SCISPACE
formerly Typeset

Open access • Proceedings Article • DOI:10.1109/ICPR.1996.546796

# Tracking human motion using multiple cameras — Source link

Qin Cai, Jake K. Aggarwal

**Published on:** 25 Aug 1996 - International Conference on Pattern Recognition

Related papers:

- Pfinder: real-time tracking of the human body
- A camera-based system for tracking people in real time
- 3-D model-based tracking of humans in action: a multi-view approach
- Human motion analysis: a review
- Tracking human motion in an indoor environment

Share this paper:

View more about this paper here: https://typeset.io/papers/tracking-human-motion-using-multiple-cameras-3akvtnqj5q

# Tracking Human Motion Using Multiple Cameras *

Q. Cai and J. K. Aggarwal
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
email: aggarwaljk@mail.utexas.edu

## Abstract

*This paper presents a framework for tracking human motion in an indoor environment from sequences of monocular grayscale images obtained from multiple fixed cameras. Multivariate Gaussian models are applied to find the most likely matches of human subjects between consecutive frames taken by cameras mounted in various locations. Experimental results from real data show the robustness of the algorithm and its potential for real time applications.*

## 1. Introduction

Tracking human motion in an indoor environment is of interest in applications of surveillance. In particular, we are developing a methodology to track individuals at sites such as corridors, airports, borders, and secured buildings. This requires that the viewing system be able to image the tracked subject in a broad area over a long period of time. In pursuit of this goal, our work has evolved from studying human walking using a fixed camera [1, 2] to tracking non-background objects in a single moving camera [3]. The studies in tracking using a fixed single camera [4, 2, 5] are limited to a very narrow area due to the restricted viewing angle of the system. A moving camera with a substantial degree of rotational freedom [3] increases the viewing angle to certain degree, however, it complicates the implementation by adding the motion estimation of both the viewing system and the subject of interest, and is still limited in the amount of viewing area. In this work, we chose to use multiple fixed cameras mounted in the area of interest to track and monitor the motion of individuals in sequences of monocular grayscale images. As long as the subject is within the area monitored by the fixed cameras, the image of this subject will be contained in the view of at least one camera. Based on this scenario, the problem of monitoring a subject becomes that of tracking the subject of interest in one camera view and matching that subject across different camera views, where the cameras' intrinsic parameters and relative positions are assumed to be known *a priori*.

To establish correspondence between consecutive frames from different cameras, conventional tracking methods based on the similarity of the object shape, such as cross-correlation and line-edge matching, are not applicable because the shape of an object image varies drastically from view to view of different cameras, and the whole body of a moving human usually goes through complicated changes during motion. Deformable template tracking for non-rigid objects does not fit either because the contours of the human shape are not always complete in cluttered indoor scenes. In addition, the continuity of the motion flow does not retain in the views of multiple cameras. Optical flow methods [6], which are widely used for featureless motion tracking, demand small and smooth motion between frames, a restriction that also does not hold in our case. In this paper, we propose to track a moving human in different camera views based on low level recognition of human motion [5]. A simpler form of a 2D human model [7] is applied to detect moving human subjects. Tracking between consecutive frames is mainly based on the consistency of the position, velocity, and average intensity of feature points formulated by multivariate Gaussian models, considered in the views of various cameras. The proposed algorithm is computationally efficient and can be readily used in real time applications.

## 2. Pre-processing

Three stages of pre-processing are performed before tracking begins: 1) segmentation of the non-background objects from the still background, 2) detection of human subjects from the segmented non-background objects, and 3) feature extraction from the segmented human subjects. The quality of object segmentation plays a critical role in later processing. If an non-background object is missed at

*The research reported in this paper was supported in part by the Army Research Office, Contract DAAH-94-G-0417 and Texas Advanced Technology Program, Contract ATP-442.

this stage, the system later will not be able to track this particular object. Detection of human subjects based on coarse human features reduces the ambiguity of matching in consecutive frames, decreases the search space, and makes the system robust and efficient. Finally, points belonging to the stick figure of a subject head and trunk are selected as the feature for matching in consecutive frames, due to its robustness in views of different cameras.

### 2.1. Segmentation

The proposed segmentation method takes advantage of the property of time-dependent data. Since we are using fixed cameras, the background image from the same camera view remains relatively unchanged. We recover and update the background image dynamically [3], as we are not able to obtain a "pure" background without including some non-background objects. Once the background image is recovered, images of the non-background objects can be separated from the background image by differencing and thresholding [8]. The next step is to obtain images of non-background objects within different bounding boxes. We apply the window slicing technique [9] to the thresholded binary image in a coarse to fine manner. The binary image is first smoothed by a $5 \times 5$ mean filter before we calculate its corresponding horizontal and vertical profiles. Then the valleys of the smoothed profiles are considered to be the boundaries of the rectangle boxes containing non-background objects.

### 2.2. Coarse Human Detection

Various techniques for modeling the human body have been developed by past researchers. Generally, the human body is represented either as a stick figure or as a volumetric model [10]. In this work, we use a coarse 2D model which is most related to Leung and Yang [7], but in a much simplified form. Based on the observation that the human head and trunk do not change as drastically as the hands and legs during motion, our method attempts to locate the head and trunk using a coarse 2D model of the human body. The human head is modeled as an ecllipse with a height to width ratio of 1 to 1.5. The human trunk is represented as a rectangle with a height to width ratio between 1 and 3 considering the different angles of body projections to the viewing camera. The ratio of the height of the trunk to that of the head is about 4 to 1. All the ratios are learnt from experimental study of images of humans from different points of view. To start with, we look for the location of the head considering the area of a blob which is consistent with $(\pi ab)/4$ where $a$ and $b$ are the axes of the ecllipse. If the area of any top subregion inside each bounding box is relatively consistent with the above relationship, we declare that this might be a head. Otherwise, we exclude it from further consideration. Any object that has a head region and a rectangle trunk region consistent with the coarse human model is considered a human subject. This step helps to remove most non-human objects.

### 2.3. Feature Extraction

We select $N$ points belonging to the medial axis of the upper body as the feature for tracking. Using multiple feature points rather than a single point [5] makes the matching of the same subject between consecutive frames more reliable. Compared to the height, the width of a subject is more likely to be complete from observations. Based on this assumption, we always treat the width of a bounding box as true information and adjust the height accordingly so that all the bounding boxes for comparison have the same width to height ratio. The locations of $N$ feature points in the medial axis of a upper body at time $t$ form a geometric feature vector $\mathbf{X}_t = [\mathbf{x}_{1t}, \mathbf{x}_{2t}, ..., \mathbf{x}_{mt}, ..., \mathbf{x}_{Nt}]^T = [(u_1, v_1), (u_2, v_2), ..., (u_m, v_m), ..., (u_N, v_N)]^T$. As for the visual features, we use an $N$ dimensional feature vector $\mathbf{Y}_t = \{y_{1t}, y_{2t}, ..., y_{Nt}\}^T$, in which $y_{mt}$ is the average intensity of the neighborhood of the $m$th feature points.

## 3. Tracking

To begin with, we monitor the target within the view of one fixed camera. Then the system follows the subject moving across the viewing boundary of one camera to another. As long as the target is within the field of view of the system cameras, it can always be tracked across various video streams captured from the cameras. Thus the tracking task in this setup consists of two major parts: 1) tracking a human in the view of one fixed camera, and 2) tracking a human across different camera views.

### 3.1. The Basic Tracking Scheme

Tracking a subject between adjacent frames can be achieved by finding the closest match in the next frame based on the consistency of certain features, such as geometric and visual features. Using Bayes' rule in the uniform *a priori* distribution case, searching a subject of interest reduces to finding the maximum of the $P(\mathbf{Z_t}|\Theta)$, where $\mathbf{Z}_t$ denotes a feature vector of a subject at time $t$, and $\Theta$ denotes the feature parameters corresponding to the tracked subject at time $t-1$. The whole feature space can be partitioned into two sub-spaces, namely, geometric and visual spaces, i.e.,

$$P(\mathbf{Z}_t|\Theta) = P_x(\mathbf{X}_t|\Theta_x)P_y(\mathbf{Y}_t|\Theta_y) \quad (1)$$

where $P_x(\mathbf{X}_t|\Theta_x)$ and $P_y(\mathbf{Y}_t|\Theta_y)$ are *pdf*s corresponding to the geometric and visual features respectively; and

$\mathbf{Z}_t = [\mathbf{X}_t \ \ \mathbf{Y}_t]$. In this work, we use position and velocity as the geometric features and average intensity of the neighborhood of a feature point as the visual feature. Multivariate Gaussian models are applied for the various feature distributions. Thus, maximizing $P(\mathbf{Z}_t|\Theta)$ is equivalent to minimizing its corresponding *Mahalanobis* distance, which is the sum of the *Mahalanobis* distances in sub-spaces, i.e.,

$$D_t = \sum_k w_k D_{k,xt} + \sum_l w_l D_{l,yt}$$

where $k$ and $l$ are indexes for the $k$th geometric and $l$th visual feature; $w_k$ and $w_l$ are the weights proportional to the robustness of the corresponding feature (we set them to be 1 in this work); $D_{k,xt}$ and $D_{l,yt}$ are the *Mahalanobis* distances of the $k$th geometric and $l$th visual features respectively. If multiple candidates exist for matching, we select the minimum value of $D_t$s as the best match. In cases only one candidate exists due to occlusion or disappearance of subjects, we compare the *Mahalanobis* distance to a certain threshold, and if the value is less than the threshold, it is considered to be a valid match. In the case of multiple candidates, the closest match found should also satisfy this threshold condition. In following subsections, we discuss the formulation of $P_x(\mathbf{X}_t|\Theta_x)$ and $P_y(\mathbf{Y}_t|\Theta_y)$ both in the cases of a single fixed camera and multiple fixed cameras.

### 3.2 Tracking in a Single Fixed Camera

To find the match for the subject of interest between consecutive frames. two cases are considered: 1) where there is no velocity information about the subject of interest, and 2) when velocity information is available.
Case 1) At the start of tracking when the velocity of the subject is unknown, the *pdf* of the geometric (position only) feature is approximated as

$$P_x(\mathbf{X}_t|\Theta_x) = \prod_{m=1}^{N} P_x(\mathbf{x}_{mt}|\Theta_x) \quad (2)$$
$$= \prod_{m=1}^{N} \frac{1}{2\pi\sigma_{x,m}^2} \exp[-\frac{(u_{mt}-\bar{u}_{mt})^2 + (v_{mt}-\bar{v}_{mt})^2}{2\sigma_{x,m}^2}]$$

where $(\bar{u}_{mt}, \bar{v}_{mt})$ is the $m$th feature point in the subject of interest in the previous frame at time $t$, and $\sigma_{x,m}$ is the maximum value of $\sqrt{(u_{mt}-\bar{u}_{mt})^2 + (v_{mt}-\bar{v}_{mt})^2}$ for all candidate subjects in the current frame. Similarly, the *pdf* of visual parameter space is modeled as

$$P_y(\mathbf{Y}_t|\Theta_y) = \prod_{m=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{y,m}} \exp[-\frac{(y_{mt}-\bar{y}_{mt})^2}{2\sigma_{y,m}^2}] \quad (3)$$

where $\bar{y}_{mt}$ is the $m$th visual feature point in the subject of interest in the previous frame, $t$ is time index, and $\sigma_{y,m}$ is the maximum value of $|y_m - \bar{y}_m|$ for all candidates in the current frame (set to 1 if only one candidate exists). According to Equations (2) and (3), the corresponding *Mahalanobis* distances are

$$D_{k,xt} = \sum_{m=1}^{N} \frac{(u_{k,mt}-\bar{u}_{k,mt})^2 + (v_{k,mt}-\bar{v}_{k,mt})^2}{\sigma_{xk,m}^2}$$
$$D_{l,yt} = \sum_{m=1}^{N} \frac{(y_{l,mt}-\bar{y}_{l,mt})^2}{\sigma_{yl,m}^2}.$$

Case 2) When the position and the velocity of the subject in previous frames are known, we use the same model but calculate the mean $(\bar{u}_{mt}, \bar{v}_{mt})$ for the multi-variate Gaussian model from the following equations

$$\bar{u}_{m(t-1)} - r_{t-1}\bar{u}_{m(t-2)} = \bar{u}_{mt}/r_t - \bar{u}_{m(t-1)}$$
$$\bar{v}_{m(t-1)} - r_{t-1}\bar{v}_{m(t-2)} = \bar{v}_{mt}/r_t - \bar{v}_{m(t-1)}$$

under the assumption that the width change is consistent with the depth change of the tracked subject among three consecutive frames. In the above equation, $(\bar{u}_{m(t-2)}, \bar{v}_{m(t-2)})$ and $(\bar{u}_{m(t-1)}, \bar{v}_{m(t-1)})$ are the locations of the same point in the previous two frames, $r_t$ denotes the width ratio of the tracked human upper body between time $t$ and $t-1$.

### 3.3. Matching in Multiple Fixed Cameras

To match a subject of interest in views of multiple fixed cameras, we assume that the intrinsic parameters of the cameras and relative positions between them are known *a priori*. The Gaussian model employed in tracking using visual cues is also similar to what we discussed in the previous subsection. The only difference is that we consider images from various cameras having different average lighting. Instead of using the intensity from an image directly, we normalize it by the ratio between the average intensities of various cameras. In this section, we focus on tracking based on geometric parameters, such as position and velocity of the feature point, in the views of different cameras.
• **Tracking Based on Positions.** The multi-variat Gaussian model for matching subject images between the frames taken by the previous camera and the current camera is modified from Equation 3 to

$$P_{x1}(\mathbf{X}_t|\Theta_{x1}) = \prod_{m=1}^{N} \frac{1}{2\pi\sigma_{x1,m}^2} \exp[-\frac{d_{mt}^2}{2\sigma_{x1,m}^2}],$$

where $d_{mt}$ is the distance between the $m$th feature point $(u_{mt}, v_{mt})$ in the view of the current camera and the line $a_{mt}x + b_{mt}y + c_{mt} = 0$ in the current camera which is mapped from the point $(\bar{u}_{mt}, \bar{v}_{mt})$ in the view of the previous camera, $t$ is the time index, and $\sigma_{x1,m}$ is obtained in the same

way as with a single camera, i.e., the maximum value of $d_{mt}$ for all candidates in the view of the current camera.

• **Tracking Based on Velocities.** Because all the extrinsic parameters are 3D in nature, the velocities used in tracking human subjects with multiple cameras have to be related to depth information. So the key problem is how to estimate the projection of a 3D point in the view of camera $i$ at time $t$ (denoted as $(\bar{u}_{it}, \bar{v}_{it})$), given $(\bar{u}_{i(t-1)}, \bar{v}_{i(t-1)})$, $(\bar{u}_{j(t-1)}, \bar{v}_{j(t-1)})$), and $(\bar{u}_{jt}, \bar{v}_{jt})$). Using the pinhole projection model, we have

$$\alpha_1 \, [\bar{u}_{i(t-1)} \; \bar{v}_{i(t-1)} \; f]^T = \alpha_2 R_{ij} \, [\bar{u}_{j(t-1)} \; \bar{v}_{j(t-1)} \; f]^T + T_{ij} \tag{4}$$

and

$$\beta_1 \alpha_1 \, [\bar{u}_{it} \; \bar{v}_{it} \; f]^T = \beta_2 \alpha_2 R_{ij} \, [\bar{u}_{jt} \; \bar{v}_{jt} \; f]^T + T_{ij} \tag{5}$$

where $\alpha_1$ and $\alpha_2$ are scaling factors, $\beta_1$ and $\beta_2$ are the depth ratio of the same point between time $t-1$ and time $t$, in the $i$th and $j$th cameras. These can be calculated using the width of the corresponding subject based on the justification that points belonging to the same subject have the same depths. From Equation 4 and 5, we arrive at

$$\begin{aligned}
\alpha &= \frac{(\beta_1 - 1) f}{r_{31} U + r_{32} V + r_{33}(\beta_2 - 1) f} \\
\bar{u}_{it} &= \frac{\alpha}{\beta_1}(r_{11} U + r_{12} V + r_{13}(\beta_2 - 1) f) + \frac{\bar{u}_{i(t-1)}}{\beta_1} \\
\bar{v}_{it} &= \frac{\alpha}{\beta_1}(r_{21} U + r_{22} V + r_{23}(\beta_2 - 1) f) + \frac{\bar{v}_{i(t-1)}}{\beta_1}
\end{aligned}$$

where $\alpha = \alpha_2/\alpha_1$, $U = \beta_1 \bar{u}_{jt} - \bar{u}_{j(t-1)}$, $V = \beta_1 \bar{v}_{jt} - \bar{v}_{j(t-1)}$, and $r_{kl}$ is the element of $R_{ij}$ in $k$th row and $l$th column. Finally, we substitute these $(\bar{u}_{it}, \bar{v}_{it})$s into a similar 2D Gaussian model as

$$\begin{aligned}
P_{x2}(\mathbf{X}_t | \Theta_{x2}) &= \prod_{m=1}^{N} P_{x2}(\mathbf{x}_{mt} | \Theta_{x2}) \\
&= \prod_{m=1}^{N} \frac{1}{2\pi\sigma_{x2,m}^2} \exp[-\frac{(u_{mt} - \bar{u}_{mt})^2 + (v_{mt} - \bar{v}_{mt})^2}{2\sigma_{x2,m}^2}].
\end{aligned}$$

## 4. Results and Extensions

In the system setup, we place cameras on both sides and at least one camera at the end of the scene, which most likely is a corridor. The distance between two cameras on the same side is at most twice the corridor width, restricted by the view angle of the wide angle lens (about 75 degrees) used in the experiments. In our prototype experiments, we use three cameras, with two of them mounted on each side and one at the end of the room, covering an area with a length-to-width ratio about 2, simulating a portion of a corridor. All these cameras are connected to a frame grabber which grabs, digitizes, and then sends grayscale images to a workstation for further processing. The images used in the experiments have a size of 512 × 480 pixels. The time interval between consecutive frames taken by a camera is 0.3 seconds, while the interval between consecutive frames taken by different cameras is 0.1 to 0.2 seconds. It takes about 0.3 seconds for a RISC workstation runing AIX (60MHZ) to process the tracking algorithm between consecutive frames.

(a) (b) (c)

**Figure 1. (a) The thresholded image after differencing, (b) the bounding boxes contain segmented non-background objects, (c) detected human head and trunk after the detection stage.**

Figures 1(a) shows an example of the thresholded binary image, Figure 1(b) is the result after segmentation, the detected human subjects are shown in Figure 1(c). Figure 1 presents a difficult case where one human subject is occluded by another, and the hands of the male subject sway away from the body. But we still see that the non-background objects due to the change of lighting is excluded after the detection stage. Figure 2 shows an example of tracking a light subject across two cameras. The switching between cameras are done manually. The subject is first tracked in a single fixed camera (camera 1). When the subject is about to walk out of the view of this camera, another camera (camera 2) takes over the tracking task. The white lines in the upper right images from camera 2 are the 2D lines projected from the same feature points in camera 1. The normalized *Mahalanobis Distance*s for both the correct match and other matches during camera switching are listed in Table 1. From the results listed in the table, we can see that each cue is successful in tracking the subject. However, single cue tracking is not as robust as the integration method when the subject is difficult to distinguish by either clothing intensity or the image position. For example, Table 2 shows a case where the integration method succeeds, while the matching by the single visual cue fails.

This paper has proposed a framework for tracking human motion in an indoor environment from sequences of monocular grayscale images using multiple cameras. Experimental results using real data prove the robustness of the

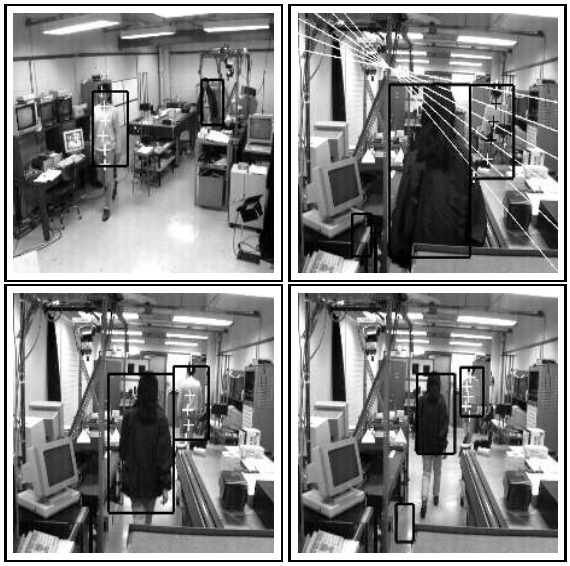

**Figure 2. An example of tracking a human subject using multiple fixed cameras.**

| | Correct Match | Other Match |
|---|---|---|
| Position | 0.074134 | 1.0 |
| Velocity | 0.254631 | 1.0 |
| Visual | 0.531236 | 1.0 |
| Sum | 0.860001 | 3.0 |

**Table 1. Normalized Mahalanobis Distances for all possible matches during camera switching.**

| | Correct Match | Other Match |
|---|---|---|
| Position | 0.373154 | 1.0 |
| Visual | 1.0 | 0.812189 |
| Sum | 1.373154 | 1.81218 |

**Table 2. Normalized Mahalanobis distances for all possible matches in the case where the single cue matching fails.**

algorithm. Future work can be extended in the following directions: 1) implementation of automatic camera switching among neighboring cameras, 2) performing a more accurate segmentation by using elaborate human models, 3) incorporation of more cues for tracking, and 4) testing on more video streams with longer sequences.

## References


[1] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed objects. In *Artificial Intelligence*, volume 19, pages 107–130, 1982.

[2] A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, Q. Cai, and J. K. Aggarwal. Lower limb kinematics of human walking with the medial axis transformation. In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 70–76, Austin, TX, 1994.

[3] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *2nd Intl. Conf. on Image Processing*, pages 215–218, Washinton, D.C., October 1995.

[4] A. Shio and J. Sklansky. Segmentation of people in motion. In *Proc. of IEEE Workshop on Visual Motion, IEEE Computer Society*, pages 325–332, October 1991.

[5] R. Polana and R. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, 1994.

[6] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–204, 1981.

[7] M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Trans. on PAMI*, 17(4):359–377, 1995.

[8] S. Yalamanchili, W. N. Martin, and J. K. Aggarwal. Extraction of moving object description via differencing. *CGIP*, (18):188–201, 1982.

[9] A. K. Jain. *Fundamental of Digital Image Processing*. Prentice-Hall International, Inc, 1989.

[10] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 16–22, Austin, TX, 1994.