Multiple People Tracking Based on Dynamic Visibility Analysis

Xinghan Luo, Robby T. Tan, Remco C. Veltkamp

Technical Report UU-CS-2011-010
May 2011

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Multiple People Tracking Based on Dynamic Visibility Analysis

Xinghan Luo, Robby T. Tan, Remco C. Veltkamp

Department of Information and Computing Sciences, Utrecht University
xinghan, robby, Remco.Veltkamp@cs.uu.nl

Keywords: Tracking, principal axis, view visibility

Abstract

Multiple people tracking from multiple cameras benefits many applications in computer vision. However, using the existing methods, various problems such as inter-person occlusions still degrade the position estimations. In this paper, we attempt to solve the problems by analyzing the view visibility and ranking the reliability of the cues from 2D views. We combine the visibility with the smoothness constraints into a probability framework, which offers a more flexible and robust estimation. Aside from that, in this paper, we also introduce 2D reference lines to estimate the 2D position of every person in the input images. These lines are able to estimate more accurate and robust 2D positions. We quantitatively evaluate our method by using both our own multiple-people data set and a public data set. The evaluation and experimental results on the standard data set show that our methods considerably improve the accuracy and the robustness.

1 Introduction

We address the problem of tracking a group of people in indoor environments, by locating their position in a 3D space using multiple views. Compared with monocular approaches, tracking a group of people using multiple cameras that have an overlapping field of view can provide more accurate estimations. The main reason is that, it has more cues from different views and possibilities to integrate those cues. The cues can be color, edge, motion, contour[2, 6, 8, 5, 1]. Most methods in the literature rely on such cues to infer the positions of the target persons. In principal, cues from two views are sufficient to infer a 3D position, however the redundant fusion of cues from all views is expected to produce more flexible and robust estimation.

For tracking moving persons in a group, Mittal and Davis [8] introduced the M2Tracker, which uses a region-based stereo algorithm to find 3D points inside an object, applies Bayesian classification to segment a view given priors, and employs occlusion analysis to combine evidences from all views pairs for the tracking. Hu et al. [5] use the principal axis as the key feature that approximates the symmetric axis of the human body, to find multi-view person correspondences without camera calibration. Kim and Davis [6] combine the principal axis and a person appearance model to segment foreground pixels. The principal axes from all views are intersected on the ground plane to represent the persons' locations.

Recently, Gupta et al. [4] introduce the COST approach applied to a variant of M2Tracker [8]. The probabilistic distribution of the position and appearance of people are used for the visibility and confusion analysis, and cameras sets are selected to minimize the computational cost. Fleuret et al. [3] estimate the probabilities of occupancy of the ground plane at individual time steps with dynamic programming to track people over time. Du and Piater [1] propose to infer target person's states in each camera and in the ground plane by collaborative particle filters. In [2], a probabilistic approach is proposed to integrating multiple

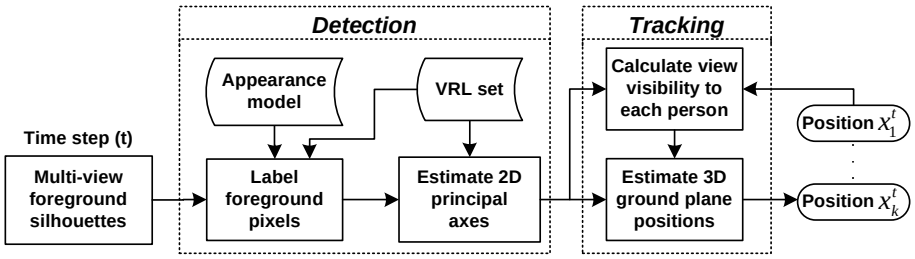

Figure 1: Overview of the components.

cues, and tracking is performed in different cues by interacting processes represented by Hidden Markov Model (HMM). Liem and Gavrila [7] obtain the position measurements by carving out the space defined by foreground regions in the overlapping camera views and projecting them onto blobs on the ground plane. Person appearance is described in terms of the color histograms in the various camera views of three vertical body regions (head-shoulder, torso, legs).

The common problem of tracking multiple people using multiple views is that different cameras provide different information about the location of the same person. This can happen because cues in one or more cameras are affected by occlusions, outliers, etc., causing the estimation to be erroneous. Therefore, integrating all cues from all views (e.g. [2]), in some cases, can lead to inaccurate and less robust estimations. To overcome the problem, we propose a solution that evaluates and ranks the visibility of the views of a target person. We only fuse two views that have higher visibility than the other views, and reject the information from other views. Previous experiments [4] also showed that it is sufficient to infer 3D position using a small number of cues. In addition, we add a smoothness constraint of the motion of the position of each target person, to reject problematic cues.

Estimating persons' position by principal axes is simple and efficient. The previous methods [6, 8, 5, 1] employ lines parallel to the vertical image columns as the principal axes. However, such lines are not equivalent to lines perpendicular to the ground plane in the 3D world. We introduce Vertical Reference Line (VRL) set, which is a set of the 2D correspondences of selected 3D lines perpendicular to the ground floor projected on the image planes. To estimate the persons' principal axes in image planes, the previous methods use the least mean squares [6] or the least median squares [5] distance between a certain pixel and the axis. We acquire foreground pixels based on the VRL set and combine both the appearance and geometric consistencies to infer principal axes of persons.

We combine the visibility cues and smoothness constraint into an MAP probability framework to decide every person's position from a set of position candidates, which are generated by intersecting the principal axes from two views or two image planes. See Fig.1 for the pipeline of our approach.

2 General Probability Framwork

In this section we discuss the general probability framework of our method, which validates the position candidates generated from two views using visibility and motion smoothness as the main criteria.

Given the estimated positions of the N persons in the previous time step $t-1$, $\{x_k^{t-1}\}_{k=1}^{N}$, including the position x_k^{t-1} of target person k, and the positions $\{x_s^{t-1}\}_s^{N-1}$ of the other person s, and the foreground pixels d_i^t ($i \in 1...L$) of each view i ($i \in 1...L$)in the current time step t, our basic algorithm is as follows:

1. Estimate the principal axes $\{y_{k,i}^t\}_{i=1}^{L}$(VRL set, Section 3.2) of person k in all views according to the function $\{y_{k,i}^t\} = h(d_i^t)$ ($i \in 1...L$), where h is based on the appearance model, VRL set and the foreground pixels. (section 3)

2. Generate position candidates $\{x_{k,j}^t\}_{j=1}^{m} = f(\{y_{k,i}^t\}_{i=1}^{L})$, which are the intersection points of VRL from all m stereo view pairs.

3. Obtain the estimated position $(x_k^t)^* \in \{x_{k,j}^t\}_{j=1}^{m}$, that maximizes the posterior probability $P(x_k^t|\{x_s^{t-1}\}_s^{N-1}, x_k^{t-1})$ which is proportional to $P(\{x_s^{t-1}\}_s^{N-1}|x_k^t)P(x_k^t|x_k^{t-1})$.

$$(x_k^t)^* = \arg \max_{\{x_{k,j}^t\}_{j=1}^m} P(\{x_s^{t-1}\}_s^{N-1}|x_{k,j}^t)P(x_{k,j}^t|x_k^{t-1})$$
$$\propto \arg \max_{\{x_{k,j}^t\}_{j=1}^m} \lambda \log P(\{x_s^{t-1}\}_s^{N-1}|x_{k,j}^t) + (1-\lambda) \log P(x_{k,j}^t|x_k^{t-1}) \quad (1)$$

where, the likelihood $P(\{x_{N-1}^{t-1}\}|x_{k,j}^t)$ denotes the probability of person k is well visible in the view pair indexed by j, while the candidate position and previous location of the other persons are known. We define this likelihood to be proportional to the joint visibility of each view pair (Section 4.1). The prior $P(x_{k,j}^t|x_k^{t-1})$ denotes the smoothness constraint, that gives a higher probability to the candidates close to the previously estimated position, and a lower probability to non-smooth motions (Section 4.2). The parameter λ is the weighting coefficient between the visibility and smoothness constraints, which is set to 0.6 in our experiments to strengthen the factor of view pair visibility. The final estimated position of each person in the current time will be the one that has a high visibility value and is relatively close to the previous position.

3 Person Detection in 2D

Having the foreground regions (silhouettes), the aim of this section is to discuss how to estimate the 2D principal axes in 2D images. We propose several basic steps: (1) computing the appearance model of each person; (2) labeling foreground pixels from both the appearance model and VRL set (Section 3.2); (3) estimating the 2D principal axes from the labeled foreground pixels and VRL set. The details of each step are as follows.

3.1 Person Appearance Model

The Gaussian kernel based KDE as in [6, 8] is used to model each person's whole-body appearance in HSV color space. To avoid the influence of the illumination differences between views and color-calibrate of cameras, for each person we combine all pixels from all views.

3.2 Vertical Reference Line Set

The 2D principal axis of a person is commonly represented by the vertical symmetric axis of the body perpendicular to the image's horizontal axis (e.g. [6, 5, 9]). However we found this inaccurate, particularly when we intend to use the 2D principal axes to determine a 3D vertical line in the world coordinate that approximately represents a person. To improve the vertical symmetric axis, we incorporate the information of the 3D world in our method (Fig.2).

We set further a constraint in the principal axis of a person by creating 3D reference lines that perpendicular to the ground plane (Fig.2 (a), colored lines) and on the line (Fig.2 (a), magenta line on the ground plane) parallel to the mapping line of a camera image plane on the ground plane (Section 4.2). We project these lines onto the image plane, producing lines we call a 'Vertical Reference Line' (VRL) set, which is shown in Fig.2 (b). For each view, the VRL set is setup by the following steps:

1. Select a common reference point (e.g. the world origin) on the ground plane that is visible from all cameras. See the black dot in Fig.2 (a).

2. Create a line that passes through the common reference point and is parallel to the intersection line of the camera's image plane to the ground plane. See in Fig.2 (a), the magenta line located in the middle of Fig.2 (a) where the colored vertical lines stand on it.

3. Generate 3D vertical lines on top of the line created in step 2, with uniform distance to each other. For example, in Fig.2 (a), the lines are represented green, blue, and red lines in the middle of the figure.

4. Compute the VRL set by projecting the 3D vertical lines onto the image planes of all the views, see the gray lines in Fig.2 (b). Apparently these 2D projection lines are generally neither parallel to the image's vertical-axis nor to each other. And the number of VRL lines on the view determines the resolution of the principal axes candidates.

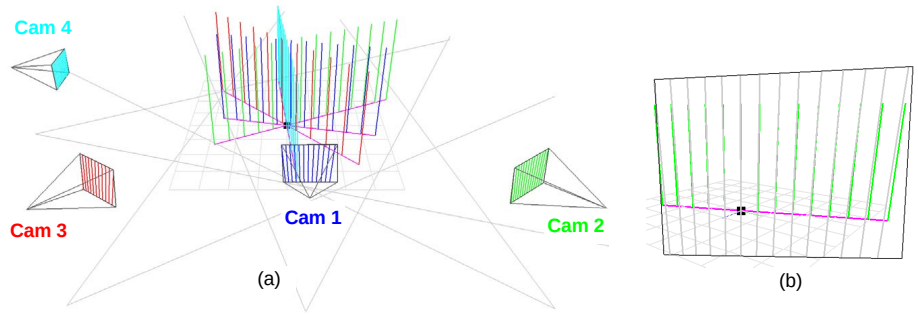

Figure 2: VRL set (for better visualization, out of 20 lines only 1 representative VRL line is shown). (a) 3D vertical lines and VRL sets for a 4 views setup (best view in color). Reference point (black), reference line(Magenta lines) on the ground plane, 3D reference lines perpendicular to the ground plane (colored with respect to views) and corresponding VRL set on camera image plane. (b) Viewing 3D reference lines (green lines) from the projection center of camera 2 (VRLs on the image plane are gray lines).

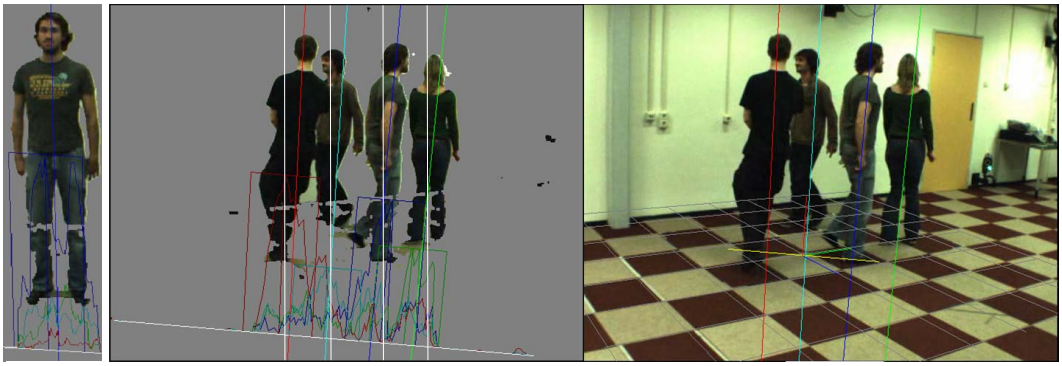

Figure 3: Principal axes estimation on views. (a) Principal axes estimation on example person. (b) Estimate multi-person principal axes with VRL on example view. VRL histogram, evaluation window and refined principal axis are superimposed on the foreground regions. The white vertical lines resembles the lines from image columns, that do not well fit the persons. (c) Multi-person tracking by VRL set-based principal axes.

3.3 Principal Axis Estimation

Having obtained the foreground blobs and the VRL set in the previous steps, we then first label pixels that lie on the VRL. Compare to [5, 6] which require proper segmentation of foreground, sampling the foreground pixels using VRL set is robust given foreground of very poor quality. Because the VRL set ensures the geometric consistency of the sampled pixels in the approximated 3D vertical direction, even if the pixels are from separate foreground pieces of the same person, shown in Fig.3 (a). After the sampling, we assign each sampled pixel a person label that gives the highest probability based on KDE (Section 3.1).

Based on the pixel labels along each VRL, the number of pixels labeled to the same person is summed up to get the histograms of the pixels for the same person, shown in the bottom of Fig.3 (a)(b). Aside from the histograms, we also incorporate the positions of the 2D principal axes of the previous time frame.

3.4 Position Refinement

The initial estimation of the 2D principal axes will tend to be in the right or left side of the person's body central axis, since the histogram's peak often corresponds to the VRL that passes through one of the

person's legs and reach the top of the head. To find a better VRL that is as close as possible to the body central axes, we refine the location of the principal axis by analyzing the distribution of pixels in the left and right side of the initial estimation of the 2D principal axes. We refine a 2D principal axis by balancing the area of the left and right side, which is based on the following equation:

$$p = p' + \frac{1}{ab(p')}\Big(\sum_{r \in right(p')} b(r) - \sum_{l \in left(p')} b(l)\Big) \tag{2}$$

where p is the refined position, p' is the initial position, a is the distance between a VRL to the next VRL, $b(r)$ and $b(l)$ are the counts at r^{th} and l^{th} bins. Functions $left$ and $right$ give the set of bins that belongs to the same person of bp' and are located at the left and right side of p', respectively.

After the position refinement, the final estimated principal axis is most likely to pass through each person's symmetric center, as shown in Fig. 3.

4 Tracking

As illustrated in our pipeline (Fig.1), having estimated the 2D principal axes for all persons in all views, the next step is to estimate the 3D position of each person. This is done by projecting the 2D principal axes onto the 3D ground plane and then computing the intersection points.

However, like in general tracking, the main problem of our tracking is inter-person occlusions. When people in a scene occlude each other with respect to the camera positions, it is likely that the estimation of the occluded persons' position on the view is not correct. To solve this problem, we introduce an algorithm of computing the visibility of the view (Section 4.1) to each person based on analyzing the geometric scene structure of person-person and person-camera relative positions. Estimating the person's position by high visibility views is more reliable due to the comparatively good visibility and least interference from the other persons. However, incorrect estimations solely using a high visibility views are still possible. To reduce the inaccuracy, we introduce the smoothness constraint (Section 4.2) that measures the distance between previously estimated position and current position candidates. The idea is to identify a jump that breaks the assumption of continuous movement. We combine visibility and smoothness constraints into the probability framework as shown in Section 2.

4.1 Computation of View Visibility

The algorithm for computing the view visibility is as follows. First, we calculate the camera's Field Of View (FOV) on the ground plane. To do this, we map each camera projection center and four corners of its image plane to the ground plane. Among the 4 mapping corners, we select a point pair with maximum distance to each other. See in Fig.4 (a), points $I^*_{c,1}, I^*_{c,2}$ define a line that approximates the projection of the camera image plane on the ground plane. Connecting these two points and the camera projection center we get the approximation of the camera's FOV, see the gray lines in Fig.4 (c).

A person inside the camera's FOV is supposed to be visible, however with respect to the camera position, the person can be either partially or completely occluded by any other persons that are in front of him. With the presence of possible occluders, we propose to quantitatively measure the visibility of a person from every camera. Note that, in our framework, the computation of the view visibility is based on all persons' positions in the previous time step. Given the camera positions $\{c_i\}_{i=1}^{L}$ and camera FOVs as constants, the quantitative measurement of the visibility to person k from the camera c_i in an N person group can be expressed as:

$$V^t_{k,i}(x^t_{k,j}, \{x^{t-1}_s\}^Q_s) = \alpha||\frac{D_E(c_f, x^t_{k,j})}{D_E(c_i, x^t_{k,j})}|| + (1-\alpha)||N - Q - 1 + \sum_{s=1}^{q_1}\frac{D_P(x^{t-1}_s), x^t_{k,j}}{D_P(I^{t-1}_{s,i}, x^t_{k,j})} - T\sum_{s=1}^{q_2}\frac{W_{k,s} - D_P(x^{t-1}_s, x^t_{k,j})}{W_{k,s}}|| \tag{3}$$

The visibility equation implies that there are two normalized factors determining the degree of visibility ($V^t_{k,i}$): (1) the person's distance to the camera; (2) the occlusion degree from the other persons. The weight α (where $0 < \alpha < 1$) balances the two factors.

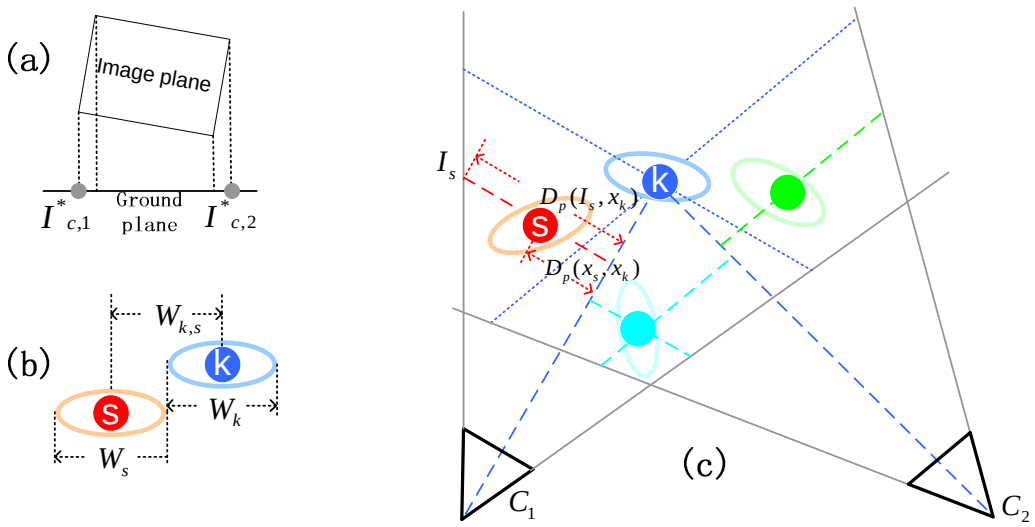

Figure 4: Computation of view visibility. (a) Map camera image plane to ground plane. (b) Compute the occlusion threshold distance $W_{k,s} = \frac{W_k + W_s}{2}$. (c) Compute camera FOV on the ground plane (gray lines), and measure the inter-person occlusion degree by the geometric analysis.

Regarding the first factor, function $D_E()$ measures the Euclidean distance between the candidate position $x^t_{k,j}$ of person k and the camera c_i, and the camera c_f that is the furthest from the person. Since the closer the person to the camera, the more cues (i.e. larger foreground blob) will be available, and the more robust the estimation will be.

Regarding the second factor, the function $D_P()$ denotes that the distance is measured by the line that is parallel to the project line of the image plane of camera c_i on the ground plane, see Fig.4 (c). $W_{k,s}$ is a constant which defines the sum of the estimated half width of the two persons, used as the occlusion threshold, see Fig.4 (b). Based on $W_{k,s}$, the occluder candidates are divided in to two groups: (i) q_1 close neighbor while $D_P(x^t_{k,j}, x^{t-1}_s) > W_{k,s}$; (ii) q_2 occluders while $D_P(x^t_{k,j}, x^{t-1}_s) < W_{k,s}$; $Q = q_1 + q_2$ denotes the total number of occluder candidates that are in the front of the target. See in Fig.4 (c), $D_P(I^{t-1}_{s,i}, x^t_{k,j})$ denotes the distance between the intersection point of the parallel line (defined by the close neighbor) to the FOV and the intersection point of the the parallel line to the camera-target reference line. For evaluation the visibility with inter-person occlusion: (i) All the non-occluders contribute $N-Q-1$, (ii) All the q_1 close neighbors contribute $\sum_{s=1}^{q_1} \frac{D_P(x^{t-1}_s, x^t_{k,j})}{D_P(I^{t-1}_{s,i}, x^t_{k,j})}$, (iii) Crucially, all the q_2 occluders reduce the overall visibility T by $T \sum_{s=1}^{q_2} \frac{W_{k,s} - D_P(x^{t-1}_s, x^t_{k,j})}{W_{k,s}}$, where T numerically equals to N.

We define the occulders' area $O_{k,i}$ as the triangular area formed by camera FOV and the target person. See the triangular areas formed by the gray lines and dotted blue lines in Fig.4 (c).

To have one candidate intersection point we need lines from at least two views. The visibility of the person on two views is therefore set to be the joint visibility as $V_{k,j_1} * V_{k,j_2}$, which is proportional to the likelihood:

$$P(\{x^{t-1}_s\}^{N-1}_s | x^t_{k,j}) \propto V_{k,j_1} * V_{k,j_2} \tag{4}$$

where j_1,j_2 are the index of the views of view pair j.

4.2 Smoothness Constraint

The smoothness of the motion is simply measured by calculating the Euclidean distance from the candidate to the previous estimated position in the exponential function $\exp(-|D_E(x^t_{k,j}, x^{t-1}_k)|)$, which is proportional to the prior:

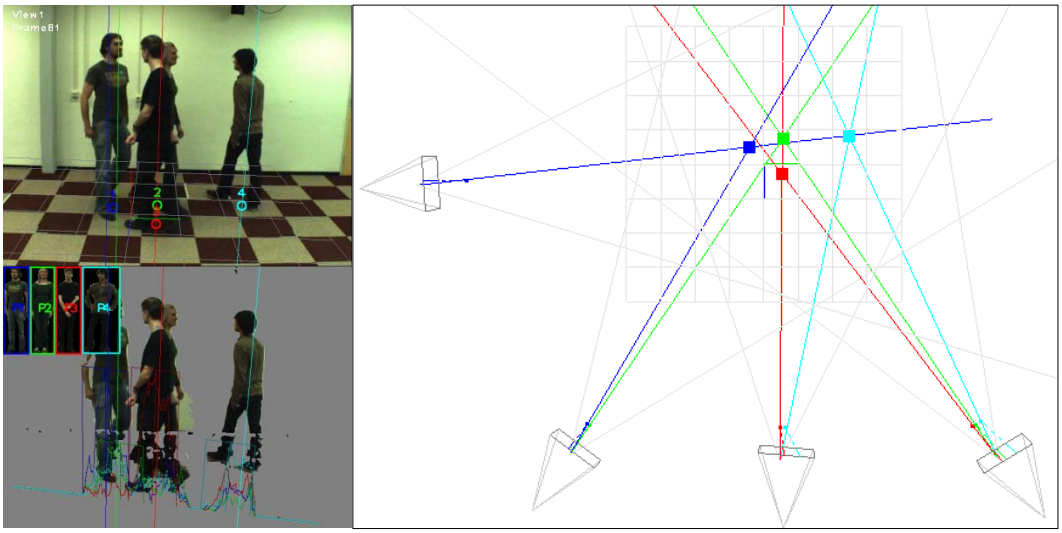

Figure 5: Tracking 4PWALK sequences (best view in color). Top: principal axes estimation in 2D view. Bottom: the estimate 3D positions, the colored lines indicate the view pairs that are selected. The choice is based on view-person visibility and level of motion smoothness.

$$P(x_{k,j}^t | x_k^{t-1}) \propto \exp(-|D_E(x_{k,j}^t, x_k^{t-1})|) \tag{5}$$

which gives low probability for non-smooth motions.

5 Experiments and Evaluations

In our experimentations [1], we focus on the indoor tracking of a group of people in a relatively small scene. We assume the cameras have clear views to the group of people without being blocked by other objects within the acquisition space.

At the initial stage, people are well visible without inter-person occlusions to provide reliable appearance information to estimate principal axes on the views. Person correspondences are found across views and initial 3D positions are estimated for the tracking.

To validate our approaches, we have tested our algorithms with two multi-view sequences: 4PWALK and UMD 4-person LAB sequences. the 30-second 4PWALK sequences (Fig.5) were captured in our lab by four calibrated 644×484 Basler PiA cameras (frontal, left, right and side).

4PWALK: It is a challenging sequence recording four freely walking persons in a dense group with severe inter-person occlusions. Given the manually cropped person appearances, all 4 persons are properly tracked and volumes are recovered in the whole sequences. See Fig.5.

UMD LAB: We also tested our tracking approach on the fifteen-view LAB dataset [4] recording four moving persons. Unlike [8] that use eight cameras or more, for equal setup we selected four cameras (index 00,03,06,12) that placed around the acquisition space. Given the manually cropped person appearances and the initial positions, our approach can properly track the 4 persons of in this challenging video that is set with poor image contrast and severe inter-person occlusions, see Fig.6. Moreover, due to the advantage of employing VRL set built by actual 3D vertical lines as principal axes, the ground positions estimated by our approach is generally closer to the manually marked ground truth of the LAB set. Using video sequences from only 4 cameras, we achieved apparently better tracking accuracies than [4] and [8]. See Fig.7 (a)(b) for the comparison of mean error in position estimation, and mean error standard deviation.

[1] http://people.cs.uu.nl/xinghan/index.html

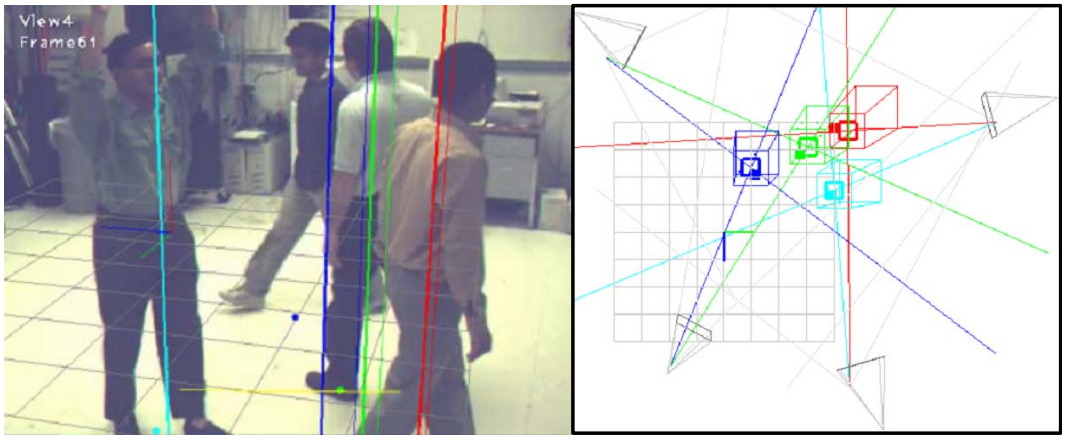

Figure 6: Tracking LAB sequences and performance comparison. (a) The estimated 2D principal axis (thinner line) and 3D position projection lines (thicker lines) in time step 61. (b) The top view scene, with ground truth shown as colored solid squares, and the 3D estimated positions by our approach are bounding boxes.

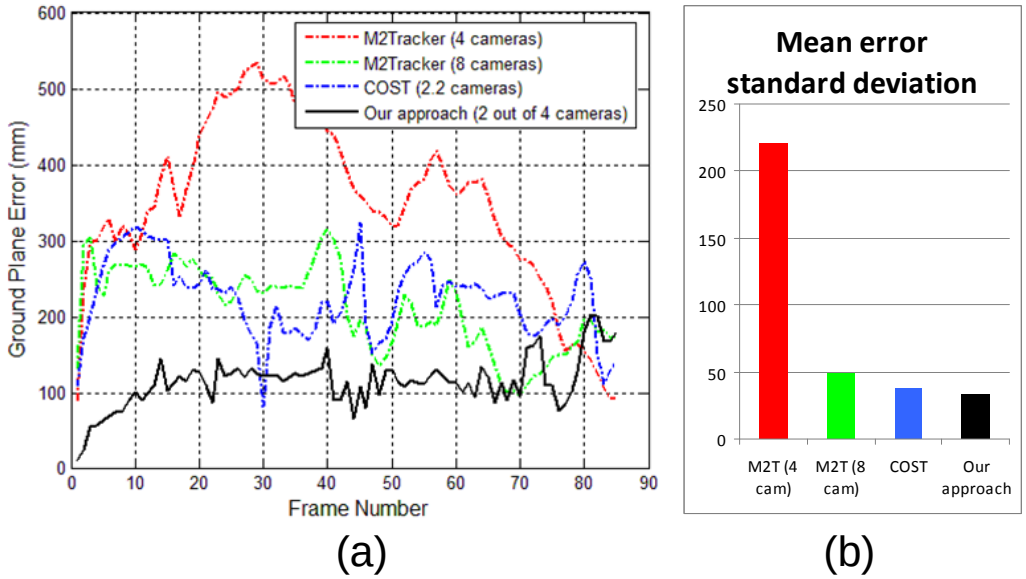

Figure 7: Compare our approach with [4] and [8] by (a) Mean error plot. (b) Mean error standard deviation.

6 Conclusions

We have introduced a geometric analysis-based multi-person tracking framework. Using symmetric body principal axis as the key feature, the persons' positions on views are approximated from foreground silhouettes by the VRL set. A novel view-person visibility evaluation algorithm is proposed to obtain the reliable cues from different views. Persons' 3D ground plane position are estimated within a probability framework that ensures the choice of position candidate is the one generated from the views of higher visibility visibility and smooth motion. In particular, the VRL set fundamentally improves the accuracy of principal axis-based person position estimation. Moreover, the visibility evaluation algorithm benefits not only the multi-person tracking under severe occlusion, but also can be a promising platform for tracking person body parts. Testing on the benchmark sets, our approaches have better accuracies to the previous methods. In the future, we will further extend the propose methods to a more flexible framework, to enable e.g. automatic appearance initialization, handling of people who enter or go out of the scene, etc.

Acknowledgments This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

References

[1] W. Du and J. Piater. Multi-camera people tracking by collaborative particle filters and principal axis-based integration. *ACCV*, pages 365–374, 2007.

[2] W. Du and J. Piater. A probabilistic approach to integrating multiple cues in visual tracking. *ECCV*, pages 225–238, 2008.

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *TPAMI*, 30(2):267–282, 2008.

[4] A. Gupta, A. Mittal, and L. S. Davis. Cost: An approach for camera selection and multi-object inference ordering in dynamic scenes. *ICCV*, pages 1–8, 2007.

[5] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *PAMI*, 28(4):663–671, 2006.

[6] K. Kim and L. S. Davis. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. *ECCV*, pages 98–109, 2006.

[7] M. Liem and D. M. Gavrila. Multi-person tracking with overlapping cameras in complex, dynamic environments. *BMVC*, 2009.

[8] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *IJCV*, 51(3):189–203, 2003.

[9] Y.-T. Tsai, H.-C. Shih, and C.-L. Huang. Multiple human objects tracking in crowded scenes. *ICPR*, 3:51–54, 2006.