MULTI-PERSON TRACKING BASED ON VERTICAL REFERENCE LINES AND DYNAMIC VISIBILITY ANALYSIS

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

Department of Information and Computing Sciences, Utrecht University
PO Box 80.089, 3508 TB Utrecht, The Netherlands

ABSTRACT

Multiple people tracking from multiple cameras can suffer from various problems, particularly from inter-person occlusions. This paper attempts to solve the problems by analyzing the view visibility and ranking the reliability of the cues from 2D views. It combines the visibility with the smoothness constraints into a probability framework, which offers a more flexible and robust estimation. Moreover, it introduces 3D reference lines to estimate the 2D position of every individual in the input images. These lines can estimate more accurate and robust 2D positions. The experimental results and quantitative evaluations on the standard data set show the effectiveness of the method.

Index Terms— Principal axis, vertical reference line, view visibility, multi-person tracking

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, multiple cameras of overlapping Field of View (FOV) provide more cues for tracking multiple people. Most methods in the literature rely on cues like color, edge and motion [1, 2, 3, 4, 5, 6] to infer the positions of the target persons.

The common problem of tracking multiple persons 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 tracking to be erroneous. Therefore, integrating all cues from all views (e.g. [1]) can, in some cases, lead to inaccurate and less robust estimations. To overcome this 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 disregard the information from other views. Previous experiments [7] also showed that it is sufficient to infer 3D position using a small number of cues. In addition, we add a smoothness constraint to the motion of the position of each target person, to reject problematic cues.

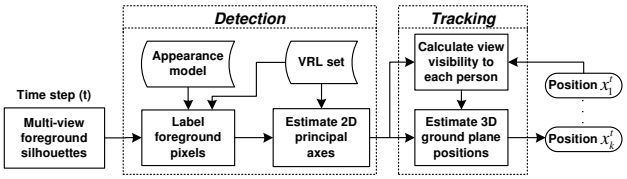

Fig. 1. Overview of the components.

The previous methods [2, 3, 4, 5] 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 the Vertical Reference Line (VRL) set, which consists of the 2D correspondences of selected 3D lines perpendicular to the ground floor projected on the image planes. 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 a probability framework to decide every person's position from a set of position candidates. See Fig.1 for the pipeline of our approach.

The paper is organized as follows. Section 2 shows the general probability framework. Section 3 describes the estimation of persons' principal axes in 2D views. Section 4 discusses the evaluation of view visibility and smoothness constraint. Experimental results are reported in Section 5. Finally, in Section 6, we conclude our paper.

2. FORMULATION

This section discusses the MAP (Maximum a Posteriori) probability framework of the proposed method in validating the position candidates from two views by using visibility and motion smoothness as the main criteria.

Given the estimated positions of N persons in the previous time step $t-1$, stated by $\{x_k^{t-1}\}_{k=1}^{N}$, 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, the basic algorithm is as follows. Note that, we define the foreground pixels as d_i^t of each view i ($i \in 1...L$) in the current time step t.

1. Estimate the principal axes $\{y^t_{k,i}\}^L_{i=1}$(VRL set, Section 3.1) of person k in all views according to the function $\{y^t_{k,i}\} = h(d^t_i)$ $(i \in 1...L)$, where h is based on the appearance model, VRL set and the foreground pixels.

2. Generate position candidates $\{x^t_{k,j}\}^m_{j=1} = f(\{y^t_{k,i}\}^L_{i=1})$, which are the intersection points of VRL from all m stereo view pairs.

3. Obtain the estimated position $(x^t_k)^* \in \{x^t_{k,j}\}^m_{j=1}$, that maximizes the posterior probability $P(x^t_k|\{x^{t-1}_s\}^{N-1}_s, x^{t-1}_k)$ which is proportional to $P(\{x^{t-1}_s\}^{N-1}_s|x^t_k)P(x^t_k|x^{t-1}_k)$.

$$
\begin{aligned}
(x^t_k)^* &= \arg \max_{\{x^t_{k,j}\}^m_{j=1}} P(\{x^{t-1}_s\}^{N-1}_s|x^t_{k,j})P(x^t_{k,j}|x^{t-1}_k) \\
&\propto \arg \max_{\{x^t_{k,j}\}^m_{j=1}} \lambda \log P(\{x^{t-1}_s\}^{N-1}_s|x^t_{k,j}) \\
&\quad +(1-\lambda) \log P(x^t_{k,j}|x^{t-1}_k) \quad (1)
\end{aligned}
$$

where the likelihood $P(\{x^{t-1}_s\}^{N-1}_s|x^t_{k,j})$ 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 $P(x^t_{k,j}|x^{t-1}_k)$ 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. 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; (3) estimating the 2D principal axes from the labeled foreground pixels and VRL set. The details of each step are as follows.

3.1. Vertical Reference Line Set

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 Fig.2 (a), the black dot in the center of the ground plane.

2. Create a line that passes through the common reference point and is parallel to the projection line of the camera's image plane (Section 4.1) to the ground plane. See Fig.2 (a), the magenta line on the ground plane.

3. Generate 3D vertical lines on top of the line created in step 2, with uniform distance to each other. See Fig.2 (a), the colored 3D vertical lines.

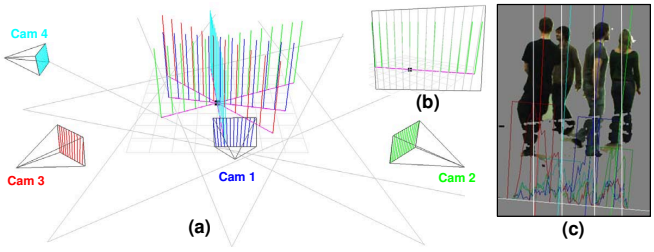

Fig. 2. VRL set and principal axis example (best view in color). (a) 3D vertical lines and VRL sets for a 4 views setup. (b) Viewing 3D reference lines (green lines) from the projection center of camera 2 (VRLs on the image plane are gray lines). (c) VRL histogram, evaluation window and refined principal axis are superimposed on the foreground regions. The white lines represent image columns.

4. Compute the VRL set by projecting the 3D vertical lines from step 3 onto the image planes of all the views. See Fig.2 (b), the gray lines on the image plane. The number of VRL lines determines the resolution of the principal axes candidates.

3.2. 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. Compared with [4, 2] which require proper segmentation of foreground, sampling the foreground pixels using VRL set is robust against poor quality foregrounds. 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 regions of the same person, shown in Fig.2 (c). The Gaussian kernel based KDE as in [2, 3] is used to model each person's whole-body appearance in HSV color space. After the sampling, we assign each sampled pixel a person label that gives the highest probability based on KDE.

Along each VRL, the number of pixels labeled as the same person is summed up to obtain the VRL histograms of the pixels for the same person, shown in the bottom of Fig.2 (c). Aside from the histograms, we also incorporate the positions of the 2D principal axes of the previous time frame.

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

The position of the principal axis p is further refined via Eq. 2, where p' is the initial position placed at the highest bin in the VRL histogram. $b(p')$, $b(r)$ and $b(l)$ are the counts at the p'^{th}, r^{th} and l^{th} bins. Function $left$ and $right$ provide the set of bins that belong to the same person and located at the left and right side of p'. See Fig.2 (c).

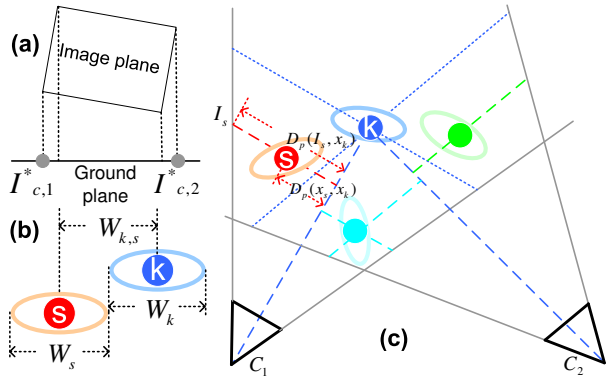

Fig. 3. Computation of view visibility. (a) Map camera image plane to ground plane. (b) Compute the occlusion threshold distance $W_{k,s} = \frac{W_k+W_s}{2}$. (c) Camera FOV on the ground plane (gray lines), and measure the inter-person occlusion degree by the geometric analysis.

4. TRACKING

As illustrated in our pipeline (Fig.1), to setup the criteria for ranking the cue reliability by the MAP framework (Eq.1), we compute the visibility of the views (Section 4.1) to each person and measure the smoothness of the motion (Section 4.2).

4.1. Computation of View Visibility

First, each camera's projection center and four image plane corners are mapped onto the ground plane. Among the 4 mapping corners, we select a point pair with maximum distance to each other. As one can see in Fig.3 (a), points $I^*_{c,1}, I^*_{c,2}$ define a line that approximates the projection of the camera image plane on the ground plane. Lines connecting these two points and the camera projection center approximate the camera's FOV on the ground plane (Fig.3 (c)).

A person inside the camera's FOV can be either partially or completely occluded by other persons. We propose to quantitatively measure every camera's visibility to a certain person. 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 constant, the quantitative measurement of the visibility of person k from the camera c_i in an N person group can be expressed as:

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

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.

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 the distance 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.3 (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.3 (b). Based on $W_{k,s}$, the occluder candidates are divided into 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 $D_P(x^{t-1}_s, x^t_{k,j}) < W_{k,s}$; $Q = q_1 + q_2$ denotes the total number of occluder candidates that are in the front of the target. See Fig.3 (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 $T\sum_{s=1}^{q_2} \frac{W_{k,s} - D_P(x^{t-1}_s, x^t_{k,j})}{W_{k,s}}$, and T numerically equals to N.

See Fig.3 (c), we define the occulders' area $O_{k,i}$ as the triangular area formed by camera FOV (the gray lines) and the target person (dotted blue lines). To have one candidate intersection point we need lines from at least two views. The visibility of the person on two views is set to be the joint visibility as $V_{k,j,1} * V_{k,j,2}$, which is proportional to the likelihood:

$$P(\{x^{t-1}_s\}^{N-1}_s | x^t_{k,j}) \propto V_{k,j1} * V_{k,j2} \quad (4)$$

where $j1$,$j2$ 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 $\exp(-|D_E(x^t_{k,j}, x^{t-1}_k)|)$, and proportional to the prior:

$$P(x^t_{k,j} | x^{t-1}_k) \propto \exp(-|D_E(x^t_{k,j}, x^{t-1}_k)|) \quad (5)$$

which gives low probability for non-smooth motions.

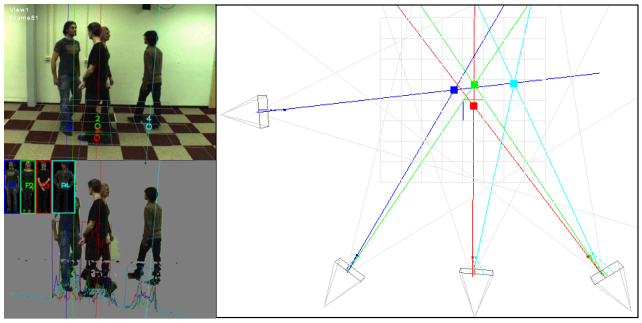

Fig. 4. Tracking 4PWALK sequences. Left: principal axes estimation in 2D view. Right: the estimate 3D positions, the colored lines indicate the selected view pairs.

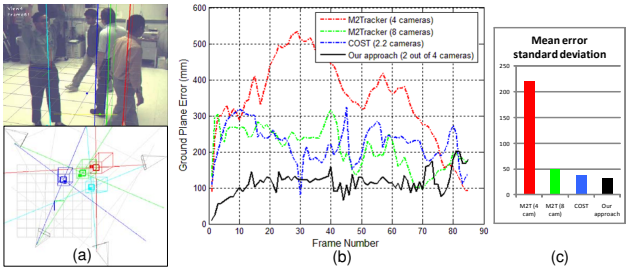

Fig. 5. Tracking LAB sequences. (a) Top: the estimated 2D principal axis (thinner line) and 3D position projection lines (thicker lines) in time step 61. Bottom: the colored solid squares are ground truth, the bounding boxes are estimated 3D positions. (b) Compare our approach with [7] and [3]: mean error plot. (c) Mean error standard deviation.

5. EXPERIMENTS AND EVALUATIONS

Our approach has been tested on a number of calibrated multi-view sequences, here are two challenging examples:

4PWALK: This sequence records four freely walking persons with severe inter-person occlusions. The tracking often fails if the positions are estimated by fusing the principal axes from all the views. By applying our framework approach with the trivial cost of visibility and motion smoothness calculation, for each time step, 2 best views (out of 4) are selected for each person, all 4 persons are properly tracked in the whole sequences. See Fig.4.

UMD LAB: The fifteen-view LAB dataset [7] records four moving persons with manually marked ground truth. Unlike [3] that uses eight cameras or more, we selected four cameras (index 00,03,06,12). Given the initial positions, our approach can properly track the 4 persons of severe inter-person occlusions, see Fig.5 (a). 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 are generally closer to the ground truth. Using only 2 best views, we achieved better tracking accuracies than [7] and [3]. See Fig.5 (b)(c) for the comparison of mean error in position estimation, and 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 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 the position candidate is the one generated from the views of higher visibility and smooth motion. In particular, the VRL set fundamentally improves the accuracy of principal axis-based person position estimation. Testing on the benchmark sets, our approach achieved better accuracies than the previous method. For more details see [8]. In the future, we will further extend the proposed method to a more flexible framework, to handle e.g. people of similar appearances, and people enter or go out of the scene.

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

7. REFERENCES

[1] Wei Du and Justus Piater, "A probabilistic approach to integrating multiple cues in visual tracking," *ECCV*, pp. 225–238, 2008.

[2] Kyungnam Kim and Larry S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," *ECCV*, pp. 98–109, 2006.

[3] Anurag Mittal and Larry S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *IJCV*, vol. 51, no. 3, pp. 189–203, 2003.

[4] W Hu, M Hu, X Zhou, T Tan, J Lou, and S Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *PAMI*, vol. 28, no. 4, pp. 663–671, 2006.

[5] Wei Du and Justus Piater, "Multi-camera people tracking by collaborative particle filters and principal axis-based integration," *ACCV*, pp. 365–374, 2007.

[6] C. Canton-Ferrer, J.R. Casas, M. Pard'as, and R. Sblendido, "Particle filtering and sparse sampling for multi-person 3d tracking," *ICIP*, pp. 2644–2647, 2008.

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

[8] Xinghan Luo, Robby T. Tan, and Remco C. Veltkamp, "Multi-person tracking based on vertical reference lines and dynamic visibility analysis," Tech. Rep. UU-CS-2011-010, ICS, Utrecht University, 2011.