Tracking People in Broadcast Sports

Angela Yao[1], Dominique Uebersax[1], Juergen Gall[1] and Luc Van Gool[1,2]

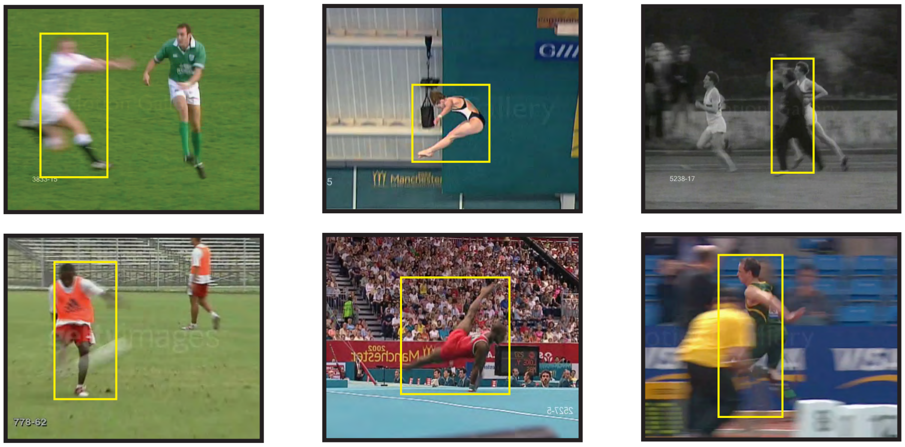
[1] ETH Zurich, Switzerland [2] KU Leuven, Belgium
{yaoa,gall,vangool}@vision.ee.ethz.ch, duebersa@ee.ethz.ch

Abstract. We present a method for tracking people in monocular broadcast sports videos by coupling a particle filter with a vote-based confidence map of athletes, appearance features and optical flow for motion estimation. The confidence map provides a continuous estimate of possible target locations in each frame and outperforms tracking with discrete target detections. We demonstrate the tracker on sports videos, tracking fast and articulated movements of athletes such as divers and gymnasts and on non-sports videos, tracking pedestrians in a PETS2009 sequence.

1 Introduction

Object tracking in video is a long-standing computer vision problem; in particular, tracking people has captured the interest of many researchers due to its potential for applications such as intelligent surveillance, automotive safety and sports analysis. State-of-the-art people trackers have predominantly focused on pedestrians for traffic or surveillance scenarios. For sports analysis, however, standard pedestrian trackers face significant challenges since in many broadcast sports, the camera moves and zooms to follow the movements of the athlete. Furthermore, in some sports, the athlete may perform abrupt movements and have extensive body articulations that result in rapid appearance changes and heavy motion blur. As such, sports tracking to date [1–6] has been limited to team sports such as football and hockey, in which there is wide view of the playing field and athletes remain relatively upright. In addition, these works are primarily focused on the data-association problem of multi-target tracking and do not deviate substantially from the pedestrian tracking scenario.

In the current work, we present a method for tracking people in monocular broadcast sports videos by coupling a standard particle filter [7] with a vote-based confidence map of an "athlete"-detector [8]. We target sporting disciplines in which the athletes perform fast and highly articulated movements, e.g. diving and gymnastics. Tracking in these types of sports is particularly difficult since the athletes do not remain in an upright configuration. Our confidence map, built from the Hough accumulator of a generalized Hough transform designed for people detection, is well suited for handling pose and appearance changes and athlete occlusions, as it is generated from a vote-based method. While we focus on tracking in broadcast sports clips, as they provide a challenging testbed, our method is applicable to generic people tracking in unconstrained videos. We demonstrate the tracker's effectiveness on the UCF Sports Dataset [9], a collection of footage from the 2008 Olympics and a PETS 2009 sequence [10].



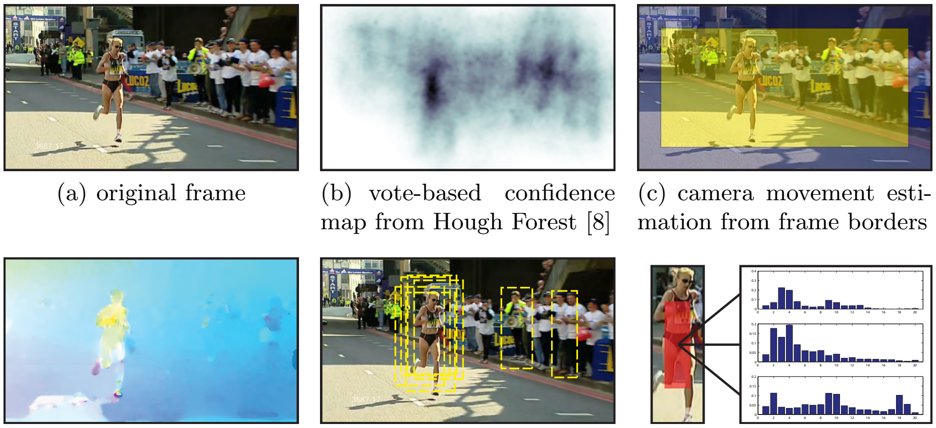
(a) motion blur (b) extensive articulation (c) occlusion

Fig. 1. Select frames from the UCF Sports Dataset [9], showing challenges of tracking in sports videos such as *(a)* motion blur, *(b)* extensive body articulation and *(c)* occlusions.

2 Related Works

Early approaches in sports tracking began with background extraction and then morphological operations to isolate foreground areas which may represent the athlete [1, 2, 11]. Tracking was then performed by enforcing spatial continuity through either Kalman or particle filtering. These approaches, both single- and multi-camera, relied heavily on colour as a cue for separating the athletes from the background as well as for tracking, though shape and motion information of the athletes have also been used [12, 4]. Most of the proposed algorithms, however, have been designed for specific sports, such as soccer [1, 2], speed-skating [13] or hockey [3] and rely on sport-specific scene-knowledge, such as distances between field lines [14].

Accurate modelling of target observations, be it athletes, pedestrians or generic objects has been the focus of several current tracking works. One line of approach learns and adapts appearance models online [15–17]; these methods cope well with appearance changes and are not limited to tracking specific object classes, but are susceptible to drift as tracking errors accumulate. Another line of approach uses pre-trained models of the targets. Tracking-by-detection methods follow this type of paradigm, in which object detectors are first trained offline and detections across the sequence are then associated together to form the track, e.g. by particle filtering. Tracking-by-detection has been used for pedestrians [5, 18, 19] and in specific sports such as hockey [3, 5] and soccer [5, 6]. All these approaches, however, assume that the humans remain upright - an assumption that does not hold for broadcast sports videos in general.



(a) original frame

(b) vote-based confidence map from Hough Forest [8]

(c) camera movement estimation from frame borders

(d) optical flow

(e) particle distribution

(f) colour and texture appearance evaluation

Fig. 2. Components of the sports tracker. From the original frame *(a)*, the vote-based confidence map *(b)* is computed using a Hough Forest [8]. The dynamical model estimates camera motion from the frame border *(c)* and motion of the tracked athlete from the frame interior using optical flow *(d)*. Each particle in the particle distribution *(e)* is weighted according to the confidence map and appearance features such as colour and texture *(f)*.

The key component of our tracker is the use of a vote-based confidence map to estimate the location of the targets. It is similar in spirit to the Fragment Tracker in [20], which tracks object fragments or patches that vote for an object center. Our work differs from [20] in that we track possible object centres from the accumulated votes in the confidence map rather than the individual patches that vote for a center.

3 Sports Tracker

The sports tracker is a tracking-by-detection approach with three components: *(1)*a continuous vote-based confidence map to estimate the target location (see 3.2), *(2)*appearance matching of the target based on feature templates (see 3.3) and *(3)*motion estimation of the camera and the target from optical flow (see 3.4).

3.1 Tracking Overview

Tracking in the sports videos is done using a particle filter [7]. We model the state $\boldsymbol{s} = \{x, y, c, u, v, d\} \in \mathbb{R}^6$ of a human by the image position and scale (x, y, c) and velocity and change in scale (u, v, d). For particle i, the weight at frame t is assigned as follows:

$$w_t^i = \frac{1}{Z} \exp\Big(-K \cdot \big(\alpha \cdot V_1(\boldsymbol{s}_t^i) + (1-\alpha) \cdot \sum_f \lambda_f V_2(\boldsymbol{s}_t^i, f)\big)\Big). \tag{1}$$

The term V_1 measures the response in the vote space (Figure 2*(b)*, see 3.2) for particle $\boldsymbol{s}_t^i$. The term V_2 measures the similarity of particle $\boldsymbol{s}_t^i$ with respect to some template appearance feature f extracted from the associated bounding box of the particle (Figure 2*(f)*, see 3.3). K is a scaling constant and $\alpha \in [0,1]$ is a weighting parameter for V_1 and V_2. λ_f are weighting parameters between the different features and sum up to 1. Z is the normalization term of the weights.

The tracker is initialized using the ground truth from the first frame of the sequence. Particles are propagated by a dynamical model accounting for camera motion (Figure 2*(c)*) and estimated athlete motion(Figure 2*(d)*, see 3.4).

3.2 Vote-Based Confidence Map

The confidence map is generated from the output of a Hough forest [8] trained for detecting athletes. The Hough forest is a random forest trained to map image feature patches to probabilistic votes in a 3D Hough accumulator H for locations and scales of the athlete. We use cropped and scale-normalized images of the athletes as positive examples, background images as negative examples, and colour and histograms of gradients [21] as features. For a detailed description of the training procedure, we refer to [8]. For detection, feature patches are densely sampled from the image and passed through the trees of the Hough forest to cast votes in H. While a detector as in [8] thresholds the local maxima in H to obtain a discrete set of object hypotheses, we consider H as a continuous confidence mapping of athlete locations and scales. From H, the vote response $V_1\left(\boldsymbol{s}_t^i\right)$ of particle $\boldsymbol{s}_t^i$ is determined by

$$V_1(\boldsymbol{s}_t^i) = -\log\Big(\sum_{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{s}_t) \cap H} G(\boldsymbol{s}_t^i - \boldsymbol{x})\Big), \tag{2}$$

i.e. we sum the votes in the neighborhood $\mathcal{N}$ of $\boldsymbol{s}_t$ weighted by a Gaussian kernel G. Note that the sum is in the range of $[0,1]$.

3.3 Appearance Model

The appearance of particle $\boldsymbol{s}_t^i$, denoted as $V_2\left(\boldsymbol{s}_t^i, f\right)$, is a measure of similarity between that particle's feature response $h^f\left(\boldsymbol{s}_t^i\right)$ and some template h_T^f for feature f. To measure similarity, we use the Bhattacharyya coefficient BC:

$$V_2\left(\boldsymbol{s}_t^i, f\right) = 1 - BC(h_T^f, h^f(\boldsymbol{s}_t^i)) \tag{3}$$

As image features, we use HSV colour histograms and local binary patterns [22] to model colour and texture respectively. For the template, we use a weighted mixture of the particle's feature response in the initial frame at t_0 and the previous frame t–1. Weighting of the individual appearance features in the final particle weight (Equation 1) is determined by λ_f, in our case λ_{colour} and $\lambda_{texture}$.

3.4 Dynamical Model

For the dynamical model, we use an estimated velocity based on optical flow. The reason for this is two-fold. First, constant-velocity models which perform well for tracking walking or running people perform poorly for actions in which the athletes move erratically, i.e. in gymnastics. Secondly, in many broadcast sports, the cinematography already provides some framing and tracking of the athlete, i.e. when the camera pans to follow the athlete across a scene. As such, the position of the athlete changes in an inconsistent manner within the frame and it is necessary to estimate the particle motion while accounting for camera motion. Particles are propagated from frame to frame by

$$(x,y,c)_t^i = (x,y,c)_{t-1}^i + (u,v,d)_{t-1}^i + \mathcal{N}\left(0, \boldsymbol{\sigma}_{tran}\right), \tag{4}$$

where $\boldsymbol{\sigma}_{tran}$ is the variance of added Gaussian noise for the transition. Velocity is estimated as a weighted mixture between camera-compensated optical flow and velocity in the previous frame, while change in scale remains constant.

$$(u,v)_t^i = \eta \cdot \left((u,v)_{t-1}^{of} - \gamma \cdot (u,v)_{t-1}^{cam} \right) + (1-\eta) \cdot (u,v)_{t-1}^i \tag{5}$$

Optical flow is computed according to [23]; camera motion is estimated as the average optical flow in the border of the frame (Figure 2*(b)*). η is a weighting parameter between estimated motion versus a constant velocity assumption, while γ serves as a scaling parameter for the estimated camera motion.

4 Experiments

4.1 Datasets

We evaluate our tracker on sports and non-sports videos. For sports, we use the UCF Sports Dataset [9] and our own collection of Olympics footage. The UCF dataset, consisting of 150 sequences (50-100 frames each) from network news videos, was originally intended for action recognition. To supplement the UCF dataset, we annotated 31 sequences (150-2000 frames each) from the 2008 Olympics, featuring sports such as diving, equestrian and various disciplines of gymnastics. The sequences are longer and more challenging than UCF, with significant motion blur and de-interlacing artifacts. For non-sports videos, we track three people from the PETS 2009 [10] sequence *S2.L1, View001.* For the sports datasets, we train on all images of annotated athletes within the dataset other than from the test sequence, in a leave-one-out fashion. For the PETS sequence, we trained on the TUD pedestrian database [18].

4.2 Evaluation

For evaluation, we use the VOC [24] criterion (the intersection over union, IOU, of the tracked bounding box and the ground truth bounding box must be greater

Experimental Variation	Affected Parameter/Variable	% of frames with IOU> 0.5
Default	NA	75.4 ± 13.4
Discrete detections	V_1 from discrete detections	26.1 ± 17.2
No detections	$\alpha = 0$	28.1 ± 17.3
No colour features	$\lambda_{colour} = 0$, $\lambda_{texture} = 1$	73.9 ± 12.3
No texture features	$\lambda_{colour} = 1$, $\lambda_{texture} = 0$	71.3 ± 10.3
No appearance features	$\alpha = 1$	70.4 ± 13.1
No camera compensation	$\gamma = 0$	71.8 ± 11.7
Constant velocity	$\eta = 0$	71.8 ± 15.0

Table 1. Average tracking performance on the Olympics sequences, where a higher % indicates better performance. There is a decrease in tracking performance with each removed component of the tracker; the most critical component seems to be the vote map, as using discrete components results in significantly lower performance.

than 0.5). We hand annotated select frames of the Olympics data and the PETS sequence and used linear interpolation to generate bounding boxes for the frames in between. For the UCF database, bounding boxes were provided as a part of the ground truth annotation released with the data.

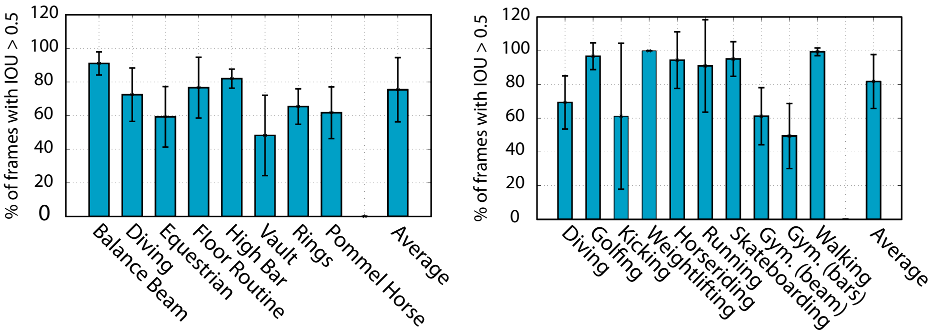
We run three experiments on the Olympics data to test the impact of each component of the tracker. First, the confidence map is compared with discrete detections; for fair comparison, we generate the discrete detections from the confidence maps by thresholding[1] the local maxima of H (see 3.2). Second, the effect of the appearance modelling is tested by removing the colour and texture features from the tracker. In the third experiment, we vary the η and γ parameters and look at the effects of removing camera compensation as well as comparing our current dynamic model to a constant velocity model. We also compare our tracker's performance on the PETS2009 sequence with the Fragment Tracker in [20], using source code provided on the author's website[2]. Run time on all datasets was around 1 second per frame for 50 particles on a standard CPU.

5 Results

Olympics Data We take the following parameter settings $\{\alpha$=0.5, λ_{colour}=0.09, $\lambda_{texture}$=0.91, η=0.3, γ=1.5$\}$ and use these as our default scenario. Parameters are set at these values for all experiments unless otherwise stated. Results for default scenario, split by discipline are shown in Figure 3 *(a)*. Tracking results from the first three experiments are shown in Table 1. From the first experiment, we see that using the vote-based confidence map in the tracker gives a significant improvement over the use of discrete detections. In fact, for the sports, having discrete detections is comparable to not using any detections (α=0). This can be attributed to the many false-positive detections with high confidences, which have the effect of attracting and clustering the particles to erroneous locations. Our second experiment shows that removing either or both appearance

[1] The threshold was set to achieve a high recall.

[2] http://www.cs.technion.ac.il/~amita/fragtrack/fragtrack.htm



(a) Olympics tracking performance

(b) UCF tracking performance

Fig. 3. Average tracking performance by sport for *(a)*Olympics Dataset and *(b)*UCF Sports Dataset, where a higher % indicates better performance.

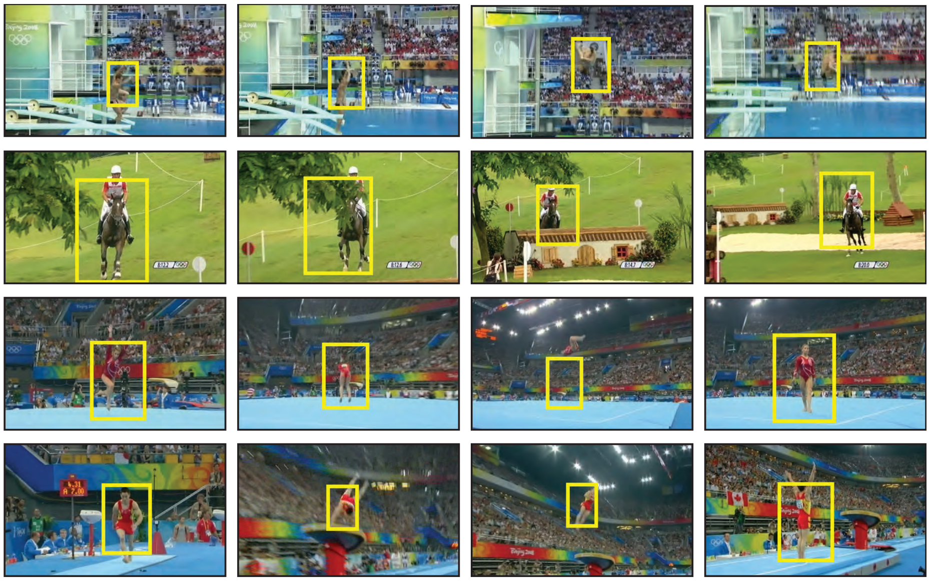


Fig. 4. Tracking on the Olympics sequences: select frames from diving (top), equestrian (second row), floor routine (third row) and vault (bottom). The tracker successfully follows the athletes but has difficulty with very fast motions, e.g. on the floor routine, in the third frame, the tracker fails to track the tumbling sequence through the air.

features results only in a slightly decreased performance, again emphasizing the importance of the confidence map in the tracker. In the last experiment, we show that the use of our motion estimate in the dynamical model outperforms a constant velocity model, particularly with having the camera compensation. Varying η and γ had little effect, with performance ranging from 71.6%-74.9%. Select frames from the tracked results are shown in Figure 4.

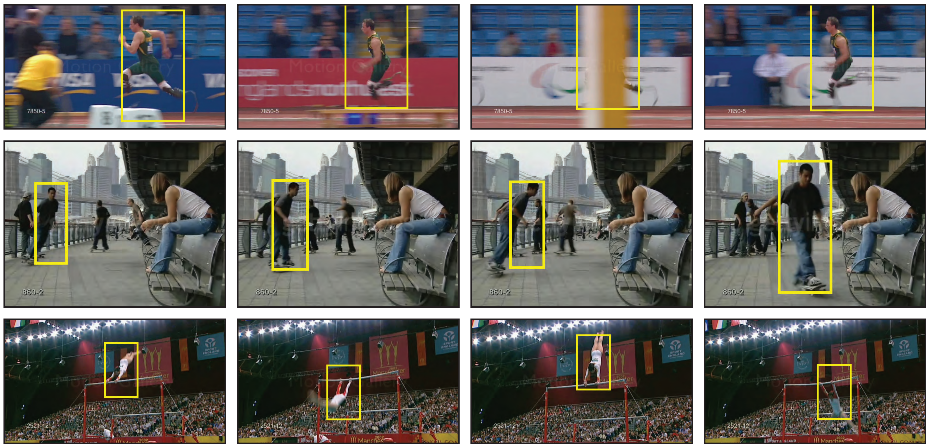


Fig. 5. Tracking on the UCF Sports Dataset, showing select frames from running (top row), skateboarding (middle row) and gymnastics (bottom row).

UCF Sports Dataset Tracking performance for the UCF Dataset are shown in Figure 3*(b)*; select frames from the tracks are shown in Figure 5. On average, 81.8% ± 16.0% of the frames have tracks with an IOU greater than 0.5. The tracker performs well in sports where people remain upright, i.e. golfing, running, and skateboarding, but faces some difficulty with sports with more extensive articulation such as diving, kicking and gymnastics. Part of the error results from ground truth being tight bounding boxes around the athletes while tracked bounding boxes are of a fixed ratio.

PETS2009 We compare the performance of our Sports Tracker with the Fragment Tracker [20] in Table 2. The Sports Tracker successfully follows two of the three tracks, but breaks down on track 3, most likely due to the lack of multiple target handling. There are two identity switches, first from the target to another person at frame 31 when several people group together and then back to the target after frame 115. Select frames are shown in Figure 6. The Fragment Tracker successfully tracks one of the three tracks, but suffers from drift on the other two tracks and around 100 frames into the tracks, loses the target completely.

Track	Frame	Sports Tracker	Fragments Tracker [20]
1	21 - 259	85.4	14.8
2	222 - 794	95.5	12.6
3	0 - 145	13.8	78.6

Table 2. Comparison of the Sports Tracker with the Fragments Tracker in [20] on the PETS2009 *S2.L1 View001* sequence. Results shown are the % of frames with IOU>0.5, where a higher % indicates better performance.

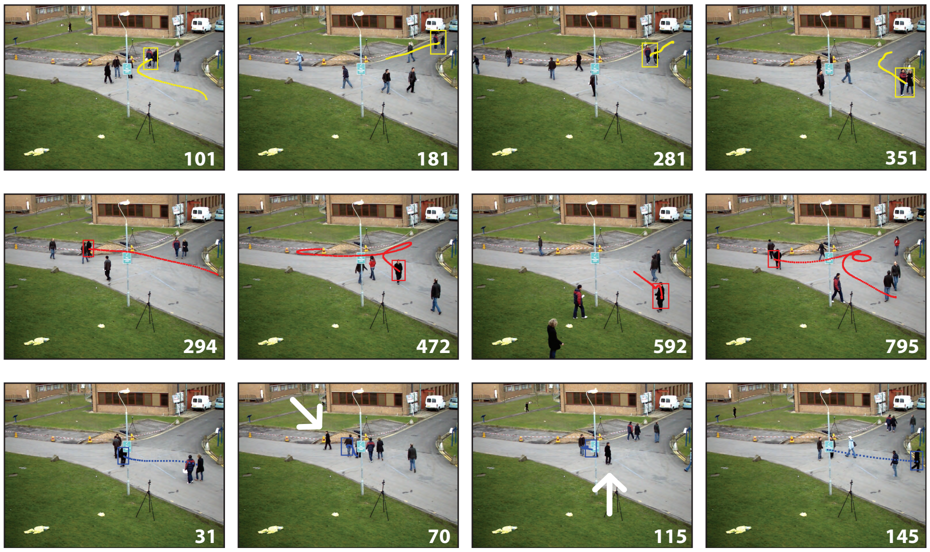


Fig. 6. Select frames from the PETS2009 sequence. The tracker successfully follows the target in track 1 and 2 (top and middle row). Track 2 is particularly challenging as it is over 500 frames long and several people including the target are all wearing black clothing. In frame 294 of track 2, the tracker handles occlusion of the target by another person wearing similar coloured clothing. In frame 31 of track 3 (bottom row), there is an identity switch (true target is indicated by the white arrow); in frame 115, the tracker switches back onto the correct target. Figure is best viewed in colour

6 Conclusion

We have presented a method for tracking athletes in broadcast sports videos. Our sports tracker combines a particle filter with the vote-based confidence map of an object detector. The use of feature templates and target motion estimates add to the performance of the tracker, but the strength of the tracker lies in the confidence map. By providing a continuous estimate of possible target locations in each frame, the confidence map greatly outperforms tracking with discrete detections. Possible extensions to the tracker include making voting for the confidence map adaptive and online, so that tracked bounding boxes are of varying ratios to yield tight bounding boxes around the athlete's body, and making a multi-target version of the tracker to better handle team sports.

Acknowledgements The authors would like to thank Thomas Brox for his optical flow code. This research was funded by the Swiss National Foundation NCCR project IM2 and the EC project IURO; Angela Yao was also funded by NSERC Canada.

References

1. Kang, J., Cohen, I., Medioni, G.: Soccer player tracking across uncalibrated camera streams. In: IEEE International Workshop on Visual Surveillance and Performance

Evaluation of Tracking and Surveillance (VS-PETS). (2003)

2. Choi, K., Seo, Y., Lee, S.: Probabilistic tracking of soccer players and ball. In: ACCV. (2004)
3. Okuma, K., Taleghani, A., de Freitas, N., Little, J., Lowe, D.: A boosted particle filter: Multitarget detection and tracking. In: ECCV. (2004)
4. Kristan, M., Pers, J., Perse, M., Kovacic, S.: Closed-world tracking of multiple interacting targets for indoor-sports applications. CVIU **113**(5) (2009) 598 – 611
5. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV. (2009)
6. Hess, R., Fern, A.: Discriminatively trained particle filters for complex multi-object tracking. In: CVPR. (2009)
7. Doucet, A., Freitas, N.D., Gordon, N., eds.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
8. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR. (2009)
9. M.D.Rodriguez, Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008)
10. Ferryman, J., Shahrokni, A.: Pets2009: Dataset and challenge. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. (2009)
11. Sullivan, J., Carlsson, S.: Tracking and labelling of interacting multiple targets. In: ECCV. (2006)
12. Lu, W.L., Little, J.J.: Tracking and recognizing actions at a distance. In: Proceedings of the ECCV Workshop on Computer Vision Based Analysis in Sport Environments (CVBASE '06), Graz, Austria (May 2006)
13. Liu, G., Tang, X., Cheng, H.D., Huang, J., Liu, J.: A novel approach for tracking high speed skaters in sports using a panning camera. Pattern Recogn. **42**(11) (2009) 2922–2935
14. Khatoonabadi, S.H., Rahmati, M.: Automatic soccer players tracking in goal scenes by camera motion elimination. Image Vision Comput. **27**(4) (2009) 469–479
15. Collins, R., Liu, Y., Leordeanu, M.: On-line selection of discriminative tracking features. TPAMI **27**(1) (2005) 1631 – 1643
16. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: ECCV. (2008)
17. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR. (2009)
18. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR. (2008)
19. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV **77**(1-3) (2008) 259–289
20. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR. (2006)
21. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
22. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI **24**(7) (2002) 971–987
23. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV. (2004)
24. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html