Noname manuscript No.
(will be inserted by the editor)

Tracking Multiple Players using a Single Camera

Horesh BenShitrit · Mirko Raca · François Fleuret · Pascal Fua

Received: date / Accepted: date

Abstract It has been shown that multi-people tracking could be successfullly formulated as a Linear Program to process the output of multiple fixed and synchronized cameras with overlapping fields of view. In this paper, we extend this approach to the more challenging single-camera case and show that it yields excellent performance, even when the camera moves.

We validate our approach on a number of basketball matches and argue that using a properly retrained people detector is key to producing the probabilities of presence that are used as input to the Linear Program.

Keywords People Tracking · Multi Target Tracking · Monocular Videos

1 Introduction

Early approaches to finding people in images tended to rely on frame-to-frame tracking, which involves predicting the pose in a frame given an estimate in the previous one. The emphasis has now shifted to tracking-by-detection in which people are detected in individual frames and the detections then linked across time, which prevents drift and provides robustness to occasional failures.

Most state-of-the-art approaches follow this tracking-by-detection paradigm and operate on graphs whose nodes can either be those where a detector has fired [27,17], or short temporal sequences of consecutive detections that are very likely to correspond to the same person [22,32,1,24,3]. On average, they are much more robust than the earlier tracking methods but typically require the careful setting of edge costs in the graph, the introduction of special purpose nodes to

This work was funded in part by the Swiss National Science Foundation.

H. BenShitrit, M. Raca, and P. Fua
EPFL, Lausanne, Switzerland
E-mail: FirstName.LastName@epfl.ch

F. Fleuret
IDIAP, Martigny, Switzerland
E-mail: FirstName.LastName@idiap.ch

handle occlusions, and an assumption that the appearance of people remains both unchanged and discriminative from frame to frame.

In the multi-camera case and when background subtraction provides usable information about people's location in individual images, we have shown in earlier work [5] that operating on a graph whose nodes are all the spatio-temporal locations where somebody could potentially be, many of these limitations can be removed. This yields a formulation that is robust over very long sequences and requires very few parameters.

In this paper, we extend this approach to the monocular case and demonstrate that it brings the same benefits, namely robustness and simplicity, even in cases where background detection is not an option, for example because the camera moves. However, to achieve satisfactory results it is necessary both to train the people detector for the kind of activities they actually perform and to refine it so that it takes geometric constraints into account. We will demonstrate this in the context of basketball and show that we can approach the performance of a multi-camera system using a single-camera one.

2 Related Work

People tracking is an intensively studied area of research. Most state-of-art approaches rely on a paradigm that has been dubbed "tracking by detection" [8], which implies a two-step process. Typically, a people-detector [9,14,23] is used to find potential people in individual frames and these detections are then grouped into individual trajectories.

Particle filtering [10] has been extensively used to perform this grouping. For example, it has been used to great effect to follow multiple hockey players [21] or to track multiple people in the ground and image planes simultaneously [11].

However, in recent years, this approach has been superseded by one in which detections are first connected into short tracks or *tracklets*, which are then linked together using a higher-level method [22,32,1,24,3]. It derives its power from the fact that, in many cases, consecutive detections can be unambiguously linked and that grouping the resulting tracklets can then be done far more reliably than grouping individual detections.

However, while yielding good results in many situations, these tracking-by-detection methods rely on an ad-hoc mathematical formulation, which does not guarantee convergence to a global optimum. They are therefore prone to mistakes such as identity switches. In our own earlier work [5], we showed that reformulating the linking step as a constrained flow optimization results in a convex problem that fits into a standard Linear Programming framework. It can be solved very efficiently using the k-shortest paths algorithm [28], which yields real-time performance on realistically-sized problems. Our method does not present any of the limitations mentioned above, nor does it require an appearance model. We demonstrated excellent performance with respect to the-state-of-the-art in the multi-camera case [6] and our approach has served as a reference in several recent papers.

In this paper, we show that the same framework can be used to good effect in the single-camera case, even when that camera is moving and we cannot rely anymore on background subtraction as we did before. We use basketball sequences that

involve severe occlusions and sudden motions to demonstrate our approach. This specific application has been tackled in a recent paper [20]. As in our own work, this involved training a Deformable Part Model [14]. However, this was done by manually cropping positive and negative examples from a specific match whereas our approach to training is completely automated and applies across matches.

3 Approach

In our earlier multi-camera work [5, 26], we assumed that the ground plane was represented by a discrete grid and that, at each time step over a potentially long period of time, we were given as input a Probabilistic Occupancy Map [15] (POM) containing probabilities of presence of people in each grid cell, which were generated using multiple cameras. Inferring trajectories from these potentially noisy POMs was formulated as a Linear Program, which we solved using the K-Shortest Paths algorithm (KSP) [5].

In this paper, we replace the POM maps by the output of a people detector [13] which we modify for our purpose and adapt the KSP algorithm to process this new output most efficiently. In the remainder of this section, we first summarize the original Linear Programming Formalism and then describe our modifications.

3.1 Multi Target Tracking as Linear Programming

As in [5], we model people's trajectories as continuous flows going through an area of interest.

More specifically, we discretize it into K grid locations, and the time interval into T instants. For any location k, let $\mathcal{N}(k) \subset \{1, \ldots, K\}$ denote its neighborhood, that is, the locations a person located at k at time t can reach at time $t+1$. To model occupancy over time, let us consider a labeled directed acyclic graph with $K \times T$ vertices such as the one depicted by Fig. 1(a), which represents every location at every instant. Its edges correspond to admissible motions, which means that there is one edge $e^t_{i,j}$ from (t, i) to $(t+1, j)$ if, and only if, $j \in \mathcal{N}(i)$. Note that to allow people to remain static, we have $\forall i, i \in \mathcal{N}(i)$, hence there is always an edge from a location at time t to itself at time $t+1$.

As shown in Fig. 1(b), each vertex is labeled with a discrete variable m^t_i standing for the number of persons located at i at time t. Each edge is labeled with a discrete variable $f^t_{i,j}$ standing for the number of persons moving from location i at time t to location j at time $t+1$. For instance, the fact that a person remains at location i between times t and $t+1$ is represented by $f^t_{i,i} = 1$.

In general, the number of people being tracked may vary over time, meaning some may appear inside the tracking area and others may leave. Thus, we introduce two additional nodes υ_{source} and υ_{sink} into our graph. They are linked to all the nodes representing positions through which people can respectively enter or exit the area, such as borders of the camera field of view. In addition, a flow goes from υ_{source} to all the nodes of the first frame to allow, the presence of people anywhere in that frame, and reciprocally a flow goes from all the nodes of the last frame to υ_{sink}, to allow for people to still be present in that frame. In the case of a small area of interest that can be modeled using only three locations, this yields the

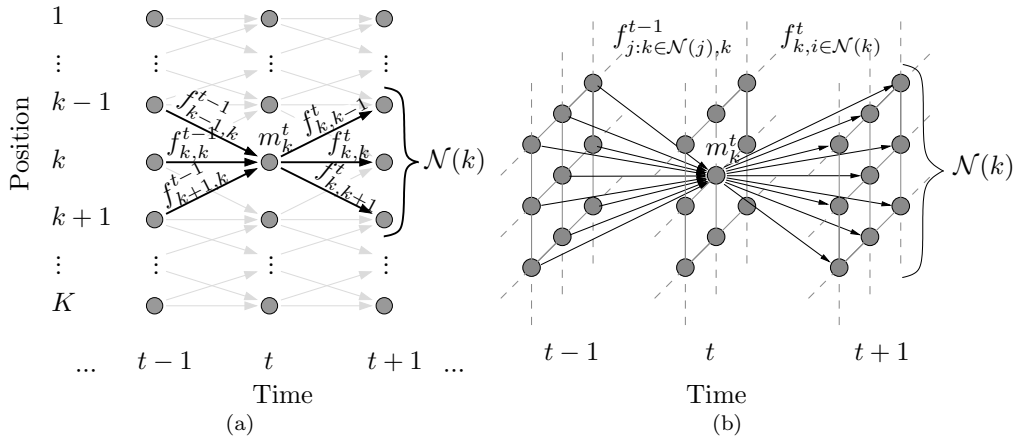


Fig. 1 Directed Acyclic Graph and corresponding flows. (a) Positions are arranged on one dimension and edges created between vertices corresponding to neighboring locations at consecutive time instants. (b) Basic flow model used for tracking people moving on a 2D grid. For the sake of readability, only the flows to and from location k at time t are printed.

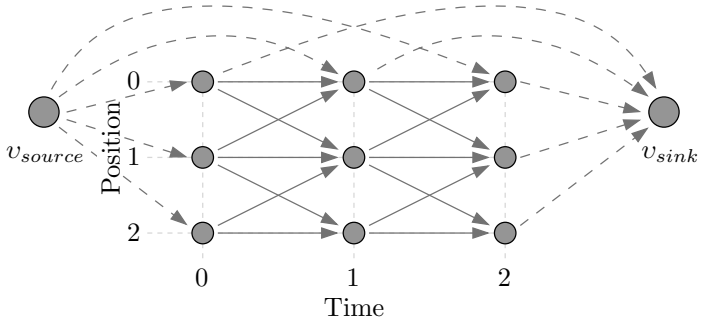


Fig. 2 Complete graph for a small area of interest consisting only of 3 positions and 3 time frames. Here, we assume that position 0 is connected to the virtual positions and therefore a possible entrance and exit point. Flows to and from the virtual positions are shown as dashed lines while flows between physical positions are shown as solid lines.

directed acyclic graph depicted by Fig. 2. v_{source} and v_{sink} are *virtual locations*, because, unlike the other nodes of the graph, they do not represent any physical place.

Under the constraints that people may not enter or leave the area of interest by any other locations than those connected to v_{sink} or v_{source} and that there can never be more than one single person at each location, it was shown in [5] that

the flow with the maximum a posteriori probability is the solution of

$$\begin{aligned}
\text{Maximize} \quad & \sum_{t,i} \log\left(\frac{\rho_i^t}{1-\rho_i^t}\right) \sum_{j\in\mathcal{N}(i)} f_{i,j}^t \\
\text{subject to} \quad & \forall t,i,j, \;\; f_{i,j}^t \geq 0 \\
& \forall t,i, \quad \sum_{j\in\mathcal{N}(i)} f_{i,j}^t \leq 1 \\
& \forall t,i, \quad \sum_{j\in\mathcal{N}(i)} f_{i,j}^t - \sum_{k:i\in\mathcal{N}(k)} f_{k,i}^{t-1} \leq 0 \\
& \qquad \sum_{j\in\mathcal{N}(\upsilon_{\text{source}})} f_{\upsilon_{\text{source}},j} - \sum_{k:\upsilon_{\text{sink}}\in\mathcal{N}(k)} f_{k,\upsilon_{\text{sink}}} \leq 0 \;\; .
\end{aligned} \tag{1}$$

where ρ_i^t is the probability that someone is present at location i at time t returned by a people detector, such as POM [15].

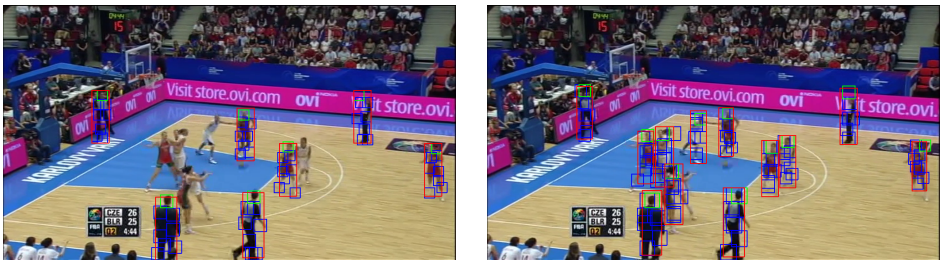
3.2 Creating the Probability Occupancy Maps

In the multi-camera version of our algorithm [5], the probability of presence ρ_i^t at location i at time t of Eq. 1 was computed a generative model operating on the output of a background subtraction algorithm [15]. As we will see in the results section, this approach can still be used in the monocular case but only if the camera is static so that background subtraction remains effective.

To handle a moving camera, we replaced background subtraction by the output of a state-of-the-art people detector, known as a Deformable Part Model (DPM) [13], which has been consistently been found to outperform many others in numerous competitions. Nevertheless, we found that directly using its output to estimate the ρ_i^t resulted in poor performances and we had to modify it in the following ways.

Re-training the DPM Model The original DPM model [13] was trained using videos and images of pedestrians whose range of motion is very limited. By contrast and as shown in Fig. 3, the basketball players tend to perform large amplitude motions. Thus, they do not look like typical pedestrians and are often missed. A similar phenomenon was observed in [30, 25] and it was shown that adding synthetic training data could boost performance.

However, creating synthetic data and ensuring that it truly matches the behavior of basketball players, is a cumbersome task. Instead, we used our multi camera setup [26] to acquire additional training data from two basketball matches for which we have multiple synchronized views, which we add to the standard INRIA pedestrian database [9]. By combining background subtraction information with the ability to occasionally read the numbers on the players jerseys and to take into account the color of their uniforms, we can track individual players over a whole period and therefore obtain long trajectories. The system models humans as cylinders of uniform height h_{std} defined by their positions in the 2D ground place. We take their image projections to be the bounding boxes for our detected humans. Of course, real players are either shorter or taller than the standard height but



Vanila DPM Trained DPM (ours)

Fig. 3 Detection results of the DPM trained using only the INRIA pedestrian database (left) vs. our retrained DPM (right). In both cases, we use the same parameters at run-time and obtain clearly better results with the retained DPM.

the system is very robust to such deviations. Furthermore, because the players are positioned on the ground plane, we know exactly who is in front of who in any given view, and consider a player to be un-occluded in a specific camera view when his bounding box is fully visible and does not intersect that of any other player or referee.

We use the bounding boxes corresponding to these un-occluded players as positive examples and images of empty courts as negative ones. However, clean frames of the empty court are not always available because they are of very little interest to the broadcaster who acquired the images. Since this is the case for the data in our possession, we created a virtual empty court by taking a short video sequence in which the players are moving and creating a color histogram for each pixel. The dominant color is then taken to be the background one. In fact, we used the same approach to learn the background model for our multi-camera tracking system [26].

This yields a fully automated retraining procedure, which could be duplicated for any other sport or activity that can be filmed using multiple cameras.

Imposing Geometrical Constraints. It has often been shown [31,7] that imposing geometrical consistency constraints on the output of a people detector significantly improves detection accuracy. In the case of basketball, we can use the court markings to accurately compute the camera intrinsic and extrinsic parameters [29].

This allows us to reject all detections that are clearly out of our area of interest, the court in this case. In addition, people's head typically are dominant features that tend to be easier to detect and locate accurately than other body parts, as can be seen in Fig. 4. We use this feature of the DPM to estimate the height corresponding to each candidate person and we keep only those that appear to be between heights $h_{\min}$ and $h_{\max}$. In all the experiments reported in the following section, we assign them the values listed in Table. 1 as do we for all other parameters introduced below.

Adapted Non-Maximum Suppression. Non-Maximum Suppression (NMS) is widely used to post-process the output of object detectors that rely on a sliding window search. This is necessary because their responses for windows translated by numbers of pixels are virtually identical, which usually results in multiple detections for a single person. In the specific case of the DPM we use, the head usually is the

most accurately detected part and, in the presence of occlusions, it is not uncommon for detection responses to correspond to the same head but different bodies. In our NMS procedure, we therefore first sort the detections based on their score. We then eliminate all those whose head overlaps by more by a fraction larger than τ_{head} with that of a higher scoring one or whose body overlaps by more than τ_{body}.

A further refinement is to compute the overlap not in terms of intersection of rectangular bounding boxes as is often done but, instead, in terms of overlapping ellipses as was proposed in [2] to handle occlusions. We take the two ellipse diameters to be the height and width of the rectangular bounding box produced by the DPM and have observed that this leads to an increase in performance, especially when players occlude each other.

From Detections to Maps Given a set of reliable detections, which correspond to players on the court, we create an occupancy map by projecting their bounding boxes to the ground. We project the middle point of the top of the bounding box to the ground, which, as discussed before, is assumed to be at height h_{std} above the ground. It is worth noting, that in the case of the DPM detector, the top of the head is usually correctly placed as opposed to the legs that often are wrongly placed. As at this point, most of the detections are reliable, we fill the occupancy map with a fixed score of high probability $\rho_{\text{detection}}$. The occupancy probability at locations where no one as been detected are set to a low value $\rho_{\text{no\_detection}}$ to account for the fact that the detector could have failed to detect somebody who was actually there. For all our experiments, we set $\rho_{\text{detection}}$ and $\rho_{\text{no\_detection}}$ to the values listed in Table. 1. Note, that these probabilities could also be learned in an automated fashion given sufficient amounts of multi-camera video sequences.

3.3 Modifying the Directed Graph

In our earlier work [5], the KSP tracker received as input very reliable detections from our multi-camera algorithm using background subtraction results as input [15]. It was rare for a player to be occluded in *all* views and occlusions in specific ones therefore had little impact. It therefore made sense to build graphs such as the one of Fig. 2, which force trajectories to begin and end at the boundary of the area of interest.

However, in the single-camera case, this is not true anymore. Occlusions become significant and a player may be missed for several frames in a row. Furthermore, even after retraining, the DPM detector remains more sensitive to unusual poses than background subtraction, which may also result in consecutive missed detections. In theory, this could result in whole segments of trajectories being lost because the player only begins to be seen once he is already in the middle of the court. In such cases, hypothesizing trajectory segments across consistently low probability graph nodes to connect high-probability ones to the borders can be more expensive than simply ignoring them. In practice, what happens even more often is that trajectory fragments corresponding to different players can be mistakenly connected to form trajectories that start and end of the area of interest. They would then have to be broken up as part of a post-processing step if one wished player identity to be preserved along individual trajectories.

Parameter	Role	Value
h_{std}	Standard height of a player	1.85m
h_{min}	Maximum distance of head above floor	2.20m
h_{max}	Minimum distance of head above floor	1.60m
$\rho_{\text{detection}}$	Probability of presence if detector has fired	0.97
$\rho_{\text{no\_detection}}$	Probability of presence if detector has not fired	0.03
w_{penalty}	Weight of edges connecting virtual to physical locations	+5
τ_{body}	Threshold of body overlap in NMS	0.75
τ_{head}	Threshold of head overlap in NMS	0.75

Table 1 Numerical values of the parameters defined in Sections 3.2 and 3.3 used to produce all the results of Section 4.

To avoid this, we connect *all* the nodes of our graph to the v_{sink} and v_{source} virtual locations of Fig. 2 so that a trajectory can begin or end anywhere. Unlike edges connecting the exit and entry locations to the source and sink whose weight is zero, these new edges are given the high weight w_{penalty} listed in Table. 1. As a result, only several successive detections can be grouped into a trajectory that starts or ends in the middle of the scene. In other words, this produces long tracklets and is similar in spirit to what is done in [24] while still guaranteeing that we find a global optimum of our Linear Program.

In practice, this prevents different players from being lumped into the same trajectory, or even missed, at only marginal increase in computational costs. It may result in disconnected trajectory fragments for the same player but they can then be reconnected on the basis of appearance [26].

4 Results

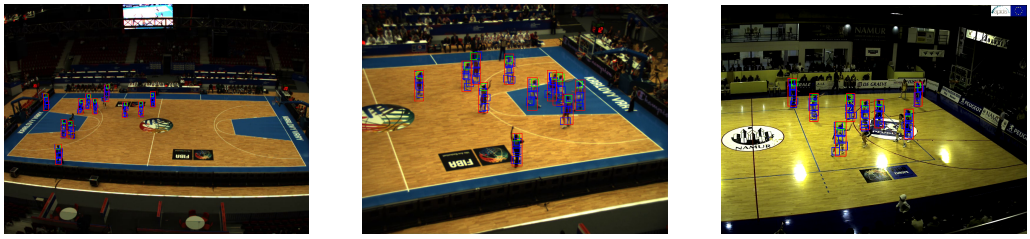
We first validate our results on long basketball sequences acquired using several fixed and synchronized cameras so that we can compare our single-camera results against those obtained with our earlier multi-camera approach [5]. The latter is a meaningful reference because it has been shown to perform well when compared to state-of-the-art approaches on the PETS'09 dataset [6].

We then show that the results of this comparison still hold in the more interesting case of a single moving camera, which precludes both the multi-camera approach and the use of background subtraction.

4.1 Fixed Camera

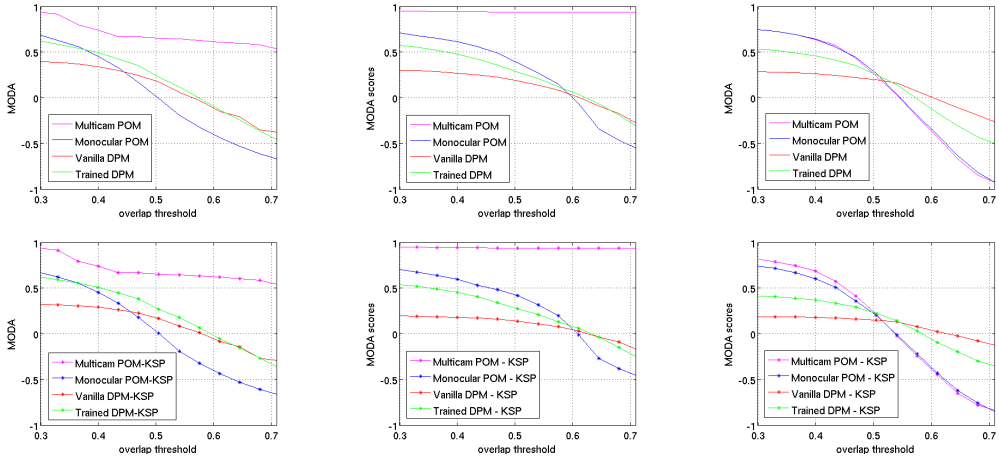
To compare the different approaches, we used two very different datasets:

– The FIBA dataset comprises several multi-view basketball sequences captured matches at the 2010 women's world championship. We manually annotated the court location of the players and the referees on 1000 frames of the Mali vs. Senegal match and 6000 frames of the Czech Republic vs. Belarus match. One frame from each match is depicted on the left and center of Fig. 4. The same ten cameras were used to film all sequences but their locations was changed from match to match.



FIBA (Mali vs. Senegal) FIBA (Czech Republic vs. Belarus) APIDIS

Fig. 4 Representative detection results of our retrained DPM on images acquired using static cameras. We remove detections that correspond to locations outside of the court. The players' bounding boxes are overlaid in red, the head in green, and the remaining body parts in blue. The heads tend to be located much more accurately than the other body parts.



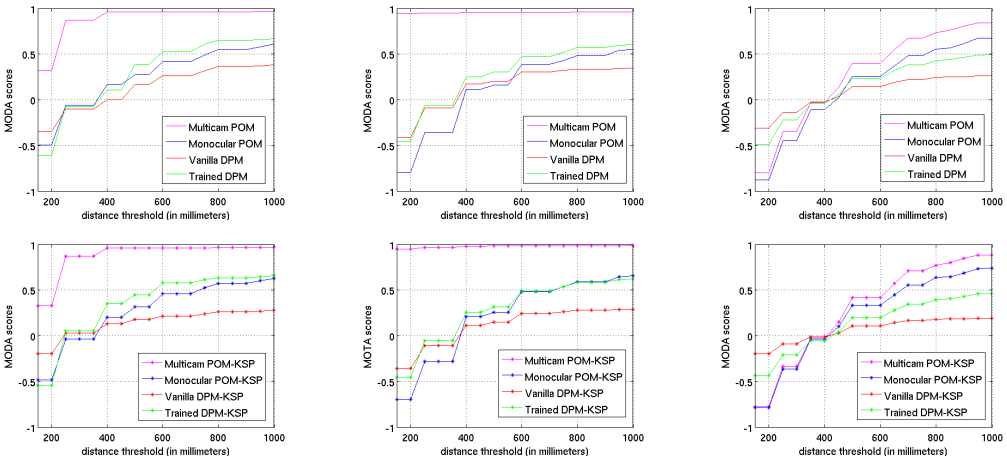
FIBA (Mali vs. Senegal) FIBA (Czech Republic vs. Belarus) APIDIS

Fig. 5 MODA scores for different player detectors, static camera. Multi-camera POM, Monocular POM, DPM trained with INRIA pedestrian dataset, DPM trained with pedestrian and basketball dataset. The MODA scores were calculated with respect to the overlap threshold in the camera view. **Top** Detections alone. **Bottom** Linked detections. The corresponding videos are supplied as supplementary material.

– The APIDIS dataset [4] is a publicly available set of video sequences of a basketball match captured by seven stationary unsynchronized cameras placed above and around the court. It features challenging lightning conditions produced by the many direct light sources that are reflected on the court while other regions are shaded. We present monocular results on the video acquired by Camera #6 that captures half of the court, as can be seen at the right of Fig. 4.

We ran several versions of our detection algorithms on these video sequences:

– **Multicam POM**: Since we have synchronized video sequences from multiple views, we ran our earlier multi-camera approach [5], which relies on background subtraction in each view to compute probability occupancy maps in each frame



FIBA (Mali vs. Senegal) FIBA (Czech Republic vs. Belarus) APIDIS

Fig. 6 Detection results: MODA scores for different player detectors, static camera. Multi-camera POM, Monocular POM, DPM trained with INRIA pedestrian dataset, DPM trained with pedestrian and basketball dataset. The MODA scores were calculated with respect to a distance threshold on the ground plane. **Top** Detections alone. **Bottom** Linked detections.

independently [15]. An alternative would have been to use the approach of [19], which also creates a probability occupancy map from background/foreground likelihood maps.

- **Monocular POM**: The formalism of [15] does not require multiple views setting and also supports the single-camera case, which is what we use to produce probability occupancy maps. They are less peaked than the multi-view ones, however, they can still be used as input to the KSP algorithm of [5].
- **Vanilla DPM**: We use the vanilla DPM algorithm [13] trained on the INRIA pedestrian database [12], which is the way it is often used the in Computer Vision literature, to instantiate our probability occupancy maps as described in Section 3.2. They are then fed to the modified KSP algorithm, as also described in Section 3.2.
- **Trained DPM**: We replace the vanilla DPM algorithm by one we have re-trained, as described in Section 3.2. To this end, we used two matches from the FIBA dataset other than those we used for testing, Czech Republic vs. Australia and Spain vs. Belarus, each of which was filmed using 10 static cameras. We automatically extracted 2000 samples of non-occluded players and referees, and added them as positive examples to the INRIA pedestrian database [9]. In addition, we used 20 images of the empty court, one for each camera in each match to produce negative examples. The DPM was then retrained using this augmented training set and publicly available code [16]. This yields detection results such as those of Fig. 4.

We operate on grids whose cells are $25cm \times 25cm$ and set the parameters introduced in Sections 3.2 and 3.3 to the values given in Table. 1. We will express our results over complete sequences in terms of the standard MODA CLEAR

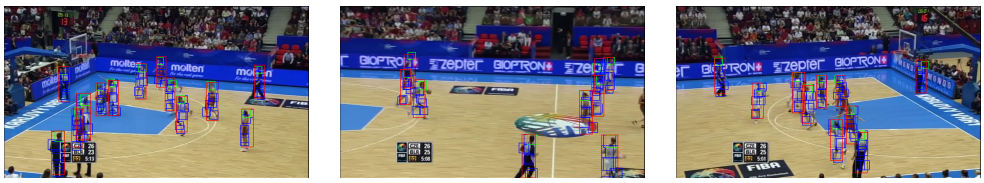


Fig. 7 Representative detection results of our retrained DPM on images acquired using a moving broadcast TV camera. As before, we remove detections that correspond to locations outside of the court.

metric [18], which stands for *Multiple Object Detection Accuracy* and is defined as

$$\text{MODA} = 1 - \frac{\sum_t (m_t + fp_t)}{\sum_t g_t} \ , \tag{2}$$

where g_t is the number of ground truth detections at time t, m_t the number of miss-detections, fp_t the false positive count.

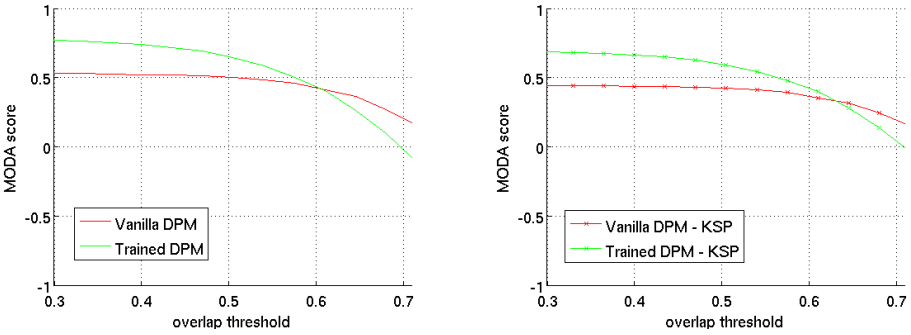
Following standard Computer Vision practice, we decide whether two detections correspond to the same person on the basis of whether the overlap of the corresponding bounding boxes is greater or smaller than a fraction of their area, which is usually taken to be between 0.3 and 0.7. In Fig. 5, we therefore plot our results as functions of this threshold both for the straight output of the people detectors discussed above and for the detections that end up belonging to selected trajectories. Because some valid detections could not be linked into long enough trajectories, the latter is slightly lower than the former in most cases. The corresponding tracking videos are supplied as supplementary material.

Since we operate on the ground plane, an alternative way to compute MODA scores is to pick a distance threshold and consider that a detection corresponds to a ground-truth person if the two are within some Euclidean distance of each other. Fig. 6 depicts the resulting scores as a function of this threshold, both before and after linking as in Fig. 5.

Unsurprisingly, in all cases multicam POM does best and provides an upper limit of what can be done without appearance or motion models. Trained DPM systematically outperforms Vanilla DPM. On the FIBA dataset, it performs similarly to monocular POM, that is, slightly better when using the metric of Fig. 6 and slightly worse when using that of Fig. 5. In other words, we can give up background subtraction without ill-effects, which is essential when dealing with a moving camera as discussed below. On the APIDIS dataset, all methods perform worse than in the FIBA dataset, because of the numerous highlights that disrupt both background subtraction and people detection. In the latter case, it affects the ground-accuracy of the detections and, as result, more of them remain unconnected at linking time.

4.2 Moving Camera

For the FIBA dataset, in addition to the images acquired using static cameras, we have access to those acquired by broadcast TV cameras. As shown in Fig. 7, the



(a) (b)

Fig. 8 MODA scores for a sequence acquired using a moving camera, which makes it impractical to rely on background subtraction and use either versions of POM. We therefore only plot results obtained using either the original DPM or the retrained one as functions of the bounding-box overlap value used to decide whether two detections correspond to the same person. As in Fig. 5, the MODA scores were calculated with respect to the overlap threshold in the camera view. (a) Detections alone. (b) Linked detections. The corresponding videos are supplied as supplementary material.

camera moves to follow the action. Out of a sequence of 1000 images from the Czech Republic vs. Belarus match, we calibrated 10 keyframes by manually supplying a few interest points such as the corners of the court. We then established SIFT correspondences between each of the remaining frames and the closest keyframe and used these to calibrate those frames as well. The resulting camera models are not particularly accurate but nevertheless sufficient enforce geometrical constraints on the detections prior to Non-Maxima Suppression. For evaluation purposes, we also manually annotated 500 frames.

In this case, we cannot use either Multicam POM or Monocular POM since we have a single moving camera, which precludes the use of background subtraction. As shown in Fig. 8, retraining the DPM detector and using the same parameters as before brings about a substantial performance improvement. Furthermore, the MODA scores we obtain is better than the ones for the static cameras and depicted by Fig. 5, mostly because the cameraman is zooming on the action thereby increasing the resolution.

5 Conclusions

We have demonstrated that we can achieve excellent people tracking accuracy from a single video sequence by formulating the problem as an Integer Program on sequences of probability of occupancy maps. The key to good performance is to properly train the people detector that creates the required maps for the specific activity the subjects are engaging in and the specific range of body poses it entails. We have shown how to do this in the specific case of basketball but the approach we advocate is generic and would apply to many other activities.

The current approach does not take people's appearance into account to produce the trajectories, which is both good and bad. It is good because it can handle similar-looking people such as basketball teammates wearing the same uniform and

bad because someone's trajectory can be broken into several disconnected fragments. Future research will therefore focus in incorporating appearance cues to overcome this problem.

References

1. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D Pose Estimation and Tracking by Detection. In: CVPR (2010)
2. Andriyenko, A., Roth, S., Schindler, K.: An Analytical Formulation of Global Occlusion Reasoning for Multi-Target Tracking. In: ICCV Workshop (2011)
3. Andriyenko, A., Schindler, K., Roth, S.: Discrete-Continuous Optimization for Multi-Target Tracking. In: CVPR (2012)
4. APIDIS European Project FP7-ICT-216023: (2008–2010). "www.apidis.org"
5. Berclaz, J., Fleuret, F., Türetken, E., Fua, P.: Multiple Object Tracking Using K-Shortest Paths Optimization. PAMI **33**, 1806–1819 (2011)
6. Berclaz, J., Shahrokni, A., Fleuret, F., Ferryman, J., Fua, P.: Evaluation of Probabilistic Occupancy Map People Detection for Surveillance Systems. In: PETS, pp. 117–124 (2009)
7. Bimbo, A.D., Lisanti, G., Masi, I., Pernici, F.: Person Detection Using Temporal and Geometric Context with a Pan Tilt Zoom Camera. In: ICPR (2010)
8. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online Multi-Person Tracking-By-Detection from a Single Uncalibrated Camera. PAMI **99** (2010)
9. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR (2005)
10. Deutscher, J., Blake, A., Reid, I.: Articulated Body Motion Capture by Annealed Particle Filtering. In: CVPR, pp. 2126–2133 (2000)
11. Du, W., Piater, J.: Multi-Camera People Tracking by Collaborative Particle Filters and Principal Axis-Based Integration. In: ACCV, pp. 365–374 (2007)
12. Ellis, A., Shahrokni, A., Ferryman, J.: Pets 2009 and Winter Pets 2009 Results, a Combined Evaluation. In: PETS (2009)
13. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. PAMI (2009)
14. Felzenszwalb, P., Mcallester, D., Ramanan, D.: A Discriminatively Trained, Multiscale, Deformable Part Model. In: CVPR (2008)
15. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-Camera People Tracking with a Probabilistic Occupancy Map. PAMI **30**(2), 267–282 (2008)
16. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/
17. Jiang, H., Fels, S., Little, J.: A Linear Programming Approach for Multiple Object Tracking. In: CVPR, pp. 744–750 (2007)
18. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Boonstra, M., Korzhova, V., Zhang, J.: Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol. PAMI **31**(2), 319–336 (2009)
19. Khan, S., Shah, M.: A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint. In: ECCV, pp. 133–146 (2006)
20. Lu, W.L., Ting, J.A., Murphy, K.P., Little, J.J.: Identifying Players in Broadcast Sports Videos Using Conditional Random Fields. In: CVPR (2011)
21. Okuma, K., Taleghani, A., de Freitas, N., Little, J., Lowe, D.: A Boosted Particle Filter: Multitarget Detection and Tracking. In: ECCV (2004)
22. Perera, A., Srinivas, C., Hoogs, A., Brooksby, G., Wensheng, H.: Multi-Object Tracking through Simultaneous Long Occlusions and Split-Merge Conditions. In: CVPR, pp. 666–673 (2006)
23. Pirsiavash, H., Ramanan, D.: Steerable Part Models. In: CVPR (2012)
24. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects. In: CVPR (2011)
25. Pishchulin, L., Jain, A., Mykhaylo, A., Thormaehlen, T., Schiele, B.: Articulated People Detection and Pose Estimation: Reshaping the Future. In: CVPR (2012)
26. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking Multiple People Under Global Apperance Constraints. In: ICCV (2011)

27. Storms, P.P.A., Spieksma, F.C.R.: An LP-Based Algorithm for the Data Association Problem in Multitarget Tracking. Computers & Operations Research **30**(7), 1067–1085 (2003)
28. Suurballe, J.W.: Disjoint Paths in a Network. Networks **4**, 125–145 (1974)
29. Tsai, R.: A Versatile Cameras Calibration Technique for High Accuracy 3D Machine Vision Metrology Using Off-The-Shelf Tv Cameras and Lenses. JRA **3**(4), 323–344 (1987)
30. Yu, J., Farin, D., Krueger, C., Schiele, B.: Improving person detection using synthetic training data. In: ICIP (2010)
31. Yuan, L., Bo, W., Nevatia, R.: Human Detection by Searching in 3D Space Using Camera and Scene Knowledge. In: ICPR (2008)
32. Zhang, L., Li, Y., Nevatia, R.: Global Data Association for Multi-Object Tracking Using Network Flows. In: CVPR (2008)