Towards Detection of Human Motion

Yang Song†, Xiaolin Feng† and Pietro Perona†‡
† California Institute of Technology, 136-93, Pasadena, CA 91125, USA
‡ Università di Padova, Italy
{yangs,xlfeng,perona}@vision.caltech.edu

Abstract

Detecting humans in images is a useful application of computer vision. Loose and textured clothing, occlusion and scene clutter make it a difficult problem because bottom-up segmentation and grouping do not always work. We address the problem of detecting humans from their motion pattern in monocular image sequences; extraneous motions and occlusion may be present. We assume that we may not rely on segmentation, nor grouping and that the vision front-end is limited to observing the motion of key points and textured patches in between pairs of frames. We do not assume that we are able to track features for more than two frames. Our method is based on learning an approximate probabilistic model of the joint position and velocity of different body features. Detection is performed by hypothesis testing on the maximum a posteriori estimate of the pose and motion of the body. Our experiments on a dozen of walking sequences indicate that our algorithm is accurate and efficient.

1. Introduction

Perceiving the motion of the human body is difficult. First of all, the human body is richly articulated – even a simple stick model describing the pose of arms, legs, torso and head requires more than 20 degrees of freedom. The body moves in 3D which makes the estimation of these degrees of freedom a challenge in a monocular setting [3, 5]. Image processing is also a challenge: humans typically wear clothing which may be loose and textured, and part of the body is typically self-occluded. This makes it difficult to identify limb boundaries, and even more so to segment the main parts of the body. In a general setting all that can be extracted reliably from the images is patches of texture in motion. It is not so surprising after all that the human visual system has evolved to be so good at perceiving Johansson's stimuli [6, 7] where each joint of the body is shown as a moving dot.

Human motion perception may be divided into two phases: first detection and, possibly, segmentation; then tracking. Of the two, tracking has recently been object of much attention and considerable progress has been made [9, 8, 3, 4, 2]. Detection (given two frames: is there a human, where?), on the contrary, remains an open problem so that current trackers have either to be initialized by hand, or by ad-hoc heuristics. Song et al. [10] have focused on detection in the context of Johannson stimuli. A method was proposed based on probabilistic modeling of human motion and on modeling the dependency of the motion of body parts with a triangulated graph, which makes it possible to solve the combinatorial problem of labeling body parts in polynomial time. Excellent and efficient performance of the method has been demonstrated on a number of motion sequences. However, that work is limited to Johansson stimuli with no clutter (the only moving parts belong to the body, as in Johansson's displays) and very limited occlusion. In a realistic situation there is no guarantee that the joints of the body will constitute good features to be tracked by the early-vision front-end. Moreover: significant occlusion and possibly large amounts of moving clutter may be present. We propose a scheme which extends this work to real images. The localization results from our algorithm may be used to compute 3D pose as in [3, 5].

2. System Overview

Given two consecutive image frames, our goal is to detect whether a moving human body is present. As shown in Figure 1, our system requires a training phase. To this effect we first hand-construct a training set containing position and velocity of labeled features on the human body in a number of motion sequences. A model of human motion is learned from the training set. The model contains the joint position and velocity probability density function of triplets of features.

At runtime the system has a feature-tracking front-end measuring the position and velocity of all the observable features between two frames. From these features, we first

detect whether there is a person in the scene by maximizing the appropriate a posterior probability. Localization is further done by finding the labeling which maximizes the likelihood of the probabilistic model.

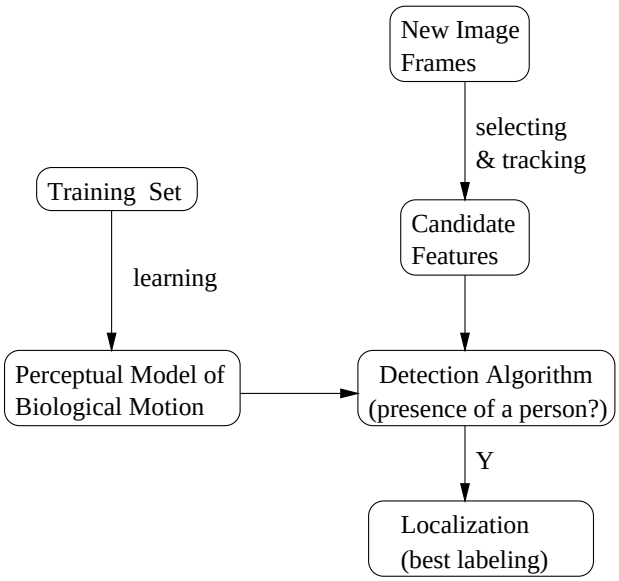


Figure 1. **System Overview**

3. Approach

The set of dots and associated velocities can be obtained from a motion detector/feature tracker applied to the entire image (Figure 2). In the following, we will address two problems: *detection* - if there is a person in the scene; *localization* - finding the most human-like configuration, i.e, the best labeling, given a set of features.

3.1. Notation

Suppose that we observe N points (as in Figure 2), and $\overline{X} = [X_1, \ldots, X_N]$ is the vector of measurements. Let O_1 denote a person present in the image, and O_0 absent. The detection task is to determine whether the ratio

$$\begin{aligned} \frac{P(O_1|\overline{X})}{P(O_0|\overline{X})} &= \frac{P(\overline{X}|O_1)P(O_1)/P(\overline{X})}{P(\overline{X}|O_0)P(O_0)/P(\overline{X})} \\ &= \frac{P(\overline{X}|O_1)}{P(\overline{X}|O_0)} \cdot \frac{P(O_1)}{P(O_0)} \end{aligned} \quad (1)$$

is greater than 1. If we assume the priors are equal, the second term of the above equation is 1. Let $\mathcal{S}_{body} = \{LW, LE, LS, H \ldots RT\}$ be the set of M body parts, for example, LW is the left wrist, RT is the right toe, etc, and BG be the background label. Let $\overline{L} = [L_1, \ldots, L_N]$ denote a possible labeling, where $L_i \in \mathcal{S}_{body} \cup \{BG\}$ is the label of X_i, $1 \leq i \leq N$. Assume $\mathcal{L}$ is all the possible labelings when a person present (O_1), then

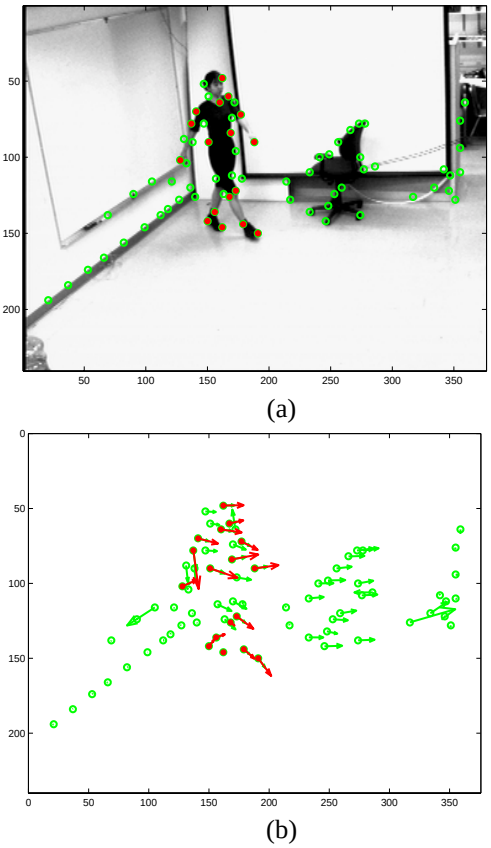


Figure 2. **Illustration of the approach** For a given image (a), features are first selected and tracked to the next frame. Dots in (a) are the features, and (b) shows the features with velocities. From all the candidate feature points (with positions and velocities), we want to first detect whether there is a person in the scene and then find the best labeling – the most human-like configuration (dark dots in (a) and (b)) according to a learned probabilistic model.

$$\begin{aligned} P(\overline{X}|O_1) &= \sum_{\overline{L} \in \mathcal{L}} P(\overline{X}, \overline{L}|O_1) \\ &= \sum_{\overline{L} \in \mathcal{L}} P(\overline{X}|\overline{L}, O_1) P(\overline{L}|O_1) \end{aligned} \quad (2)$$

When there is no person in the scene, the only possible labeling is $\overline{L}_0 = [BG, BG, \ldots, BG]$. Then,

$$\begin{aligned} P(\overline{X}|O_0) &= P(\overline{X}, \overline{L}_0|O_0) \\ &= P(\overline{X}|\overline{L}_0, O_0) P(\overline{L}_0|O_0) \\ &= P(\overline{X}|\overline{L}_0, O_0) \end{aligned} \quad (3)$$

If we don't have any prior information about the labeling, then we can assume in equation (2), for any labeling $\overline{L}$,

$P(\overline{L}|O_1) = 1/|\mathcal{L}|$, where $|\mathcal{L}|$ is the number of possible labelings. To compute equation (1), we still need to estimate $\sum_{\overline{L}\in\mathcal{L}} P(\overline{X}|\overline{L}, O_1)$ and $P(\overline{X}|\overline{L}_0, O_0)$.

Given a labeling $\overline{L}$, each point feature i has a corresponding label L_i. Therefore each measurement X_i corresponding to body labels may also be written as X_{L_i}, i.e. the measurements corresponding to specific body part associated with label L_i. For example if $L_i = LW$, i.e. the i^{th} label is associated to the left wrist, then $X_i = X_{LW}$ is the position and velocity of the left wrist.

Let's define

$$\begin{aligned}
\overline{\mathcal{L}}_{body} &= \{L_i; i \in 1, \ldots, N\} \cap \mathcal{S}_{body} \\
&\quad \text{set of body parts appearing in } \overline{L} \\
\overline{X}_{body} &= [X_{i_1}, \ldots, X_{i_K}] \\
&\quad \text{such that } \{L_{i_1}, \ldots, L_{i_K}\} = \overline{\mathcal{L}}_{body} \\
\overline{X}_{bg} &= [X_{j_1}, \ldots, X_{j_{N-K}}] \\
&\quad \text{such that } L_{j_1} = \cdots = L_{j_{N-K}} = BG
\end{aligned}$$

where K is the number of body parts appearing in $\overline{L}$. If we assume that the position and velocity of the visible body parts is independent of position and velocity of clutter points, then,

$$P(\overline{X}|\overline{L}, O_1) = P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body}) \cdot P_{bg}(\overline{X}_{bg}) \quad (4)$$

$$P(\overline{X}|\overline{L}_0, O_0) = P_{bg}(\overline{X}) \quad (5)$$

where $P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body})$ is the marginalized probability density of the whole body according to $\overline{\mathcal{L}}_{body}$. If independent uniform background noise is assumed, $P_{bg}(\overline{X}_{bg}) = (1/S)^{N-K}$, where $N - K$ is the number of background points, and S is the volume of the space X_i can be in. We will use this assumption about background features throughout this paper. Under this assumption, part of the background terms in $P(\overline{X}|\overline{L}, O_1)$ and $P(\overline{X}|\overline{L}_0, O_0)$ can be cancelled out so that detection can be performed by thresholding the summation of the 'modified' foreground likelihoods without accurately estimating background probabilities. More details of the procedure will be explained below.

3.2. Summation of likelihoods

We first consider the problem where there are no missing body parts, i.e., if a person is present, then all the body parts can be seen. In this case, from the above subsection, we know that if background (clutter) features are assumed to be independent and uniform, then the detection depends on $(1/|\mathcal{L}|)\cdot\sum_{\overline{L}\in\mathcal{L}} P_{\mathcal{S}_{body}}(\overline{X}_{body})$. If the summation is done in a brute-force way, the computational cost would be exponential with regard to the number of body parts (M), which is computationally prohibitive. The method proposed in [10] provides a way to approximate the foreground probability density $P_{\mathcal{S}_{body}}(\overline{X}_{body})$ so that we can do the summation efficiently. By using the kinematic chain structure of human body, the whole body can be decomposed as in Figure 3. If the appropriate conditional independence (Markov property) is valid, then

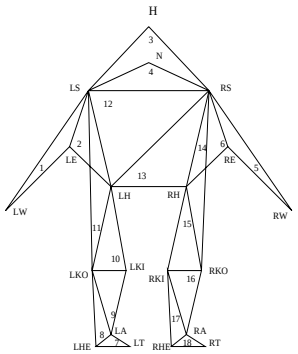


Figure 3. **Decompositions of the human body into triangles.** 'L' and 'R' in label names indicate left and right. H:head, N:neck, S:shoulder, E:elbow, W:wrist, H:hip, KI:inside knee, KO:outside knee, A:ankle, HE:heel and T:toe. The numbers inside triangles give the order in which the algorithm proceeds.

$$\begin{aligned}
& P_{\mathcal{S}_{body}}(\overline{X}_{body}) \\
= \; & P_{LW,LE,LS}(X_{LW}|X_{LE}, X_{LS}) P_{LE,LS,LH}(X_{LE}|\ldots) \\
& \ldots P_{RA,RHE,RT}(X_{RA}, X_{RHE}, X_{RT}) \\
= \; & \prod_{t=1}^{T-1} P_t(X_{A_t}|X_{B_t}, X_{C_t}) \cdot P_T(X_{A_T}, X_{B_T}, X_{C_T})
\end{aligned} \quad (6)$$

Where T is the number of triangles in the decomposed graph in Figure 3, t is the triangle index, and A_t is the first label associated to triangle t, etc. The structure of the decomposable graph ([1, 10]) allows us to do the summation as follows,

$$\begin{aligned}
& \sum_{\overline{L}\in\mathcal{L}} P_{\mathcal{S}_{body}}(\overline{X}_{body}) \\
= \; & \sum_{\overline{L}\in\mathcal{L}} \prod_{t=1}^{T-1} P_t(X_{A_t}|X_{B_t}, X_{C_t}) P_T(X_{A_T}, X_{B_T}, X_{C_T}) \\
= \; & \sum_{X_{A_T}, X_{B_T}, X_{C_T}} P_T(X_{A_T}, X_{B_T}, X_{C_T}) \sum_{X_{A_{T-1}}} \cdots \\
& \sum_{X_{A_2}} P_2(X_{A_2}|X_{B_2}, X_{C_2}) \sum_{X_{A_1}} P_1(X_{A_1}|X_{B_1}, X_{C_1})
\end{aligned} \quad (7)$$

The summation in equation (7) can be done by an algorithm similar to dynamic programming ([10, 1]). Let

$$\Psi_t(X_{A_t}, X_{B_t}, X_{C_t}) = P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t}), \quad for\ 1 \le t \le T-1 \quad (8)$$

$$\Psi_t(X_{A_t}, X_{B_t}, X_{C_t}) = P_{A_T B_T C_T}(X_{A_T}, X_{B_T}, X_{C_T}), \quad for\ t = T \tag{9}$$

be the cost function associate with each triangle, then the summation algorithm can be described as follows:

Stage 1: for every pair (X_{B_1}, X_{C_1}),

Compute $\Psi_1(X_{A_1}, X_{B_1}, X_{C_1})$ for all possible X_{A_1}

Define $T_1(X_{A_1}, X_{B_1}, X_{C_1})$ the total value so far.

Let $T_1(X_{A_1}, X_{B_1}, X_{C_1}) = \Psi_1(X_{A_1}, X_{B_1}, X_{C_1})$

Store $\Gamma_1(X_{B_1}, X_{C_1}) = \sum_{X_{A_1}} T_1(X_{A_1}, X_{B_1}, X_{C_1})$

Stage t, $2 \le t \le T$: for every pair (X_{B_t}, X_{C_t}),

Compute $\Psi_t(X_{A_t}, X_{B_t}, X_{C_t})$ for all possible X_{A_t}

Compute the total value so far (till stage t):

– Define $T_t(X_{A_t}, X_{B_t}, X_{C_t})$ the total value so far. Initialize $T_t(X_{A_t}, X_{B_t}, X_{C_t}) = \Psi_t(X_{A_t}, X_{B_t}, X_{C_t})$

– If edge (A_t, B_t) is contained in a previous stage and τ is the latest such stage, multiply $\Gamma_\tau(X_{A_t}, X_{B_t})$ (or $\Gamma_\tau(X_{B_t}, X_{A_t})$ if the edge was reversed) to $T_t(X_{A_t}, X_{B_t}, X_{C_t})$

– Likewise, multiply the values of the latest previous stages containing respectively edge (A_t, C_t) and edge (B_t, C_t) to $T_t(X_{A_t}, X_{B_t}, X_{C_t})$

Store $\Gamma_t(X_{B_t}, X_{C_t}) = \sum_{X_{A_t}} T_t(X_{A_t}, X_{B_t}, X_{C_t})$

When stage T calculation is complete, the overall sum can be obtained by

$$\sum_{\overline{L} \in \mathcal{L}} P_{\mathcal{S}_{body}}(\overline{X}_{body}) = \sum_{X_{B_T}, X_{C_T}} \Gamma_T(X_{B_T}, X_{C_T}) \tag{10}$$

The computational complexity of the above method is on the order of $M * N^3$.

3.3. Detection and localization - with occlusion

From the above subsection, in the case of no occlusion, detection can be done by thresholding $(1/|\mathcal{L}|) \cdot \sum_{\overline{L} \in \mathcal{L}} P_{\mathcal{S}_{body}}(\overline{X}_{body})$. Assuming equal priors and independent and uniform background features, localization and labeling can be obtained by finding the labeling $\overline{L}^*$,

$$\begin{aligned} \overline{L}^* &= \arg\max_{\overline{L} \in \mathcal{L}} P(\overline{L}|\overline{X}, O_1) \\ &= \arg\max_{\overline{L} \in \mathcal{L}} P(\overline{X}|\overline{L}, O_1) P(\overline{L}|O_1)/P(\overline{X}) \\ &= \arg\max_{\overline{L} \in \mathcal{L}} P(\overline{X}|\overline{L}, O_1) \\ &= \arg\max_{\overline{L} \in \mathcal{L}} P_{\mathcal{S}_{body}}(\overline{X}_{body}) \end{aligned} \tag{11}$$

The above optimization can be done by dynamic programming as in [10, 11].

When some body parts are occluded, the foreground probability $P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body})$ is the marginalized version of $P_{\mathcal{S}_{body}}(\overline{X}_{body})$ – marginalization over the missing body parts. If we assume that the background features are independent and uniformly distributed, detection can be done by thresholding

$$(1/|\mathcal{L}|) \cdot \sum_{\overline{L} \in \mathcal{L}} P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M - K_{\overline{L}}} \tag{12}$$

where M is the total number of body parts, $K_{\overline{L}}$ is the number of body parts present in labeling $\overline{L}$, and $1/S$ is the volume of the space X_i lies in. If the local cost function $\Psi_t(X_{A_t}, X_{B_t}, X_{C_t})$ associate with triangle t, ($1 \le t \le T-1$), is defined as

- if all the three body parts observed, it is $P_{A_t|B_t,C_t}(X_{A_t}|X_{B_t}, X_{C_t})$;
- if A_t is missing or two or three of A_t, B_t, C_t are missing, it is $1/S$;
- if B_t or C_t is missing and the other two body parts observed, it is $P_{A_t|C_t}(X_{A_t}|X_{C_t})$ or $P_{A_t|B_t}(X_{A_t}|X_{B_t})$.

(the same idea can be applied to the last triangle T), then the summation algorithm described in section 3.2 can be used to obtain equation (12).

Similar to equation (11), the localization and labeling can be found by

$$\begin{aligned} \overline{L}^* &= \arg\max_{\overline{L} \in \mathcal{L}} P(\overline{L}|\overline{X}, O_1) \\ &= \arg\max_{\overline{L} \in \mathcal{L}} P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M - K_{\overline{L}}} \end{aligned} \tag{13}$$

Under the above described local cost function, dynamic programming can be used to get the optimum labeling.

The detailed analysis and explanation of equation (12) to (13) can be found in [11]. One intuitive explanation is that for each triangle, the dimensions of the local cost function are the same for different number of missing body parts, which makes it reasonable to sum (or get the maximum of) them locally. Also, the dimension of the domain of $P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M - K_{\overline{L}}}$ is fixed regardless of the number of candidate features and the number of missing body parts in the labeling $\overline{L}$, so we can directly compare the likelihood of different hypotheses, even hypotheses from different images.

Another way to perform detection [11] is to first get the most likely labeling (the labeling with highest $P_{\overline{\mathcal{L}}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M - K_{\overline{L}}}$), then compare the likelihood of such labeling to a threshold. If the likelihood is higher than the threshold, then we will declare that a person is there. We did experiments on both methods and compare their performances in the experiments section.

3.4 Using information from multiple frames

So far, we have only assumed that we may use information from two consecutive frames, from which we obtain position and velocity of a number of features. In this section we would like to extend our previous results to the case where multiple frames are available. However, in order to maintain generality we will assume that tracking features across more than 2 frames is impossible. This is a simplified model of the situation where, due to extreme body motion or to loose and textured clothing, tracking is extremely unreliable and each individual feature's lifetime is short. Neri et al. [7] used similar assumption when conducting their psychophysical investigation of biological motion perception in the human visual system.

Let $P(O|\overline{X})$ denote the probability of the existence of a person given $\overline{X}$ observed. From previous subsections, we use the approximation: $P(O|\overline{X})$ is proportional to $\Phi(\overline{X})$ defined as $\Phi(\overline{X}) \stackrel{\text{def}}{=} (1/|\mathcal{L}|) \cdot \sum_{\overline{L}\in\mathcal{L}} P_{\mathcal{L}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M-K_{\overline{L}}}$. Now if we have n observations $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$, then the decision depends on:

$$\begin{aligned} & P(O|\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n) \\ = & \frac{P(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n|O) \cdot P(O)}{P(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)} \\ = & \frac{P(\overline{X}_1|O)P(\overline{X}_2|O)\ldots P(\overline{X}_n|O) \cdot P(O)}{P(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)} \end{aligned} \quad (14)$$

The last line of the above equation holds if we assume that $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$ are independent observations. Assuming the priors are equal, $P(O|\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)$ can be represented by $P(\overline{X}_1|O)P(\overline{X}_2|O)\ldots P(\overline{X}_n|O)$, which is proportional to $\prod_{i=1}^{n}\Phi(\overline{X}_i)$. If we set up a threshold for $\prod_{i=1}^{n}\Phi(\overline{X}_i|\overline{L}_i^*)$, then we can do detection given $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$.

4. Experiments

The image sequences used in the experiments were captured by a CCD camera at 30 Hz. There are three different types of motion: (1). A subject walks from the left back corner to the right front corner, facing about 60 degrees away from the front view (middle row of Figure 4). For this motion, we have about 1000 frames (8 sequences, around 120 frames each) as training set, and another 1500 frames (12 sequences) as testing set. (2). A chair moves from left to right, about 1000 frames (8 sequences) (bottom row of Figure 4). (3). While a subject walks as in the motion type (1), a chair also moves as a background moving object (top row of Figure 4). 2000 frames (16 sequences) were collected. The goal is to detect if there is a person walking in the scene and further localize and label the person.

4.1. Training of the probabilistic models

We chose 20 features to represent the human body configuration. Most of these features are close to the main joints of the body. The dark dots in Figure 2 show 17 of them (being correctly labeled in that frame), the other 3 are missing: two at the left knee and one at the right heel.

On the 8 training sequences with about 1000 frames in total, we hand-construct the ground truth of feature positions and velocities in the following way: on the first frame of each sequence, we manually select the positions of all the visible model features. Then the features are tracked automatically to the next frame using the Lucas-Tomasi-Kanade tracking algorithm ([12]) and their velocities between the two frames are computed. At each frame after tracking, we monitor the result and discard the features which have obvious tracking errors. The correct positions of these discarded features and some newly appeared ones after occlusion are given by hand, so that we have again the positions of all the features appearing in this frame and we may track them to the next frame and get their velocities. The features are also hand-labeled at the same time. Occlusion is common in our training set: each feature is present in approximately 85% of frames (see Figure 5 (a)).

The training was done by estimating the joint (or conditional) probabilistic density functions (pdf) for all the triplets as described in section 3. As in [10, 11], we assumed all the pdfs were Gaussian, and the parameters for the Gaussian distribution were estimated from the training set.

4.2. Testing Set

For the testing sequences, the system automatically selects features at each frame, and tracks them to the next frame. The feature selection and tracking algorithm is the standard Tomasi-Kanade version. We don't track features over more than 2 frames, but reselect all the features at the next frame after tracking. Thus, there is no feature correspondence between sequential frames and each frame has a unique set of features and velocities, which is arguably the most difficult conditions under which to perform labeling and detection, as mentioned in section 3.4. The dots in Figures 2 and 4 are the features from the automatic selection and tracking.

4.3. Test of probabilistic model

To test the triangulated probabilistic model (Figure 3), we first did experiments on the manually tracked data (with ground truth, as in section 4.1). We have a total of 8 sequences (with 120 frames each). To test a sequence, frames

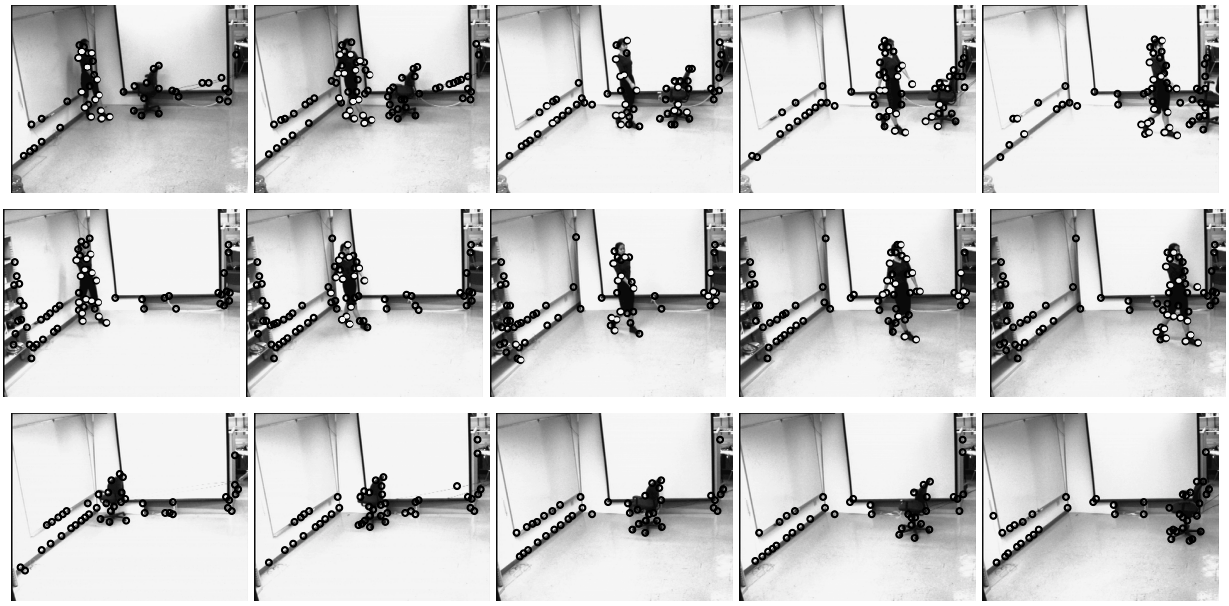
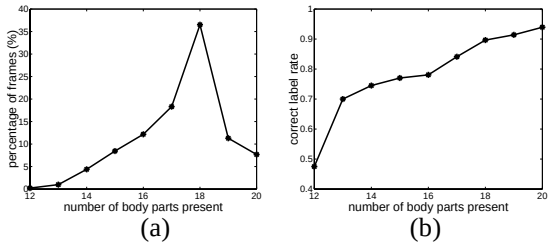


Figure 4. **Sample frames** from body and chair moving sequences (top row), body moving sequences (middle row), and chair moving sequences (bottom row). The dots (either in black or in white) are the features selected by Tomasi-Kanade algorithm on two frames. The white dots are the most human-like configuration found by our algorithm.

from all the other seven sequences were used as the training set. A label error happens when a body part appears but is labeled as either a different part, or as background, and when a body part is missing but its label is assigned to another point. Figure 5 (a) shows the statistics of the number of body parts present in all the sequences used in this experiment. Since the data were manually tracked, not a big number of body parts were missing. Figure 5 (b) shows the correct labeling rate vs. the number of body parts present, with the overall (considering all the frames) correct labeling rate being 85.89%. If the average number of features detected is N, ($N \approx 17$ in this experiment), the chance level of a body part being assigned a correct candidate feature by random selection is $1/(N+1)$ (with one more background point). The correct rate here is much higher than that. From Figure 5 (b), we see that the correct label rate goes up as the number of detected body parts increases, which is consistent with the fact that with more body parts present, the probability decomposition as in equation (6) is a more accurate approximation.



(a) (b)

Figure 5. **(a)** percentage of frames corresponding to the number of body parts present in the hand-constructed (as in section 4.1) data set; **(b)** correct labeling rate vs. the number of body parts present. The chance level of a body part being assigned a correct candidate feature is around 0.06. The correct rates here are much higher than that.

4.4. Detection

The detection task is: for a given image pair, to decide whether or not there is a moving person in the scene. We performed detection experiments using three types of sequences: body moving (middle row of Figure 4), body and chair moving (top row of Figure 4), and chair moving (bottom row of Figure 4). Figure 4 shows sample frames from the three types of sequences. The white and black dots are features detected by the Tomasi-Kanade tracker. The sequences with only the person walking had a total of 1500 frames and with an average of 64 features detected per frame. The sequences with both the person and the chair moving had 2000 frames total and average 58 features per frame, and the sequences with only the chair had 1000 frames and 46 features.

Figure 6 shows two receiver operating characteristics (ROC) curves constructed from the summation of likelihoods as in equation (12). The solid curve is the ROC when the sequences with the body and chair and the sequences with the chair only were combined to compute the false alarm and detection rates. With $P_{detect} = 1 - P_{false-alarm}$, the detection rate was 87.54%. The dashed curve is the

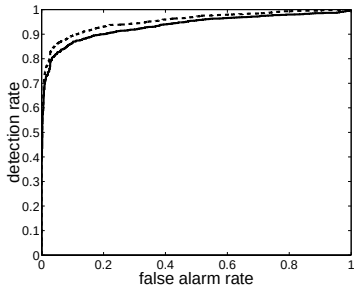


Figure 6. **ROC curves** of the detection experiment in section 4.4. Solid line: images with body and chair vs. images with chair only, the detection rate is 87.54% when $P_{detect} = 1 - P_{false-alarm}$; Dashed line: images with body only vs. images with chair only, the detection rate is 89.61% when $P_{detect} = 1 - P_{false-alarm}$.

ROC when the results of the sequences with the body only and with the chair only are combined. In this case the detection rate at $P_{detect} = 1 - P_{false-alarm}$ is 89.61%. The two curves are very similar, showing that adding a distractor (moving chair) to the scene does not degrade the performance of the person detector much. In fact, the difference between the two ROCs is more likely attributable to the facts that the backgrounds were slightly different in the sequences (and many more features on the shelf (left front in the images of middle row of Figure 4) were tracked in the sequences with only the person moving), and the variability in the motion and path of the person.

Figure 4 also gives the localization results. For each image, the white dots correspond to the best labeling $\overline{L}^*$ as in equation (13). For most frames, the person is localized correctly. Notice that for an image, the white dots consisting of the best configuration can be far away from each other. For example for the frame in the middle of the top row (Figure 4), except the white dots on the body, two white dots are on the wall, and four white dots are on the chair. A detailed study finds that the program took the two dots on the wall as 'left elbow and left wrist', and the four dots on the chair as 'left outside knee, left ankle, left toe and left heel'. The reason for this is that for a triangulated body decomposition such as the one we use, shown in Figure 3, if, say, 'left shoulder and left hip' are missing, then both 'left elbow and left wrist' and 'left outside knee, left ankle, left toe and left heel' are disconnected with other body parts. Therefore, the optimal labeling is composed of several independent components, possibly far away from each other. It is clear that in this case the conditional independence required by equation (6) is not a good approximation any longer.

Experiments were also conducted to compare the performance of thresholding the summation of likelihood of all the possible labelings (as in section 3.3) and thresholding the likelihood of the most human like configuration (as in [11]). Solid curves in Figure 7 show the results of using the method in section 3.3, and dashed lines are of [11]. Figure

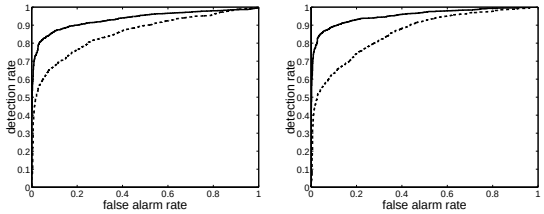


Figure 7. **ROC curves**. (a) Results of images with body and chair vs. images with chair only. (b) Results of images with body only vs. images with chair only. Solid line: using method in section 3.3; dashed line: using method in [11].

7 (a) and (b) are respectively of images with body and chair vs. images with chair only and of images with body only vs. images with chair only. From Figure 7, we see that our method here works better.

4.5. Using information from multiple frames

Here we tested how the detection rate improved by integrating information over time, using the approach described in section 3.4. The sequences with the body and chair and the sequences with the chair only were used. Figure 8(a) shows ROC curves of using 1 to 4 frames respectively. Figure 8(b) plots the detection rates (with $P_{detect} = 1 - P_{false-alarm}$) vs. the number of frames integrated. With more frames used, the detection rate gets higher. The detection rate is more than 98% when more than 7 frames (around 200 ms) were used.

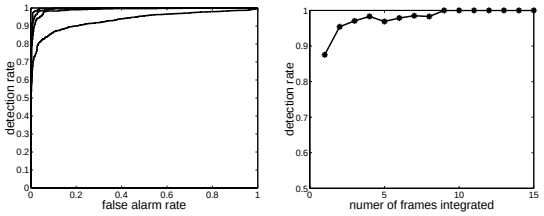
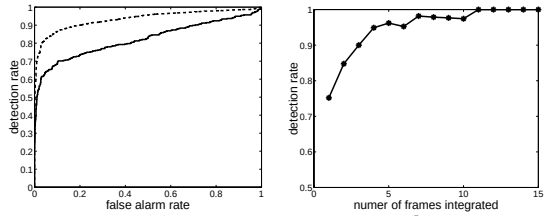


Figure 8. **Results of integrating multiple frames. (a)** Four curves are ROCs of integrating 1 to 4 frames respectively. The more frames integrated, the better the ROC curve is. **(b)** detection rate (when $P_{detect} = 1 - P_{false-alarm}$) vs. number of frames used.

4.6. Experiments on different subjects

In the previous experiments, the sequences for training and testing were from the same subject. In this section we test the performance on another subject, who was also walking with a chair moving in the scene. Four sequences, around 120 frames each, were used. Figure 9(a) shows the comparison result. The solid line is the ROC curve for the new subject, with 75.19% detection rate (when $P_{detect} = 1 - P_{false-alarm}$), and the dashed line is that of the subject of the training set. Figure 9(b) shows the detection rates (with $P_{detect} = 1 - P_{false-alarm}$) vs. the



(a) (b)

Figure 9. **Results on a different subject. (a)** ROC curves. Solid line: another subject (different from training set); Dashed line: Subject of the training set (the same as the solid line of Figure 6). **(b)** detection rate (when $P_{detect} = 1 - P_{false-alarm}$) vs. number of frames used for the new subject.

number of frames integrated for the new subject. The detection performance improves with more frames integrated: it is almost perfect with more than 10 frames used.

5. Discussion and conclusions

We have presented a method for detecting and labeling human motion in monocular image sequences. The method takes as its input the position and velocity of the most salient features in the image, as computed by the Lucas-Kanade feature tracker. No prior image segmentation is required. The method is based on modeling human motion with an approximation of the joint probability density of the position and motion of features that are associated with the human body. Given a (possibly cluttered) motion sequence, the detection is performed by summation of the likelihoods of all the possible labelings. Localization is done by finding the subset of detected features that is most likely to be associated with a human body. The model is trained on a hand-labeled training set.

We have tested our method on a number of image sequences containing either a walking pedestrian, or some non-human motion, or both. The results are encouraging: the detection rate is around 90% on 2 frames, or 60ms, and in excess of 98% on 7 frames, or 200ms. It also appears to generalize well when training and testing are done on two different people. Both labeling and detection take less than 1 second per frame in a Matlab implementation running on a 450MHz Pentium PC giving hope for a real-time implementation in C.

Our ideas may be extended and improved upon in a number of directions. For instance, currently human motion is modeled using Gaussians; this choice is arbitrary and needs to be re-examined in the light of our training data. Also, we did not experiment with different structures for the triangulated model – many reasonable choices exist. Furthermore, some form of hierarchical modeling will be needed to account for long-range dependency of body parts; this is critical in the case of occlusion as discussed in the experimental section. One last issue: Song et al. [10] have demonstrated that their system generalizes well to viewpoint changes and to different types of motion when using unoccluded Johansson stimuli and this gives reason to believe that our system would be equally robust. However, systematic testing needs to be done on a variety of body motions and under a number of viewing conditions in order to assess the limits.

Acknowledgments

Funded by the NSF Engineering Research Center for Neuromorphic Systems Engineering (CNSE) at Caltech (NSF9402726), and by an NSF National Young Investigator Award to PP (NSF9457618). We are grateful to Luis Goncalves and Yair Weiss for helpful comments.

References

[1] Y. Amit and A. Kong. Graphical templates for model registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:225–236, 1996.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE CVPR*, pages 8–15, 1998.

[3] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. In *Proc. 5^{th} Int. Conf. Computer Vision*, pages 764–770, Cambridge, Mass, June 1995.

[4] I. Haritaoglu, D.Harwood, and L.Davis. Who, when, where, what: A real time system for detecting and tracking people. In *Proceedings of the Third Face and Gesture Recognition Conference*, pages 222–227, 1998.

[5] N. Howe, M. Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Tech. Rep. TR-99-37, a Mitsubishi Electric Research Lab*, 1999.

[6] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[7] P. Neri, M.C.Morrone, and D.C.Burr. Seeing biological motion. *Nature*, 395:894–896, 1998.

[8] J. Rehg and T. Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Proceedings of the workshop on Motion of Non-Rigid and Articulated Bodies*, pages 16–24, November 1994.

[9] K. Rohr. Incremental recognition of pedestrians from image sequences. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 8–13, New York City, June, 1993.

[10] Y. Song, L. Goncalves, E. D. Bernardo, and P. Perona. Monocular perception of biological motion - detection and labeling. In *International Conference on Computer Vision*, pages 805–812, Sept 1999.

[11] Y. Song, L. Goncalves, and P. Perona. Monocular perception of biological motion - clutter and partial occlusion. *To appear in Sixth European Conference on Computer Vision*, 2000.

[12] C. Tomasi and T. Kanade. Detection and tracking of point features. *Tech. Rep. CMU-CS-91-132,Carnegie Mellon University*, 1991.