ROBUST MULTI-OBJECT TRACKING VIA CROSS-DOMAIN CONTEXTUAL INFORMATION FOR SPORTS VIDEO ANALYSIS

Tianzhu Zhang, Bernard Ghanem

Advanced Digital Sciences Center of Illinois
Singapore

Narendra Ahuja

University of Illinois at Urbana-Champaign
Electrical and Computer Engineering
Urbana, IL USA

ABSTRACT

Multiple player tracking is one of the main building blocks needed in a sports video analysis system. In an uncalibrated camera setting, robust mutli-object tracking can be very difficult due to a number of reasons including the presence of noise, occlusion, fast camera motion, low-resolution image capture, varying viewpoints and illumination changes. To address the problem of multi-object tracking in sports videos, we go beyond the video frame domain and make use of information in a homography transform domain that is denoted the homography field domain. We propose a novel particle filter based tracking algorithm that uses both object appearance information (e.g. color and shape) in the image domain and cross-domain contextual information in the field domain to improve object tracking. In the field domain, the effect of fast camera motion is significantly alleviated since the underlying homography transform from each frame to the field domain can be accurately estimated. We use contextual trajectory information (intra-trajectory and inter-trajectory context) to further improve the prediction of object states within an particle filter framework. Here, intra-trajectory contextual information is based on history tracking results in the field domain, while inter-trajectory contextual information is extracted from a compiled trajectory dataset based on tracks computed from videos depicting the same sport. Experimental results on real world sports data show that our system is able to effectively and robustly track a variable number of targets regardless of background clutter, camera motion and frequent mutual occlusion between targets.

Index Terms— Tracking, Particle Filter, Cross-Domain, Contextual Information

1. INTRODUCTION

Tracking multiple targets has been of broad interest in the computer vision community for decades. A visual-based multi-target tracking system should be able to track a variable number of objects in a dynamic scene and maintain the correct identities of the targets regardless of occlusion and any other visual perturbations (e.g. camera motion, illumination changes, and object resolution). Extensive work has been done over the years [1, 2], as it is a very complicated and challenging problem. In this paper, we address the problem of robust multi-target tracking within sports videos (e.g. American football) by tracking players using hybrid information from both the image and field domains.

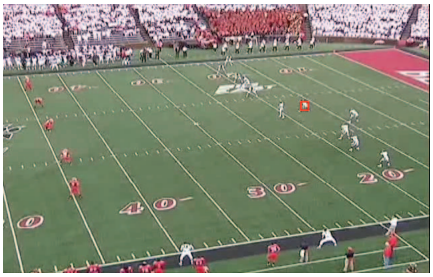
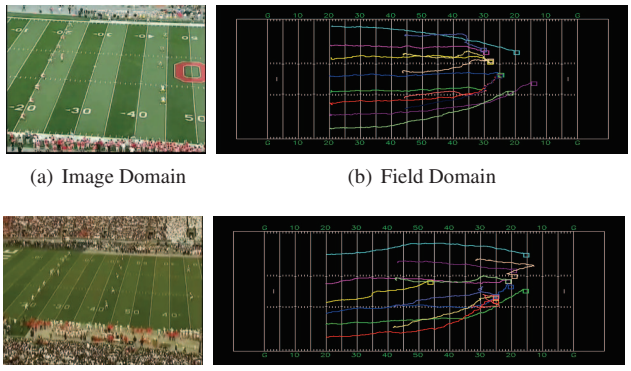


Fig. 1. An exemplar frame from an American football video clip. The red bounding box (15×11 pixels) is an initialization for object tracking. Note that players on the same team have very similar appearances and are usually of low-resolution.

Human activity analysis has been established in the fields of security surveillance and military applications, but the sports world has been extremely under-serviced. Multiple player tracking is one of the main building blocks needed in an effective sports video analysis system. Knowing the location of each player on the field at each point of the game is crucial for sports experts (e.g. coaches, trainers, and sports analysts) to better understand complex player formations and trajectory patterns, which ultimately depict the effectiveness of their teams' strategies as well as their opponents'. Being able to effectively track multiple players at one time can enable the development of reliable activity recognition and higher-level processing modules for sports video analysis. Such a tracking building block will have a positive impact on how sports experts analyze game footage, how content providers identify/display particular sports events and highlights accompanied with relevant advertisements, and how end users browse and query large collections of sports video.

Tracking players in the image domain is a difficult and challenging problem for several reasons: **(1)** Tracking players in sports is hard. Players on the same team have similar appearance information as shown in Fig. 1. This leads to the loss of a player's track when he/she is moving near other players from the same team. **(2)** For sports video, it is always recorded in far-field view. Players are blurry and are often captured in low-resolution as exemplified in Fig. 1 (the red bounding



(a) Image Domain (b) Field Domain

(c) Image Domain (d) Field Domain

Fig. 2. 2(a) and 2(c) are two exemplar frames from two video clips of American football. 2(b) and 2(d) are tracking results of multiple players. From the results, we can see that trajectories in different videos are similar due to the inherent rules of the game. Clearly, contextual information based on prior tracking knowledge can be useful to improve object tracking. Moreover, a player's motion pattern in the homography field domain can be much more informative than motion patterns in the image domain especially due to camera motion. This is a primary reason why coaches and sports experts prefer field domain trajectories over image domain ones.

box is 15×11 pixels), thus precluding robust tracking using a player's individual local information. **(3)** Tracking players in the presence of camera motion (i.e. pan, tilt, or zoom) is significantly more difficult than when the camera is static, since background subtraction becomes non-trivial. Coupled with motion blur, frequent occlusions and exit/re-entrance of players, tracking becomes quite challenging. **(4)** Motion features extracted from the image domain may not provide enough discriminative information to reliably track multiple players on the field. For instance, in the image domain and due to perspective transformation, apparent local player motion is highly impacted by motion parallax due to camera motion.

The aforementioned reasons render it difficult to realize a robust multi-player tracking system that functions solely in the image domain of sports videos. Consider state-of-the-art object tracking methods such as tracking-by-detection methods [1, 3, 4, 5]. The player has low-resolution, as shown in Fig. 1, and there is limited appearance information to train a reliable object detector, which renders these methods ineffective. Therefore, any tracking method that is only based on image domain information will tend not to be robust and fail due to limited discriminative information. Fortunately, we can resort to the information from the field domain to alleviate this problem. Our decision to use cross-domain contextual information based approach is motivated by several factors. First, in the field domain, we can eliminate fast camera motion effects (e.g. parallax) through the homographical correspondence between points in the field and image domains. Second, the trajectory of each player enjoys many characteristics that allow it to be more predictable in the field domain as shown in Fig. 2. This can help predict a player's next position, leading to more robust tracking. Third, due to game rules, players in different video clips have similar trajectories as shown in Fig. 2. This demonstrates that using prior player trajectories (e.g. from a trajectory dataset) can help improve player tracking. Therefore, we attempt to implement a robust multi-object tracking system using cross-domain contextual information from both the field and image domains. In our algorithm, we employ the particle filter framework [6] to guide the tracking process. The cross-domain contextual information is integrated into the framework and acts as a guide for particle propagation and proposal.

2. OUR PROPOSED METHOD

2.1. Particle Filter

The particle filter [7] is a Bayesian sequential importance sampling technique for estimating the posterior distribution of state variables characterizing a dynamic system. It provides a convenient framework for estimating and propagating the posterior probability density function of state variables regardless of the underlying distribution, consisting of essentially two steps: prediction and update. Let x_t denote the state variable describing the parameters (e.g. appearance or motion features) of an object at time t. The predicting distribution of x_t given all available observations $z_{1:t-1} = \{z_1, z_2, \cdots, z_{t-1}\}$ up to time $t-1$, denoted by $p(x_t|z_{1:t-1})$, is recursively computed in (1).

$$p(x_t|z_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx_{t-1} \quad (1)$$

At time t, the observation z_t is available and the state vector is updated using Bayes rule, as in (2), where $p(z_t|x_t)$ denotes the observation likelihood.

$$p(x_t|z_{1:t}) = \frac{p(z_t|x_t)p(x_t|z_{1:t-1})}{p(z_t|z_{1:t-1})} \quad (2)$$

In the particle filter framework, the posterior $p(x_t|z_{1:t})$ is approximated by a finite set of N samples $\{x_t^i\}_{i=1}^N$ (called particles) with importance weights w_i. The candidate samples x_t^i are drawn from an importance distribution $q(x_t|x_{1:t-1}, z_{1:t})$ and the weights of the samples are updated as Eq.(3). To avoid degeneracy, particles are resampled to generate a set of equally weighted particles by their importance weights.

$$w_t^i = w_{t-1}^i \frac{p(z_t|x_t^i)p(x_t^i|x_{t-1}^i)}{q(x_t|x_{1:t-1}, z_{1:t})} \quad (3)$$

Using the particle filter framework, we model the observation likelihood and the proposal distribution as follows. For the observation likelihood $p(z_t|x_t)$, we follow [1] and adopt a multi-color observation model based on Hue-Saturation-Value (HSV) color histograms and a gradient-based shape model using Histograms of Oriented Gradients (HOG). We apply the Bhattacharyya similarity coefficient to define the distance between HSV and HOG histograms respectively. Moreover, we also divide up the tracked regions into two sub-regions (2×1) in order to describe the spatial layout of color and shape features for a single player. We model the proposal distribution $q(x_t|x_{1:t-1}, z_{1:t})$ as shown in (4), by fusing information from different sources described in the subsections 2.2 and 2.3.

$$q(x_t|x_{1:t-1}, z_{1:t}) = \alpha_1 p(x_t|x_{t-1}) + \alpha_2 p(x_t|x_{t-L:t-1}) + \alpha_3 p(x_t|x_{1:t-1}, T_{1:K}). \quad (4)$$

To decide the values of α_1, α_2, and α_3, we can use a cross-validation set. For simplicity, α_1, α_2, and α_3 are equal and set to be $1/3$ by experience in our experiments.

2.2. Intra-trajectory Contextual Information

For a tracked object from frame 1 to $t-1$, we obtain $t-1$ points: $\{\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{t-1}\}$, which correspond to a short trajectory denoted as T_0. Our aim is to predict the next state at time t using the previous states in a non-trivial data-driven fashion. As shown in Fig. 2, for each object, its previous states can help to predict its next state in the field domain. For simplicity, we just consider the most recent L points in the trajectory to predict the state at time t. To obtain robust intra-trajectory information, we adopt $\hat{p}_{t-L}$ as the start point, and all other more current points to define the difference as $\nabla\hat{p}_l = (\hat{p}_{t-L+l} - \hat{p}_{t-L})/l$, where $\nabla\hat{p}_l$ is also denoted as $\nabla\hat{p}_l = (\nabla x_l, \nabla y_l)^{'}$, $l = 1, 2, \cdots, L$. In this way, given $\nabla\hat{p}_{1:L-1}$, the probability of $\nabla\hat{p}_L$ is defined as:

$$p(\nabla\hat{p}_L|\nabla\hat{p}_{1:L-1}) = \frac{e^{-\frac{1}{2}(\nabla\hat{p}_L - u_{\nabla\hat{p}_l})^T\Sigma^{-1}(\nabla\hat{p}_L - u_{\nabla\hat{p}_l})}}{2\pi|\Sigma|^{\frac{1}{2}}} \quad (5)$$

Here Σ is assumed to be diagonal matrix. To consider the temporal information, each $\nabla\hat{p}_l$ is weighted with λ_l defined as $\lambda_l = \frac{e^{-l^2/\delta^2}}{\sum_l e^{-l^2/\delta^2}}$. Based on the weight λ_l, $u_{\nabla\hat{p}_l}$ and Σ are defined as $u_{\nabla\hat{p}_l} = \sum_{l=1}^{L-1}\lambda_l\nabla\hat{p}_l$ and $\Sigma = diag(\delta^2_{\nabla x_l}, \delta^2_{\nabla y_l})$, where $\delta^2_{\nabla x_l} = \frac{\sum_{l=1}^{L-1}\lambda_l}{(\sum_{l=1}^{L-1}\lambda_l)^2 - \sum_{l=1}^{L-1}\lambda_l^2}\sum_{l=1}^{L-1}\lambda_l(\nabla x_l - u_{\nabla x_l})^2$, and $\delta^2_{\nabla y_l}$ has the same form. Finally, $p(x_t|x_{t-L:t-1})$ in Eq.(4) is defined as $p(x_t|x_{t-L:t-1}) = p(\nabla\hat{p}_L|\nabla\hat{p}_{1:L-1})$.

2.3. Inter-trajectory Contextual Information

Given the dataset introduced in Section 3.1, for the short trajectory T_0, we can obtain its K nearest neighbors by use of dynamic time warping (DTW) [8], and the K trajectories are denoted as $T_{1:K}$. For each T_k, $k = 1, \ldots, K$, we calculate the Euclidean distance between its points and $\hat{p}_{t-1}$, and select the point $\hat{p}_s$ with the smallest distance. Then, we select L points from the point $\hat{p}_s$ to $\hat{p}_{s+L-1}$ in trajectory T_k to obtain $p_k(\nabla\hat{p}_i|\nabla\hat{p}_{1:L-1})$ as the same as Eq.(5), where $\nabla\hat{p}_i = \hat{p}_i - \hat{p}_{t-1}$, and $\hat{p}_i$ is a certain point in field domain. Given T_0 and $T_{1:K}$, the probability of $\nabla\hat{p}_i$ for each point $\hat{p}_i$ in field domain is defined as:

$$p(\nabla\hat{p}_i|T_0, T_{1:K}) = \sum_{k=1}^{K}\eta_k p_k(\nabla\hat{p}_i|\nabla\hat{p}_{1:L-1}), \quad (6)$$

where η_k is the weight of the k-th trajectory and is set to be $\eta_k = \exp(-\frac{(Dist(T_k,T_0)-u_0)^2}{2\delta_0^2})$. The $Dist(T_k, T_0)$ is the distance between two trajectories, and u_0 and δ_0 are obtained from the dataset. For each trajectory in the database, we can obtain its K nearest neighbors, and calculate their distances. Then, based on all the distances, u_0 and δ_0 can be obtained.

Based on T_0 and the K nearest neighbors, $p(x_t|x_{1:t-1}, T_{1:K})$ in Eq.(4) is defined as $p(x_t|x_{1:t-1}, T_{1:K}) = p(\nabla\hat{p}_i|T_0, T_{1:K})$.

This inter-trajectory contextual information is useful and effective to improve the object tracking, because the players in different video clips have similar trajectories as shown in Fig. 2. For a trajectory T_0, if there is no similar trajectory in the dataset, the K nearest neighbours have very small weights η_k as shown in Eq.(6). As a result, the probability $p(\nabla\hat{p}_i|T_0, T_{1:K})$ is very small, and no useful inter-trajectory contextual information can be exploited. However, this happens rarely if the dataset is large-scale.

3. EXPERIMENTAL RESULTS

3.1. Dataset and Implemention Details

Our dataset contains 93 low-resolution videos of different football plays from 10 different teams, each around 400 frames long. Each video contains footage of a single football play shot from a PTZ camera with a sideline view high above the field. Fig. 1 depicts a typical view from this camera. The dataset is very complex. For each team, there are different background colors and environments as shown in Fig. 2. Every video is pre-processed to register frames to an overhead model of the football field using the method described in [9], thereby enabling us to determine players' locations in football field coordinates.

It is time-consuming to build the database manually. Therefore, we implement a simple method that does not make use of inter-trajectory context information and adopt interactive object tracking. For each video clip, we track 8 to 13 players per frame. To evaluate the performance of the proposed tracking approach, we randomly select 5 video sequences as the testing set, and the rest are used for building the database. For the testing video clips, we create a tracking ground truth bounding box of the target in each frame for quantitative evaluation by manually annotating the data.

To evaluate the performance of our tracker, we use a score based on the PASCAL challenge object detection score: Given the detected bounding box ROI_D and the ground truth bounding box ROI_GT, the overlap score evaluates as $score = area(ROI_D \cap ROI_{GT})/area(ROI_D \cup ROI_{GT})$. For each track, we get the average score. Then, we average these scores to obtain the evaluation score for the video. We compare our method with two state-of-the-art visual trackers for sports video analysis [1, 10]. For the baselines, we use publicly available code and adopt the same parameters as the authors.

3.2. Results and Analysis

Fig. 3 shows the probability map of intra_trajectory and inter_trajectory contextual information for a short trajectory in red. The pixel with high probability may be the next position

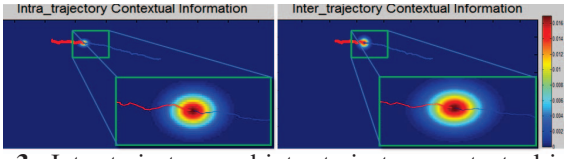
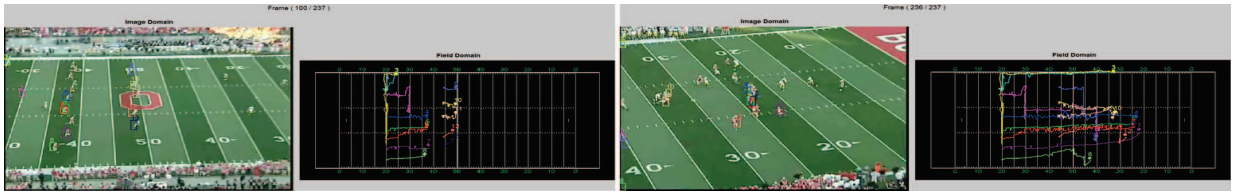
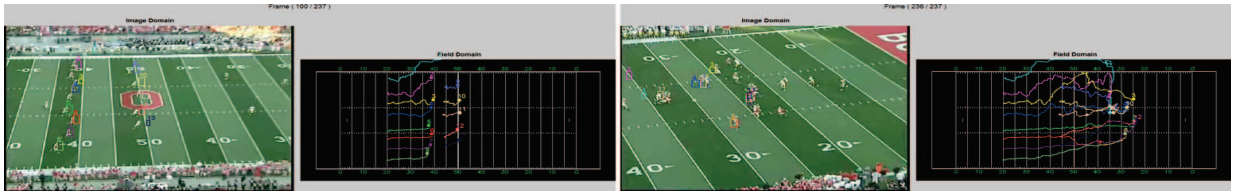


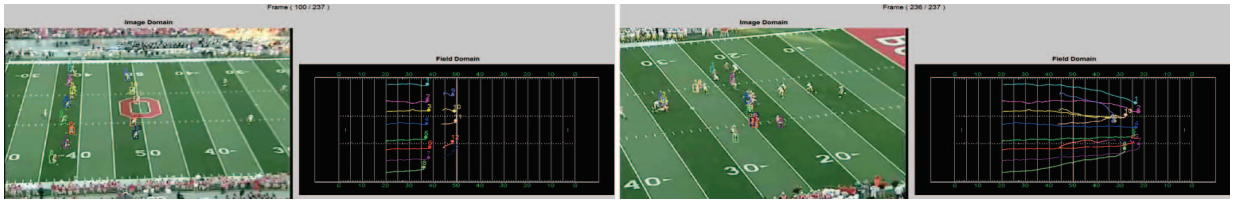
Fig. 3. Intra_trajectory and inter_trajectory contextual information for a short red trajectory in field domain. The blue trajectory is its ground truth path in future.



(a) Online boosting tracker results [10].



(b) Boosted particle filtering tracker results [1].



(c) Our proposed tracker results.

Fig. 4. Tracking results of different algorithms. The number and color of bounding boxes and trajectories show the correspondences between the image domain and the field domain. For a better view, please see the pdf file.

Table 1. The average tracking scores on five video sequences.

Method	$Video_1$	$Video_2$	$Video_3$	$Video_4$	$Video_5$
OAB[10]	0.354	0.394	0.380	0.347	0.297
BPF[1]	0.386	0.404	0.356	0.367	0.295
Ours	**0.734**	**0.756**	**0.746**	**0.728**	**0.689**

of state. Based on the probability map, we can confirm the contextual information is effective to help predict the state. Moreover, the standard deviation in x coordinate is higher, as players are more likely to run straight forward. The quantitative results are summarized in Table 1. This table gives the average tracking scores of each approach in five sequences, and our method achieves more than 30% improvement. We also show the tracking results for the three trackers in Fig. 4. From the results we can see that although the traditional tracking approaches cannot track the players in American football well, our proposed method can track the players robustly and stably. That is because there is not enough appearance information in the image domain for methods [1, 10]. However, the cross-domain contextual information is effective to improve object tracking.

4. CONCLUSION

In this paper, we propose a novel method to track multi-players in low-resolution videos of American football with cross-domain context information. Because the camera motion is eliminated in field domain, object intra-trajectory context information and inter-trajectory context information are helpful to predict the players states. Experimental results on many real-world challenging video clips demonstrate our method is effective and useful to improve the multi-object tracking performance. Our cross-domain tracker is generic, and can also be used in other fields, such as video surveillance.

Acknowledgment

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

5. REFERENCES

[1] Kenji Okuma, Ali Taleghani, Nando De Freitas, O De Freitas, James J. Little, and David G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *ECCV*, 2004, pp. 28–39.

[2] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking," in *CVPR*, June 2009, pp. 240–247.

[3] Wei-Lwun Lu, Jo-Anne Ting, Kevin P. Murphy, and James J. Little, "Identifying Players in Broadcast Sports Videos using Conditional Random Fields," in *CVPR*, 2011.

[4] Yizheng Cai and C Yizheng Cai, "Robust visual tracking for multiple targets," in *ECCV*, 2006, pp. 107–118.

[5] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *ICCV*, 2009.

[6] Michael Isard and Andrew Blake, "Condensation - conditional density propagation for visual tracking," *IJCV*, vol. 29, pp. 5–28, 1998.

[7] Arnaud Doucet, Nando De Freitas, and Neil Gordon, "Sequential monte carlo methods in practice," in *Springer-Verlag*. New York, 2001.

[8] Hiroaki Sakoe, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, 1978.

[9] Bernard Ghanem, Tianzhu Zhang, and Narendra Ahuja, "Robust video registration applied to field-sports video analysis," in *ICASSP*, 2012.

[10] Helmut Grabner, Michael Grabner, and Horst Bischof, "Real-Time Tracking via On-line Boosting," in *BMVC*, 2006.