information of the moving items in view $V_{m,t}$ and handle the possible occlusion problem in the adjacent views, we choose inter-view features in both adjacent views $V_{l,t}$ and $V_{r,t}$

$$F_s^{m,t} = F^{m,t} \cap (F^{r,t} \cup F^{l,t}). \quad (8)$$

Note, regarding the complexity of the algorithm, global SIFT matching is applied for the initialization of the algorithm. Thereafter, local SIFT matching will be used. Details are given in Section 2.3.

2.2. Feature Clustering by Using 3D Information

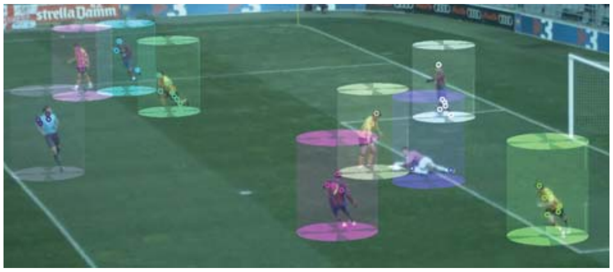

Fig. 3. Clustering of 3D features; each player is identified by one cylinder, the features are shown by small circles; the same color indicates the same cluster.

After extracting features from dynamic items, the 3D information can be exploited to identify players by clustering the features.

Conventional methods for object classification usually operate on the 2D image plane. However, due to the lack of 3D distance information, the accuracy of classification is low. In our work, we use 3D coordinates $[X_i, Y_i, Z_i]$ to describe the feature points $p_i^{m,t}$ in 3D space and use k-means [8] to cluster them:

$$\begin{aligned} \min \quad & \sum_{i=1}^{K} \sum_{p_j^{(i)} \in S_i} \|p_j^{(i)} - C_i\|_2 \\ s.t. \quad & F_s^{m,t} = \bigcup_{i}^{K} S_i, \end{aligned} \tag{9}$$

where S_i is the i-th set of features and C_i the cluster center. Note, we use the cluster center C_i to define the position of player i. Let the number of players be denoted by K. $\|p_j^{(i)} - C_i\|_2$ is the 3D distance between feature point $p_j^{(i)}$ and the cluster center C_i. In other words, each player is identified by nearby 3D features, as depicted in Fig. 3.

This approach offers three advantages: First, compared to 2D information, 3D information is more robust. Second, due to the available 3D information on the field, the estimate for the number of clusters is more reliable. Third, the computational cost is lower for object classification.

2.3. Feature Tracking in Consecutive Frames

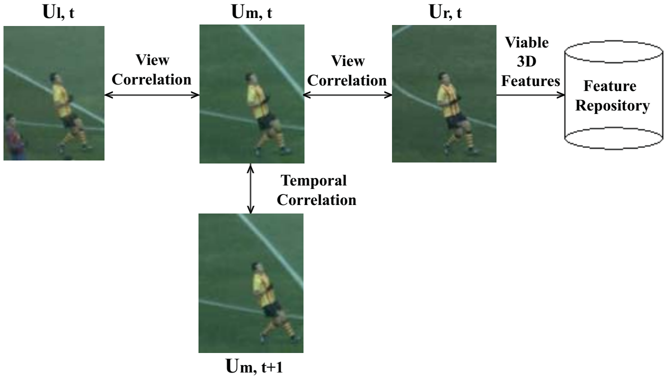

Fig. 4. Matching local SIFT features in the selected regions.

After generating the set of active players $\{C_i\}_1^K$ by clustering the features $F_s^{m,t}$, the tracking of each player can be implemented by a scheme as shown in Fig. 1.

The 3D information of player position $C_i(X, Y, Z)$ allows us to extract a cylindrical region for each player. This 3D world coordinate is projected onto the image plane of $V_{m,t}$, $V_{l,t}$ and $V_{r,t}$, resulting in the corresponding 2D player positions $c_i^{m,t}$, $c_i^{l,t}$ and $c_i^{r,t}$. Note, for view $V_{m,t+1}$, we assign the 2D player position $c_i^{m,t+1}$ also to $c_i^{m,t}$. Then, the 2D player position can be used to define one region U in each view to cover player i. Note, the local 3D features $\tilde{S}_i^t$ within cluster i can be used to extract the corresponding 2D features $\widetilde{p^{m,t}}$ for each player i at time t by using the scheme in Section 2.1, as depicted in Fig. 4.

If there is more than one player in region U, we find for each $\widetilde{p^{m,t}}$ its nearest 3D cluster C_i with respect to 3D distance, and assign $\widetilde{p^{m,t}}$ to the 3D set $\tilde{S}_i^t$. An example is shown in Fig. 5.

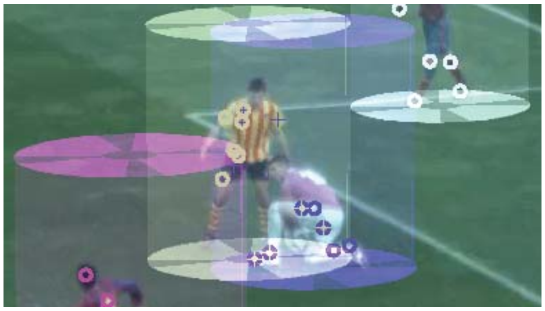

Fig. 5. Feature assignment; each color defines one region U_i, circles indicate features belonging to the 3D set $\tilde{S}_i^t$, crosses indicate features that do not belong to the same set, however have been extracted from region U_i.

There are two advantages of matching SIFT features locally: First, since the computational burden of SIFT is determined by the image size, matching SIFT features only in dynamic parts [1] avoids unnecessary complexity. Second, the selected region U efficiently provides a prior region for feature matching, which leads to more reliable results. Note, the occlusion problem will be addressed in Section 2.4.

Knowing the 2D correspondences $\widetilde{p^{m,t}} \leftrightarrow \widetilde{p^{m,t+1}}$, the motion vector v_i for player i can be determined by calculating the median of motion shifts:

$$v_i = \underset{1 \le j \le |\tilde{S}_i^t|}{\text{median}} \{\widetilde{p_j^{m,t+1}} - \widetilde{p_j^{m,t}}\}. \tag{10}$$

Note, we assume a rigid body for each player since they are relatively small in each view. Hence, the median of motion shifts can be used to represent the motion vector v_i.

With this rigid model, we are able to track the motion between two successive frames. For that, we update the position $\tilde{C}_i^t$ of each player by calculating the new cluster center

$$\tilde{C}_i^t = \frac{1}{|\tilde{S}_i^t|} \sum_{[X_j, Y_j, Z_j] \in \tilde{S}_i^t} \begin{pmatrix} X_j \\ Y_j \\ Z_j \end{pmatrix}. \tag{11}$$

Next, $\tilde{C}_i^t$ is projected onto the image plane of $V_{m,t}$ to obtain its 2D coordinate $\widetilde{c_i^{m,t}}$. Then, the 2D coordinate $\widetilde{c_i^{m,t+1}}$ in $V_{m,t+1}$ can be estimated by motion compensation

$$\widetilde{c_i^{m,t+1}} = \widetilde{c_i^{m,t}} + v_i. \tag{12}$$

Finally, the $\widetilde{c_i^{m,t+1}}$ can be back-projected into 3D space to get the player position information $\widetilde{C_i^{m,t+1}}$ in the frame $t+1$. Moreover, it can also be warped into adjacent views.

2.4. Addressing Occlusion Problems

With our multi-view cooperative tracking scheme, we are able to handle efficiently both partial and total occlusion by exploiting the perspective diversity of the multi-camera system.

Let us consider the following example: Player i is visible in $U_{m,t}$, however, it is totally occluded by other objects (i.e. other players) in $U_{m,t+1}$. Obviously, feature matching between $U_{m,t}$ and

Table 1. Comparison of the reliability of tracking schemes.

Sequence index	Player-number	Frame-number	Reliability of 2D tracking	Reliability of 3D tracking	Reliability of 3D tracking (detection error excluded)
01	1	480	100%	100%	100%
02	> 9	180	88.8%	99.2%	99.9%
03	> 11	480	86.5%	95.4%	97.5%
04	> 13	360	84.2%	98.1%	99.7%

$U_{m,t+1}$ for player i will fail, which leads to an empty 3D feature set $\widetilde{S}_i^t$. However, due to the multi-camera arrangement which captures views from different perspectives, player i is probably visible in another view, i.e., $V_{x,t}$. Since the 3D position $\widetilde{C}_i^t$ of player i is shared by all views, new local view groups $U_{x,t}$, $U_{x,t+1}$, $U_{r',t}$, $U_{l',t}$ containing player i with the structure as shown in Fig. 4 can be easily defined by calculating the projection of $\widetilde{C}_i^t$. Thus, our SIFT matching method in Section 2.3 can be reused to obtain the 3D feature set $\widetilde{S}_i^t$ for player i. This improves the reliability of tracking of occluded objects.

2.5. Detection of Appearing Objects

As we assume that soccer games are captured by arrays of fixed high-definition cameras, player appearance in $V_{m,t}$ is usually captured in other views too. Thus, our scheme in Section 2.1 with associated sets of tracked players $\{\widetilde{C_i}\}_1^K$ can be used to detect new players.

Recall that we extract the set $F_s^{m,t}$ of 3D information for the moving items in view $V_{m,t}$. For each feature in $F_s^{m,t}$, we measure its minimum 3D distance to the sets of tracked players $\{\widetilde{C_i}\}_1^K$. If the distance is larger than δ, where δ is a threshold for the Euclidean distance in 3D space, a new player is reliably detected. After clustering the new features according to Section 2.2, the label of the new player position will be added to the set of active players $\{C_i\}_1^{K'}$. Note that there exists a trade-off between early detection in all camera frames and computational complexity. A low detection frequency may reduce the reliability of capturing new players. However, a low detection frequency is favorable for reducing computational complexity.

3. EXPERIMENTAL RESULTS

We evaluate our SIFT-based multi-view cooperative tracking scheme with the soccer test video set *Barca-St. Andreu* which is provided by the MEDIAPRO group. The videos are captured by three fixed broadcast cameras. The resolution of the videos is 1080×1920 at 15 fps. For comparison purposes, conventional SIFT-based 2D tracking is utilized, which obtains the tracking information by measuring the motion of SIFT features between two successive frames.

We assess the performance of the algorithm by measuring the reliability of tracking. The measure captures the average reliability of successful tracking, i.e, the number of correctly tracked players divided by the number of players in each frame, averaged over all frames. For a fair comparison, we calculate the average reliability for the center camera view V_m. Note, to distinguish the failure of continuous tracking from the failure of appreence detection of new players, we show also results that exclude the detection error in the first frames.

As shown in Table 1, our SIFT-based multi-view cooperative tracking scheme outperforms the reference SIFT-based 2D tracking scheme for multiple player tracking (*Sequence 02-04*). The average reliability improves by more than 10%.

4. CONCLUSIONS

We discussed a SIFT-based multi-view cooperative tracking scheme for acquiring multiple player position information for soccer video. Our scheme utilizes inter-view and inter-frame correlation by extracting SIFT features. Further, the 3D information of these features is exploited to track the position of each player and to solve the occlusion problem. The experimental results show that our SIFT-based multi-view cooperative tracking scheme improves the reliability of tracking by more than 10% when compared to SIFT-based 2D tracking.

5. 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".

6. REFERENCES

[1] 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.

[2] N. Gordon M. Arulampalam, S. Maskell and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, pp. 174–188, Feb. 2002.

[3] R. Collins, "Mean-shift blob tracking through scale space," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. II – 234–40 vol.2.

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

[5] Y. Yuan H. Zhou and C. Shi, "Object tracking using SIFT features and mean shift," *Computer Vision and Image Understanding*, vol. 113, pp. 345–352, Mar. 2009.

[6] R. Kumar T. Zhao, M. Aggarwal and H. Sawhney, "Real-time wide area multi-camera stereo tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 976 – 983 vol. 1.

[7] 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.

[8] N. Netanyahu C. Piatko R. Silverman A. Wu T. Kanungo, D. Mount, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 881 –892, 2002.