

Monocular perception of biological motion in Johansson Displays *

Yang Song[†], Luis Goncalves[†], Enrico Di Bernardo[†] and Pietro Perona^{†‡}

[†] California Institute of Technology, 136-93, Pasadena, CA 91125, USA

[‡] Università di Padova, Italy

{yangs,luis,dibe,perona}@vision.caltech.edu

Running head: Monocular perception of biological motion

Contact author: Yang Song

Address: California Institute of Technology - Mail stop 136-93
Pasadena, CA 91125, USA

Tel: (626) 395-4866

Fax: (626) 795-8649

Email: yangs@vision.caltech.edu

*The work in this paper was partially published in the Proceedings of ICCV'99 and in the Proceedings of ECCV'00

Abstract

Computer perception of biological motion is key to developing convenient and powerful human-computer interfaces. Algorithms have been developed for tracking the body; however, initialization is done by hand. We propose a method for detecting a moving human body and for labeling its parts automatically in scenes that include extraneous motions and occlusion. We assume a Johansson display, i.e. that a number of moving features, some representing the unoccluded body joints and some belonging to the background are supplied in the scene. Our method is based on maximizing the joint probability density function (PDF) of the position and velocity of the body parts. The PDF is estimated from training data. Dynamic programming is used for calculating efficiently the best global labeling on an approximation of the PDF. Detection is performed by hypothesis testing on the best labeling found. The computational cost is on the order of N^4 where N is the number of features detected.

We explore the performance of our method with experiments carried on a variety of periodic and non-periodic body motions viewed monocularly for a total of approximately 30,000 frames. The algorithm is demonstrated to be accurate and efficient.

1 Introduction

Perceiving the position and the activities of humans (‘biological motion’ in the literature of human vision) is very useful for human social interactions, and is a most important technology for developing convenient and effective human-computer interfaces. A striking demonstration of the capabilities of the human visual system is provided by the experiments of Johansson [13]. Johansson filmed people acting in total darkness with small light bulbs fixed to the main joints of their body. A single frame of a Johansson movie is nothing but a cloud of identical bright dots on a dark field; however, as soon as the movie is animated one can readily detect, count, segment a number of people in a scene, and even assess their activity, age and sex [4, 15, 5]. Although such perception is completely effortless, our visual system is ostensibly solving a hard combinatorial problem (which dot should be assigned to which body part of which person?).

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 [8, 11]. Image processing is also a challenge: humans typically wear clothing which may be loose and textured. This makes it difficult to identify limb boundaries, and even more so to segment the main parts of the body. Additionally, in the same scene there may be other moving objects: cars, vegetation waving in the wind and the human body may be partially occluded. 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.

Perception of biological motion 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 [19, 18, 2, 8, 9, 3, 23, 7]. Detection

(given two frames: is there a human, where?), on the contrary, remains an open problem. In this paper we address the problem of defining and estimating a perceptual model of biological motion and use it for detecting the human body and labeling it in monocular image sequences. By ‘labeling’ we mean assigning to each region in the image a label that corresponds to the body part (shoulder, elbow etc) that is imaged in that region. We choose not to address the issue of detecting and classifying pictorial features that are associated to the body parts – for the time being this has been sufficiently explored by [12, 20, 25, 26]. Therefore our experimental setup is identical to Johansson’s experiments [17, 10, 6, 14]: we suppose that a number of markers are attached to the body of an actor. At every frame we need to attach labels to the observed features (some may be caused by noise; some body parts may have been missed). Rather than modeling the details of the mechanics of the human body, we choose to approach biological motion perception as the problem of recognizing a peculiar spatio-temporal pattern which may be learned perceptually. Therefore we approach the problem using learning and statistical inference. We observe the subject moving about in order to estimate a model of his/her stereotypical motions. This model, which we formulate as the joint probability density function (PDF) of the position and motion of the body, is used to select the best labeling. In a cluttered scene, if the hypothesis corresponding to the best labeling exceeds a likelihood threshold, we will say a human is detected.

This paper is organized as follows. In section 2, we deal with the labeling problem when there is no occlusion, no clutter points. In section 3, we explain how to extend the algorithm to perform detection and labeling in a cluttered and occluded scene. Section 4 contains a number of experiments characterizing the performance and the failure modes of our algorithm.

2 The Johansson problem

2.1 Notation and approach

We first consider the problem of labeling a set of observed points when there is no clutter and no body parts are missing. We call this the 'Johansson problem'. We choose to characterize body pose and motion by the joint probability density of the position and velocity of its parts. Our goal is to interpret monocular image sequences, hence we use part position and velocity in the image plane (Figure 1). In our Johansson scenario each part appears as a single dot (marker) in the image plane. Therefore its identity is not revealed by cues other than its relative position and velocity.

Let $\mathcal{S}_{body} = \{LW, LE, LS, H \dots RF\}$ be the set of M body parts, for example, LW is the left wrist, RF is the right foot, etc. Correspondingly, let X_{LW} be the vector representing the position and velocity of the left wrist, X_{RF} be the vector of the right foot, etc. We model the pose and motion of the body probabilistically by means of a probability density function $P_{\mathcal{S}_{body}}(X_{LW}, X_{LE}, X_{LS}, X_H, \dots, X_{RF})$.

Let $\bar{X} = [X_1, \dots, X_N]$ be the vector of measurements (each X_i , $i = 1, \dots, N$, is a vector describing position and velocity of point i). Here we assume that there are no missing body parts and no clutter, that is, $N = M$. Let $\bar{L} = [L_1, \dots, L_N]$ be a vector of labels, where $L_i \in \mathcal{S}_{body}$ is the label of X_i . The labeling problem is to find \bar{L}^* , over all possible label vectors \bar{L} , such that the posterior probability of the labeling given the observed data is maximized, that is,

$$\bar{L}^* = \arg \max_{\bar{L} \in \mathcal{L}} P(\bar{L}|\bar{X}) \quad (1)$$

where $P(\bar{L}|\bar{X})$ is the conditional probability of a labeling \bar{L} given the data \bar{X} and \mathcal{L} is the set of all possible labelings. Using Bayes' law:

$$P(\bar{L}|\bar{X}) = P(\bar{X}|\bar{L}) \frac{P(\bar{L})}{P(\bar{X})} \quad (2)$$

If we assume that the priors $P(\bar{L})$ are equal for different labelings, then,

$$\bar{L}^* = \arg \max_{\bar{L} \in \mathcal{L}} P(\bar{X}|\bar{L}) \quad (3)$$

Given a labeling \bar{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 a specific body part associated with label L_i . For example if $L_i = LW$, i.e. the label corresponding to the left wrist is assigned to the i th point, then $X_i = X_{LW}$ is the position and velocity of the left wrist. Then,

$$P(\bar{X}|\bar{L}) = P_{S_{body}}(X_{LW}, X_{LE}, X_{LS}, X_H, \dots, X_{RF}) \quad (4)$$

where $P_{S_{body}}$ is the joint probability density function of the position and velocity of all the M body parts.

Three problems face us at this point: (a) What is the structure for the probability/likelihood function to be maximized? (b) How do we estimate its parameters? (c) How do we address the combinatorial search problem of finding the optimal labeling? Problems (a) and (c) need to be addressed together: the structure of the probability density function must be such that it allows efficient optimization.

A brute force solution to the optimization problem is to search exhaustively among all $M!$ (assuming no clutter, no missing body parts) possible \bar{L} 's and find the best one. The search cost is exponential with respect to M . Assume $M = 16$, then the number of possible labelings is larger than 2×10^{13} which is computationally prohibitive.

It is useful to notice that the body is a kinematic chain: for example, the wrist is connected to the body indirectly via the elbow and the shoulder. One could assume that the position and the velocity of the wrist are, therefore, independent of the position and velocity of the rest of the body once the position and velocity of elbow and shoulder are known. This

intuition may be generalized to the whole body: once the position and velocity of a set S of body parts is known, the behavior of body parts that are separated by S is independent. Of course this intuition is only an approximation which needs to be validated experimentally.

Our intuition on how to decompose the problem may be expressed in the language of probability: consider the joint probability density function of 5 random variables $P(A, B, C, D, E)$. It may be expressed as $P(A, B, C, D, E) = P(A, B, C)P(D|A, B, C)P(E|A, B, C, D)$. If these random variables are conditionally independent as described in the graph of Figure 3, then

$$P(A, B, C, D, E) = P(A, B, C)P(D|B, C)P(E|C, D) \quad (5)$$

Thus, if the body parts can satisfy the appropriate conditional independence conditions, we can express the joint probability density of the pose and velocity of all parts as the product of conditional probability densities of n -tuples. This approximation makes the optimization step computationally efficient as will be discussed below.

What is the best decomposition for the human body? What is a reasonable size n of the groups (or cliques) of body parts? We hope to make n as small as possible to minimize the cost of the optimization. But as n gets smaller, conditional independence may not be a reasonable approximation any longer. There is a tradeoff between computational cost and algorithm performance. In this paper we use models with $n = 3$ as described in Figure 2. Optimization on triangulated graphs such as these may be efficiently performed using Dynamic Programming [1].

2.2 Algorithms

What is needed is an algorithm that will search through all the legal labelings and find the one that maximizes the global joint probability density function. Notice that this optimum cannot be obtained by optimizing independently each triplet. By the decomposition in equation (5), we know that dynamic programming can be used to solve this problem efficiently.

The key condition for using dynamic programming is that the problem exhibits optimal substructure, namely, if equation (5) holds, then

$$\max_{A,B,C,D,E} P(A, B, C, D, E) = \max_{A,B,C} (P(A, B, C) \cdot \max_D (P(D|B, C) \cdot \max_E P(E|C, D)))$$

If we take the probability density function as the cost function, a dynamic programming method similar to that described in [1] can be used, which requires the triangulated body graph to be decomposable. If all the cliques in a graph are of size three, then the decomposable property means that there always exists a free vertex to delete and the remaining subgraph is again a collection of triangles until only one triangle is left. A vertex is free when it is only contained in one triangle. Figure 2 shows two decomposable graphs of the whole body, along with the order of successive elimination of the cliques.

If the decomposed body graph is decomposable and the corresponding conditional independence holds, then,

$$P_{S_{body}}(X_{LW}, X_{LE}, X_{LS}, X_H \dots X_{RF}) = \prod_{t=1}^{T-1} P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t}) \cdot P_{A_TB_TC_T}(X_{A_T}, X_{B_T}, X_{C_T}) \quad (6)$$

where $T = M - 2$ is the total number of triangles; (A_t, B_t, C_t) , $1 \leq t \leq T$, are the body parts associated to the triangle for stage t , A_t is the vertex to be deleted in stage t , and (B_t, C_t) are the two vertices (body parts) connected to A_t when A_t is deleted.

For each triplet (A_t, B_t, C_t) , we characterize it with a 10-dimensional feature vector

$$\mathbf{x} = (v_{Ax}, v_{Bx}, v_{Cx}, v_{Ay}, v_{By}, v_{Cy}, p_{Ax}, p_{Cx}, p_{Ay}, p_{Cy})^T \quad (7)$$

The first three dimensions of \mathbf{x} are the x -direction (horizontal) velocity of body parts (A_t, B_t, C_t) , the next three are the velocity in the y -direction (vertical), and the last four dimensions are the positions of body parts A_t and C_t relative to B_t . Relative positions are used here so that we can deal with translation invariance. As a first order approximation it is convenient to assume that \mathbf{x} is jointly Gaussian-distributed and therefore its parameters may be estimated from training data using standard techniques. After the joint probability

density function is computed, the conditional one can be obtained accordingly:

$$P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t}) = \frac{P_{A_tB_tC_t}(X_{A_t}, X_{B_t}, X_{C_t})}{P_{B_tC_t}(X_{B_t}, X_{C_t})} \quad (8)$$

where $P_{B_tC_t}(X_{B_t}, X_{C_t})$ can be obtained by estimating the joint probability density function of the vector $(v_{Bx}, v_{Cx}, v_{By}, v_{Cy}, p_{Cx}, p_{Cy})^T$.

Let

$$\Psi_t(X_{A_t}, X_{B_t}, X_{C_t}) = -\log P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t}), \text{ for } 1 \leq t \leq T - 1 \quad (9)$$

$$\Psi_t(X_{A_t}, X_{B_t}, X_{C_t}) = -\log P_{A_TB_TC_T}(X_{A_T}, X_{B_T}, X_{C_T}), \quad \text{for } t = T \quad (10)$$

be the cost function associate with each triangle, then the dynamic programming 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 cost 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 $\begin{cases} X_{A_1}^*_{[X_{B_1}, X_{C_1}]} = \arg \min_{X_{A_1}} T_1(X_{A_1}, X_{B_1}, X_{C_1}) \\ T_1(X_{A_1}^*_{[X_{B_1}, X_{C_1}]}, X_{B_1}, X_{C_1}) \end{cases}$

Stage t , $2 \leq t \leq T$: 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 cost so far (till stage t):

– Define $T_t(X_{A_t}, X_{B_t}, X_{C_t})$ the total cost 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, add the cost

$T_\tau(X_{A_\tau[X_{A_t}, X_{B_t}]}, X_{A_t}, X_{B_t})$ (or $T_\tau(X_{A_\tau[X_{B_t}, X_{A_t}]}, X_{B_t}, X_{A_t})$ if the edge was reversed) to $T_t(X_{A_t}, X_{B_t}, X_{C_t})$

– Likewise, add the costs 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 $\begin{cases} X_{A_t[X_{B_t}, X_{C_t}]}, X_{A_t} = \arg \min_{X_{A_t}} T_t(X_{A_t}, X_{B_t}, X_{C_t}) \\ T_t(X_{A_t[X_{B_t}, X_{C_t}]}, X_{B_t}, X_{C_t}) \end{cases}$

When stage T calculation is complete, $T_T(X_{A_T[B_T, C_T]}, X_{B_T}, X_{C_T})$ includes the value of each Ψ_t , $1 \leq t \leq T$, exactly once. Since the Ψ_t 's are the logs of conditional (and joint) probabilities, then if equation (6) holds,

$$T_T(X_{A_T[B_T, C_T]}, X_{B_T}, X_{C_T}) = -\log P_{S_{body}}(X_{LW}, X_{LE}, X_{LS}, X_H \dots X_{RF})$$

Thus picking the pair $(X_{B_T}^*, X_{C_T}^*)$ that minimizes T_T automatically maximizes the joint probability density function.

The best labeling can now be found tracing back through each stage: the best $(X_{B_T}^*, X_{C_T}^*)$ determines $X_{A_T}^*$, then the latest previous stages with edge respectively $(X_{A_T}^*, X_{B_T}^*)$, $(X_{A_T}^*, X_{C_T}^*)$, and/or $(X_{B_T}^*, X_{C_T}^*)$ determine more labels and so forth.

A simple example of this algorithm is shown in Figure 3.

The above algorithm is computationally efficient. Assume M is the number of body part labels and N ($N = M$ for this section) is the number of candidate markers, then the total number of stages is $T = M - 2$ and in each stage the computation cost is $\mathcal{O}(N^3)$. Thus, the complexity of the whole algorithm is on the order of $M * N^3$. In our implementation, the computational cost is further reduced by enforcing the relative position constraint: if two candidate markers of a triplet are too far away from each other, its cost will not be evaluated. Since we want to get the labeling with the highest probability, and the probabilities of the triplets with too large relative positions are very small, it's unnecessary to compute them.

3 Generalized Johansson problem: clutter and partial occlusion

In the previous section we dealt with the ideal case where all the body parts are present and no clutter points. But in reality, due to other moving patterns in the scene or the noisy output of feature detector/selector, there are often clutter points being detected. Due to occlusion, some body parts cannot be observed. In this section, we extend the algorithm to handle occlusion and clutter points. We call this 'generalized Johansson problem'.

It can be formulated as follows: given the positions and velocities of many points in an image plane (Figure 4 (a)), we want to find the most likely human configuration and decide whether it is a human body (Figure 4 (b) and (c)). In practice, the set of dots and associated velocities can be obtained from a low-level motion detector/feature tracker applied to the entire image [21]. In the following, we first address the labeling problem, i.e. how to find the most human-like configuration given a set of features.

3.1 Notation and description of the problem

Similar to section 2.1, the labeling problem can be described as follows. Suppose that we observe N points (as in Figure 4(a), where $N = 38$). We assign an arbitrary index to each

point. Then:

$$i \in 1, \dots, N \quad \text{Index} \quad (11)$$

$$\bar{X} = [X_1, \dots, X_N] \quad \text{Vector of measurements} \quad (12)$$

$$\bar{L} = [L_1, \dots, L_N] \quad \text{Vector of labels} \quad (13)$$

$$L_i \in \mathcal{S}_{body} \cup \{BG\} \quad \text{Possible values for each label} \quad (14)$$

Since there exist clutter points that do not belong to the body, the background label BG is added to the label set. Due to clutter and occlusion N is not necessarily equal to M (which is the size of \mathcal{S}_{body}). If we assume that the priors $P(\bar{L})$ are equal, then as in equation (3), we want to find

$$\bar{L}^* = \arg \max_{\bar{L} \in \mathcal{L}} P(\bar{X}|\bar{L})$$

Note that the equal priors assumption is only an approximation. For instance, if we have some prior knowledge on the number of background (clutter) points, $P(\bar{L})$ can be more precisely estimated. In [24], the number of clutter points is modeled with a Poisson distribution. However, it is hard to include this kind of global term in the dynamic programming algorithm as described in section 2.2. Hence we use the simplified equal priors assumption and validate it in experiments.

Let $\bar{\mathcal{L}}_{body}$ denote the set of body parts appearing in \bar{L} , \bar{X}_{body} be the vector of measurements labeled as body parts, and \bar{X}_{bg} be the vector of measurements labeled as background (BG). More formally, we group the measurements \bar{X} in two vectors \bar{X}_{body} and \bar{X}_{bg} ,

$$\begin{aligned} \bar{\mathcal{L}}_{body} &= \{L_i; i = 1, \dots, N\} \cap \mathcal{S}_{body} \\ \bar{X}_{body} &= [X_{i_1}, \dots, X_{i_K}] \quad \text{such that } \{L_{i_1}, \dots, L_{i_K}\} = \bar{\mathcal{L}}_{body} \\ \bar{X}_{bg} &= [X_{j_1}, \dots, X_{j_{N-K}}] \quad \text{such that } L_{j_1} = \dots = L_{j_{N-K}} = BG \end{aligned} \quad (15)$$

where K is the number of points described in $\bar{\mathcal{L}}_{body}$ (i.e. the size of $\bar{\mathcal{L}}_{body}$) and $N - K$ is the number of points in \bar{X}_{bg} , i.e. the number of background points.

If we assume that the position and velocity of the visible body parts is independent of position and velocity of clutter points, then,

$$P(\bar{X}|\bar{L}) = P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body}) \cdot P_{bg}(\bar{X}_{bg}) \quad (16)$$

where $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$ is the marginalized probability density function of $P_{S_{body}}$ (as in equation (4)) according to $\bar{\mathcal{L}}_{body}$. If independent uniform background noise is assumed, $P_{bg}(\bar{X}_{bg}) = (1/S)^{N-K}$, where $N - K$ is the number of background points, and S is the volume of the space the position and velocity of a background point lies in. In the following sections, we will address the issues of estimating $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$ and further find the \bar{L}^* with the highest likelihood.

3.2 Approximation of foreground probability density function

If no body part is missing, we can use equation (6) to get an approximation of the foreground probability density $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$. By the decomposable graph in Figure 2(a),

$$\begin{aligned} & P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body}) \\ &= P_{S_{body}}(X_{LW}, X_{LE}, X_{LS}, X_H \dots X_{RF}) \\ &= P_{LW|LE,LS}(X_{LW}|X_{LE}, X_{LS}) \cdot P_{LE|LS,LH}(X_{LE}|\dots) \cdot \dots \cdot P_{LA,LF,RF}(X_{LA}, X_{LF}, X_{RF}) \\ &= \prod_{t=1}^{T-1} P_t(X_{A_t}|X_{B_t}, X_{C_t}) \cdot P_T(X_{A_T}, X_{B_T}, X_{C_T}) \end{aligned} \quad (17)$$

Where T is the number of triangles in the decomposed graph in Figure 2(a), t is the triangle index, and A_t is the first body part associated to triangle t etc.

If some body parts are missing, then the foreground probability density function is the marginalized version of the above equation – marginalization over the missing body parts. Marginalization should be performed so that it is a good approximation of the true marginal probability density function and allows efficient computation such as dynamic programming as well. We propose that doing the marginalization term by term (triangle by triangle) of equation (17) and then multiplying them together is a reasonable way to get such an

approximation. The idea can be illustrated by a simple example as in equation (5) and Figure 3. For the graph in Figure 3, if A is missing, then the marginalized PDF is $P(B, C, D, E)$. If the conditional independence as in equation (5) can hold, then,

$$P(B, C, D, E) = P(B, C) \cdot P(D|B, C) \cdot P(E|C, D) \quad (18)$$

In the case of D missing, the marginalized PDF is $P(A, B, C, E)$. If we assume that E is conditionally independent of A and B given C , which is a more demanding conditional independence requirement than that of equation (5), then,

$$P(A, B, C, E) = P(A, B, C) \cdot 1 \cdot P(E|C) \quad (19)$$

Each term on the right hand sides of equations (18) and (19) is the marginalized version of its corresponding term in equation (5). Similarly, if some stronger conditional independence can hold, we can obtain an approximation of $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$ by performing the marginalization term by term of equation (17). For example, considering triangle (A_t, B_t, C_t) , $1 \leq t \leq T-1$, if all of A_t , B_t and C_t are present, then the t th term of equation (17) is $P_{A_t|B_t, C_t}(X_{A_t}|X_{B_t}, X_{C_t})$; if A_t is missing, the marginalized version of it is 1; if A_t and C_t are observed, but B_t is missing, it becomes $P_{A_t|C_t}(X_{A_t}|X_{C_t})$; if A_t exists but both B_t and C_t missing, it is $P_{A_t}(X_{A_t})$. For the T th triangle, if some body part(s) are missing, then the corresponding marginalized version of P_T is used. The foreground probability $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$ can be approximated by the product of the above (conditional) probability densities. Note that if too many body parts are missing, the conditional independence assumptions of the graphical model may no longer hold; it is reasonable to assume that the wrist is conditionally independent of the rest of the body given the shoulder and elbow, but if both shoulder and elbow are missing, this is no longer true. All the above (conditional) probability densities can be estimated from the training data. For instance, $P_{A_t|B_t, C_t}(X_{A_t}|X_{B_t}, X_{C_t})$ can be obtained via $P_{A_t, B_t, C_t}(X_{A_t}, X_{B_t}, X_{C_t})$ and $P_{B_t, C_t}(X_{B_t}, X_{C_t})$ as in section 2.2, equation (8) and similarly $P_{A_t|C_t}(X_{A_t}|X_{C_t})$ can be computed through $P_{A_t, C_t}(X_{A_t}, X_{C_t})$ and $P_{C_t}(X_{C_t})$.

3.3 Comparison of two labelings and cost functions for dynamic programming

The best labeling (\bar{L}^*) can be found by comparing all the possible labelings. To compare two labelings \bar{L}^1 and \bar{L}^2 , if we can assume the priors $P(\bar{L}^1)$ and $P(\bar{L}^2)$ are equal, then by equations (2) and (16),

$$\begin{aligned}
\frac{P(\bar{L}^1|\bar{X})}{P(\bar{L}^2|\bar{X})} &= \frac{P(\bar{X}|\bar{L}^1)}{P(\bar{X}|\bar{L}^2)} \\
&= \frac{P_{\bar{\mathcal{L}}_{body}^1}(\bar{X}_{body}^1) \cdot P_{bg}(\bar{X}_{bg}^1)}{P_{\bar{\mathcal{L}}_{body}^2}(\bar{X}_{body}^2) \cdot P_{bg}(\bar{X}_{bg}^2)} \\
&= \frac{P_{\bar{\mathcal{L}}_{body}^1}(\bar{X}_{body}^1) \cdot (1/S)^{N-K_1}}{P_{\bar{\mathcal{L}}_{body}^2}(\bar{X}_{body}^2) \cdot (1/S)^{N-K_2}} \\
&= \frac{P_{\bar{\mathcal{L}}_{body}^1}(\bar{X}_{body}^1) \cdot (1/S)^{M-K_1}}{P_{\bar{\mathcal{L}}_{body}^2}(\bar{X}_{body}^2) \cdot (1/S)^{M-K_2}} \tag{20}
\end{aligned}$$

where $\bar{\mathcal{L}}_{body}^1$ and $\bar{\mathcal{L}}_{body}^2$ are the sets of observed body parts for \bar{L}^1 and \bar{L}^2 respectively, K_1 and K_2 are the sizes of $\bar{\mathcal{L}}_{body}^1$ and $\bar{\mathcal{L}}_{body}^2$, and M is the total number of body parts ($M = 16$ here). $P_{\bar{\mathcal{L}}_{body}^i}(\bar{X}_{body}^i)$, $i = 1, 2$, can be approximated as in section 3.2. From equation (20), the best labeling \bar{L}^* is the \bar{L} which can maximize $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body}) \cdot (1/S)^{M-K}$. This formulation makes both search by dynamic programming and detection in different frames (possibly with different numbers of candidate features N) easy, as will be explained below.

At each stage of the dynamic programming algorithm described in section 2.2, the local optima are stored according to the total cost so far $T_t(X_{A_t}, X_{B_t}, X_{C_t})$, which is the sum of the local cost of the current triangle $\Psi_t(X_{A_t}, X_{B_t}, X_{C_t})$ and the costs of all the triangles on the path of the deletion of the current triangle. This requires that the local cost function $\Psi_t(X_{A_t}, X_{B_t}, X_{C_t})$ should be comparable for different labelings: whether there are missing part(s) or not. Therefore we cannot only use the terms of $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$, because, for example, as we discussed in the previous subsection, the t th term of $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body})$ is $P_{A_t|B_t,C_t}(X_{A_t}|X_{B_t}X_{C_t})$ when all the three parts are present and it is 1 when A_t is miss-

ing. It is unfair to compare $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$ with 1 directly. At this point, it is useful to notice that in $P_{\bar{\mathcal{L}}_{body}}(\bar{X}_{body}) \cdot (1/S)^{M-K}$, for each unobserved (missing) body part ($M - K$ in total), there is a $1/S$ term. $1/S$ (S is the volume of the space the position and velocity of a background point lies in) can be a reasonable local cost for a triangle with missing vertex A_t (the vertex to be deleted) because then for the same stage, the dimension of the domain of the local cost function is the same. Also, $1/S$ can be thought of as a threshold of $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$, namely, if $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$ is smaller than $1/S$, then the hypothesis that A_t is missing will win. Therefore, the local cost function ($\exp(-\Psi_t(X_{A_t}, X_{B_t}, X_{C_t}))$) for the t th ($1 \leq t \leq T - 1$) triangle can be approximated as follows:

- if all the three body parts observed, it is $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$;
- 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 . These approximations are to be validated in experiments. Notice that when two body parts in a triangle are missing, only velocity information for the third body part is available since we use relative positions. The velocity of a point alone doesn't have much information, so for two parts missing, we use the same cost function as the case of three body parts missing.

The approximation of the local cost functions described above can be illustrated by a simple example of Figure 3 (with $M = 5$). We want to compare a labeling $\bar{\mathcal{L}}^1$ with all five vertices (A, B, C, D, E) present and another labeling $\bar{\mathcal{L}}^2$ with D missing. By equations (5), (19) and (20), we need to compute,

$$\begin{aligned}
& \frac{P(A, B, C, D, E)}{P(A, B, C, E) \cdot (1/S)} \\
= & \frac{P(A, B, C) \cdot P(D|B, C) \cdot P(E|C, D)}{P(A, B, C) \cdot 1 \cdot P(E|C) \cdot (1/S)} \\
= & \frac{P(A, B, C)}{P(A, B, C)} \cdot \frac{P(D|B, C)}{(1/S)} \cdot \frac{P(E|C, D)}{P(E|C)}
\end{aligned} \tag{21}$$

The last line of equation (21) gives the local cost for each triangle.

With the local cost functions defined above, dynamic programming can be used to find the labeling with the highest $P_{\mathcal{L}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M-K}$. The computational complexity is on the order of $M * N^3$.

3.4 Detection

Given a hypothetical labeling \overline{L} , the higher $P(\overline{X}|\overline{L})$ is, the more likely it is that the associated configuration of features represents a person. The labeling \overline{L} with the highest $P_{\mathcal{L}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M-K}$ provides us with the most human-like configuration out of all the candidate labelings. Note that since the dimension of the domain of $P_{\mathcal{L}_{body}}(\overline{X}_{body}) \cdot (1/S)^{M-K}$ is fixed regardless of the number of candidate features and the number of missing body parts in the labeling \overline{L} , we can directly compare the likelihoods of different hypotheses, even hypotheses from different images.

In order to perform detection we first get the most likely labeling, then compare the likelihood of this labeling to a threshold. If the likelihood is higher than the threshold, then we will declare that a person is present. This threshold needs to be set based on experiments, to ensure the best trade-off between false acceptance and false rejection errors.

3.5 Integrating temporal information

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 extend our previous results to the case where multiple frames are available. However, in

order to maintain generality we will assume that tracking across more than two frames is impossible and therefore that the measurements from one pair of frames to the next are uncorrelated. This is a simplified model of the situation where, due to extreme body motion or to loose and textured clothing and occlusion, tracking is extremely unreliable and each feature’s lifetime is short. Neri et al. [16] used similar assumption when conducting their psychophysical investigation of biological motion perception in the human visual system.

Let $P(O|\bar{X})$ denote the probability of the existence of a person given \bar{X} observed. From equation (20) and the previous subsection, we use the approximation: $P(O|\bar{X})$ is proportional to $\Gamma(\bar{X}|\bar{L}^*)$ defined as $\Gamma(\bar{X}|\bar{L}^*) \stackrel{\text{def}}{=} \max_{\bar{L} \in \mathcal{L}} P_{\bar{L}_{body}}(\bar{X}_{body}) \cdot (1/S)^{M-K}$, where \bar{L}^* is the best labeling found from \bar{X} . Now if we have n observations $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$, then the decision depends on:

$$\begin{aligned}
& P(O|\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n) \\
= & \frac{P(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n|O) \cdot P(O)}{P(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n)} \\
= & \frac{P(\bar{X}_1|O)P(\bar{X}_2|O) \dots P(\bar{X}_n|O) \cdot P(O)}{P(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n)} \tag{22}
\end{aligned}$$

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

3.6 Counting

Counting how many people are in the scene is also an important task since images often have multiple people in them. By the method described above, we can first get the best configuration to see if it could be a person. If so, all the points belonging to the person are deleted and the next best labeling can then be found from the remaining points. We

can keep doing this until the likelihood of the best configuration is smaller than a threshold. Then the number of configurations with likelihood greater than the threshold is the number of people in the scene.

Assume M is the number of body labels, N is the number of candidate markers, and P is the number of people in the scene. By section 2.2, the cost of detecting the first person is on the order of $M * N^3$, the cost for the second person is of $M * (N - m_1)^3$, where m_1 , ($m_1 \leq M$), is the number of body parts present in the first person, and so on. Therefore, the total cost of counting P individuals is of $P * M * N^3$.

4 Experiments

We did experiments on data obtained filming a subject moving freely in 3D; 16 light bulbs were strapped to the main joints of the subject’s body. In order to obtain ground-truth the data were first acquired, reconstructed and labeled in 3D using a 4-camera motion capture system operating at a rate of 60 samples/sec. Since our goal is to detect and label the body directly in the camera image plane, a generic camera view was simulated by orthographic projection of the 3D marker coordinates. In the following sections we will control the camera view with the azimuth viewing angle: a value of 0 degrees will correspond to a right-side view, a value of 90 to a frontal view of the subject. Six sequences were acquired each around 2 minutes long. In the next sections they will be referred as follows: Sequences W1 (7000 frames), W2 (7000 frames): relaxed walking forward and backwards along almost straight paths (with ± 20 degree deviations in heading); W3 and W4 (6000 frames each): relaxed walking, with the subject turning around now and then (Figure 5(a) shows sample frames from W3); Sequence HW (5210 frames): walking in a happy mood, moving the head, arms, hips more actively (Figure 5(b)); Sequence DA (3497 frames): dancing and jumping (Figure 5(c)), with the subject moving his legs and arms freely and much faster than in the previous four sequences. Given that the data were acquired from the same subject and that orthographic projection was used to simulate a camera view, our data were already

normalized in scale. The velocity of each candidate marker was obtained by subtracting the positions in two consecutive frames. Thus, to get velocity information, we assumed that features could be tracked for two frames but we didn't use any feature correspondence over more than two frames, which is arguably the most difficult conditions under which to perform labeling and detection, as mentioned in section 3.5.

4.1 Labeling

We first explored the labeling performance of the algorithm described in section 2 on frames with all the body parts detected and no clutter points. Among the sequences, walking sequences W1 and W2 are the relatively simple ones, so W1 and W2 were first used to test the validity of the Gaussian probabilistic model and the performance of two possible body decompositions (Figure 2). Since the heading direction of W1 and W2 was roughly along a line, these sequences were also used to study the performance as a function of viewing angle. Then experiments were conducted using W3, HW and DA to see how the model worked for more active and non-periodic motions.

4.1.1 Detection of individual triangles

In this section, the performance of the Gaussian probabilistic model for individual triangles is examined. In the training phase, the joint Gaussian parameters (mean and covariance) for each triangle in Figure 2 were estimated from walking sequence W1 (viewed with a 45 degrees viewing angle). In the test phase, for each frame in W2 (also viewed of 45 degrees), each triangle probability was evaluated for all possible combinations of markers ($16 \times 15 \times 14$ different combinations). Ideally, the correct combination of markers should produce the highest probability for each respective triangle. Otherwise, an error occurred. Figure 6(a) shows how well each triangle's joint probability model detects the correct set of markers. Figure 6(b) shows a similar result for the conditional probability densities of triangles, where for each triangle conditional probability density $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$, we

computed $P_{A_t|B_tC_t}(X_{A_t}|X_{B_t}, X_{C_t})$ for all the possible choices of A_t (14 choices), given the correct choice of markers for B_t and C_t . Figure 6 shows that the Gaussian model is very good for most triangles (in the joint case, if a triangle is chosen randomly, then the chance of getting the correct one is 3×10^{-4} and the probability models do much better than that).

It is not surprising that the performance of some triplets is much worse than others. The worst triangles in Figure 6(a) are those with left and right knees, which makes sense because the two knees are so close in some frames that it is even hard for human eyes to distinguish between them. Therefore, it is also hard for the probability model to make the correct choice.

Further investigation of the behavior of the triangle probabilities revealed that, for frames in which the correct choice of markers did not maximize a triangle probability, that probability was nevertheless quite close to the maximal value. Figure 7 shows the ratio of the probabilities of the correct choice over the maximizing choice for the two worst behaving triangles, over the set of frames where the errors occurred. Figure 7(a) shows the ratio of the joint probability distribution for triangle 10 (consisting of right hip, left knee, and right knee, as in figure 2(a)). Figure 7(b) shows the ratio of the conditional probability distribution for triangle 17 (head, neck, and left shoulder). Although these two triangles had the highest error rates, the correct marker combination was always very close to being the highest ranking, always less than a factor of 1.006 away. This is a good indication that the individual triangle probability models encode the distribution quite well.

4.1.2 Performance of different body graphs

We did experiments on the two decompositions in Figure 2. The training sequence W1 and the test sequence W2 were under the same viewing angle: 45 degrees, which is between the side view and the front view. Table 1 shows the results. The *frame-by-frame error* is the percentage of frames in which errors occurred, and *label-by-label error* is the percentage of markers wrongly labeled out of all the markers in all the testing frames. Label-by-label

error is smaller than frame-by-frame error because an error in a frame does not mean all the markers are wrongly labeled.

decomposition model	(a)	(b)
frame-by-frame error	0.27%	13.13%
label-by-label error	0.06%	1.61%

Table 1: **Error rates using the models in Figure 2**

The performance of the algorithm using the decomposition of Figure 2(a) is almost perfect and much better than that of (b), which is consistent with our expectation (by Figure 6, the local performance of decomposition Figure 2(a) is better than that of Figure 2(b)). We used the better model in the rest of the experiments.

4.1.3 Viewpoint invariance

In the previous sections the viewing angle for training and for testing was the same. Here we explore the behavior of the method when the testing viewing angle is different from that used during training. Figure 8 shows the results of three such experiments where walking sequence W1 was used as the training set and W2 as the test set .

The solid line in Figure 8(a) shows the percentage of frames labeled correctly when the training was done at a viewing angle of 90 degrees (subject facing the camera) and the testing viewing angle was varied from 0 degrees (right side view) to 180 degrees (left side view) in increments of 10 degrees. When the viewing angle was between 60 to 120 degrees almost all frames were labeled correctly, thus showing that the probabilistic model learned at 90 degrees is insensitive to changes in viewpoint by up to 30 degrees.

The solid line in Figure 8(b) shows the results of a similar experiment where the training viewpoint was at 0 degrees (right side view) and the testing angle was varied from -90 degrees (back view) to 90 degrees (front view) in 10 degree increments. A noticeable dip in the performance centered around 0 degrees is visible in the plot. Inspection of the errors which occurred at these viewing angles revealed that they consisted solely of confusions

between homologous left-right leg parts; i.e., the two hips were sometimes confused, as were the knees, the ankles, and the feet. Considering that an orthographic projection of the 3-D data was used to create the 2-D views, this result is not surprising; given an orthographic side view of a person walking (with no self-occlusions) a person viewing the motion is unable to distinguish the left and right sides of the body. Thus, modulo this left-right ambiguity, the model learned at 0 degrees viewing angle is insensitive to changes in viewpoint of up to 50 degrees.

The dashed line in Figure 8(a) shows the results of an experiment of trying to increase the invariance of the probabilistic model with respect to changes in viewpoint. The same 3-D training sequence was used to generate three 2-D data sequences with viewing angles at 30, 90, and 150 degrees. The three 2-D sequences were combined, and used all together to learn the probability density functions of the graph triangles. As shown in the plot, this procedure does in fact improve the labeling accuracy. At 0 degrees, the only errors were the above mentioned left-right ambiguity within the legs. Between 10 and 60 degrees, besides left-right errors, also the feet and ankles were confused. From 120 to 180 degrees, the errors once again consisted solely of swapped left and right body parts.

4.1.4 Performance with different motions

The previous sections show that for simple motions very good results can be achieved using the probabilistic model. Here we want to investigate how the method works for more general sets of motions. We did experiments on walking sequence W3, happy walking sequence HW and dancing sequence DA. Each sequence was divided into four segments for a total of twelve segments. To test a segment, frames from all the other eleven segments were used as the training set. The error rates for different sequences are obtained by averaging the results of the corresponding segments.

Table 2 shows the error rates for different sequences. The first column is the average result for all the three sequences, and the next three columns show the error rates for

test set	ALL	W3	HW	DA
frame-by-frame error	6.81%	3.02%	4.49%	15.95%
label-by-label error	0.69%	0.38%	0.50%	1.45%

Table 2: **Error rates for different sequences.** ALL: average over all three sequences; W3: walking sequence; HW: walking in happy mood; DA: dancing sequence

walking sequence W3, happy walking sequence HW and dancing sequence DA respectively. The results for walking sequence W3 and happy walking sequence HW are very good, with *frame-by-frame error* less than 5% and *label-by-label error* no more than 0.5%. It is not surprising that the error rates of dancing sequence are higher than the walking sequences because the motions in the dancing sequence are more random and agitated and therefore harder to model. Another possible reason is that the dancing sequence is shorter than the other sequences, so the motion of dancing has relatively less weight in the training set.

Figure 9 shows the error rate of each individual body part for each of the sequences. Notice that most errors occur at the left and right wrist (LW and RW) in the dancing sequence. This is because in the dancing sequence the wrists are very close to hips in some frames, and the program mistook the hip markers as being the wrists. The reason why the program wouldn't mistake wrist markers as hips is that hips have better motion constraints than wrists. In our decomposed body graph Figure 2(a), both left and right hip (LH and RH) appear in five triangles, but the wrists (LW and RW) are only in one triangle each.

4.2 Detection - clutter and partial occlusion

In this section we show the results of using the algorithm described in section 3 when there are clutter points and/or occlusions in the scene. We perform experiments to analyze the detection rate as a function of the number of visible body parts, with and without integration of temporal information. We also test the system with different types of clutter statistics, and analyze the performance of estimating the number of people in the scene.

4.2.1 Detection

In this experiment, we test how well our method can distinguish whether or not a person is present in the scene (Figure 4). To do so, we present the algorithm with two types of inputs (presented randomly in equal proportions); in one case only clutter (background) points are present, in the other a pre-determined number of randomly selected body parts in the set of test data are superimposed on some clutter. If there are body parts in the scene and the program thinks there is person, the person is correctly detected. If there are only background points in the scene but the program thinks there is person, a false alarm happens. We measure the frequency of correct detections and false alarms, and build receiver operating characteristics (ROC) curves for our detector.

We want to test the detection performance when only part of the whole body (with 16 body parts in total) can be seen. We generated the signal points (body parts) in a frame in the following way: for a fixed number of signal points, we randomly selected which body parts to be used for each frame (actually pair of frames, since consecutive frames are used to estimate the velocity of each body part). So in principle, each body part has an equal chance to be represented, and as far as the decomposed body graph is concerned, all kinds of graph structures (with different body parts missing) can be tested.

The positions and velocities of clutter (background) points were independently generated from uniform distributions of their corresponding ranges. For positions, we used the leftmost and rightmost positions of the whole sequence as its horizontal range, and highest and lowest body part positions as its vertical range. For velocities, the possible range is inside a circle of the velocity space (horizontal and vertical velocities) with radius of the maximum magnitude of the velocities from the real sequences. Figure 4 (a) shows a frame with 8 body parts and 30 added background points with arrows representing velocities.

Figure 10 shows the experimental results on walking sequences (sequences W3 and W4, sequence W3 was used for the training and W4 for testing). The six solid curves of Figure

10 (a) shows the receiver operating characteristics (ROCs) of 3 to 8 signal points with 30 added background points vs. 30 background points. The bigger the number of signal points observed, the better the ROC is. With 30 background points, when the number of signal points is more than 8, the ROCs are almost perfect.

In practice, when using the detector, some detection threshold needs to be set; if the likelihood of the best labeling of the scene exceeds the threshold, a person is deemed to be present. Since the number of body parts is unknown before detection, we need to fix a threshold that is independent of (and robust with respect to) the number of body parts present in the scene. The dashed line in Figure 10 (a) shows the overall ROC of all the frames used for the six ROC curves in solid lines. We took the threshold when $P_{detect} = 1 - P_{false-alarm}$ on that curve as our threshold. The star ('*') point on each solid curve shows the point corresponding to the threshold. Figure 10 (b) shows the relation between detection rate and the number of body parts displayed with regard to the fixed threshold. The corresponding false alarm rate is 12.97%. From Figure 10 we can see that even when only three body parts are present in the scene, the detection performance is much better than the chance level.

When the algorithm can correctly detect whether a person is there, it doesn't necessarily mean that all the body parts are correctly labeled. So we also studied the correct label rate (*label-by-label rate*) when a person is correctly detected. An error happens when a body part is assigned a wrong candidate feature. Figure 10 (c) shows the result. While the detection rate keeps constant (almost 1) with 8 or more body parts visible, the correct label rate goes up as the number of body parts increases. The correct label rates here are smaller than the results in section 4.1 since we have less signal points but many more background points. If the average number of features detected is N , (N is more than 30 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, more than 50% with only 3 body parts (almost 20 times above chance level) and exceeding

90% when 12 out of the 16 body parts are present.

4.2.2 Using temporal information

Here we tested how the detection rate improved by integrating information over time, using the approach described in section 3.5. We used the data of 5 signal points and 30 background points in each frame to test the performance of using information from multiple frames (the body parts present in each frame were chosen randomly and independently). Figure 11 (a) shows ROC curves of using n ($n = 1, \dots, 8$) frames. The bigger n is, the better the ROC curve is. When $n > 5$, the ROCs are almost perfect and overlapped with the axes. If Θ is the likelihood threshold of $P_{detect} = 1 - P_{false-alarm}$ when only one frame is used, then the threshold of $P_{detect} = 1 - P_{false-alarm}$ for using n frames is Θ^n . Figure 11 (b) plots the detection rate (with $P_{detect} = 1 - P_{false-alarm}$) vs. the number of frames integrated. From the plots, we see that the results get better with more frames used, and even with only 5 body parts present it is possible to get completely accurate detection after combining information from only 6 frames.

4.2.3 Biological Clutter

We also tested our method by using biological clutter, which means, the background points were generated by independently drawing points (with position and velocity) of randomly chosen frames and body parts from the walking sequence. Figure 12 shows the results. Eight solid curves in Figure 12(a) are ROCs for 3 to 10 body parts and 30 background points. The dashed line is the overall ROC for all the frames used. We choose the threshold on that curve when $P_{detect} = 1 - P_{false-alarm}$ and get the detection rates (shown by stars in 12(a)), with false alarm rate 24.19%. The solid line (with stars) in 12(b) shows the relation between the detection rate and the number of signal points. Comparing Figure 10 and Figure 12(a), we can see that the performance is better in Figure 10, which means that the detection task is easier if the background points are generated in the previous way. This is consistent with our intuition since biological clutter has the same single point statistics as the signal points.

We also did experiments with less number of background points. The dashed line (with triangles) in Figure 12 is the detection rate vs. the number of signal points with 20 added background points. The false alarm rate is 19.45%. The result of 20 background points is better than that of 30 background points. Less background points make the task easier.

4.2.4 Counting Experiments

The counting task is to find how many people are in a scene given a number of observed points (with position and velocity). A person was generated by randomly choosing a frame from the sequence, and several frames (persons) can be superimposed together into one image. In one image, the position of each person was randomly selected, but made sure not to overlap with each other. The background points were generated in a similar way to section 4.2.1, but with the positions of the background points uniformly distributed on a window which is three times as wide as the window in Figure 4 (a). Figure 13 gives an example of images used in this experiment, with three persons (six body parts each) and sixty background points.

We did experiments on up to three persons in one image. We used the threshold from Figure 10. If the likelihood of the configuration found was above the threshold, then it was counted as a person. If the number of detected people provided by the algorithm was different (either more or less) from the ground truth, an error happened. The curves in Figure 14 show the correct rate vs. the number of signal points (body parts displayed) for each person. To compare the results conveniently, we used the same number of body parts for different persons in one image (but the parts appearing were randomly chosen). The solid line with stars is the result of one person in an image, the dashed line with circles is for two persons, and the dash-dot line with triangles is for three persons. If there was no person in the image, the correct rate is 95%. From Figure 14, we see that the result for less people in an image is better than that of more people, especially when the number of body parts present is small. We can explain it as follows. If the probability of counting one person correctly is P , then the probability of counting n people correctly is P^n if the detection of

different people is independent. For example, in the case of four body parts, for one person the correct rate is 0.6, then the correct rate for counting three person is $0.6^3 = 0.216$. But since we randomly chose the position of each person, body parts from different persons may be very close, so the independence couldn't be strictly held. Furthermore, the assumption of independence is also violated since once a person is detected the corresponding body parts are removed from the scene in order to detect subsequent people.

4.2.5 Experiments on dancing sequence

In this section, we performed detection experiments on the dancing sequence DA (the first half was used for training and the second half for testing). The seven solid curves of Figure 15 (a) are the ROC curves of 4 to 10 signal points with 30 added background points. The signal points are from the dancing sequence and the background points were generated the same way as in Section 4.2.1. In Figure 15 (a), the bigger the number of signal points observed, the better the ROC is. The dashed line in Figure 15 (a) shows the overall ROC of all the frames used for the seven ROC curves in solid line. We took the threshold when $P_{detect} = 1 - P_{falsealarm}$ on that curve as our threshold and get the plot of detection rate vs. the number of signal points in Figure 15 (b). The false alarm rate is 14.67%. With more than 9 (out of 16) body parts present, the detection rate is almost 1. Comparing with the results in Figure 10, we can see that more body parts must be observed during the dancing sequence to achieve the same detection rate as with the walking sequences, which is expected since the motion of dancing sequences is more active and harder to model. Nevertheless, the ROC curve with 10 out 16 body parts present is nearly perfect.

5 Conclusions and future work

We have built a probabilistic perceptual model of biological motion. The labeling problem is solved by finding the set of labels which maximizes the likelihood of the observed data. The detection problem is solved by comparing this maximum likelihood to a threshold. The

model has a Markov-like structure, therefore dynamic programming may be used to find efficiently the globally optimal solution.

The method was tested on several types of motions and has an overall label-by-label error rate of 0.7% (with all body parts present). It detects biological motion reliably even when large portions of the body are occluded and when clutter is present. 2-frame detection rate is better than 90% when 6 out of 16 body parts are seen within 30 clutter points (see Figure 10). When the number of frames considered exceeds 5 then performance quickly reaches 100% correct (see Figure 11). Note that this is a lower bound on performance, as we assumed short feature lifetime and did not make use of time correlation in feature position and velocity. This means that even in high-noise conditions detection is almost flawless in 100ms or so, a figure comparable to the alleged performance of the human visual system. Moreover, our algorithm is computationally efficient, taking approximately 1 second in our Matlab implementation on a regular Pentium II 450 MHz computer, which gives significant hope for a real-time C implementation on the same computer.

The next step in our work is clearly the application of our system to real image sequences, rather than Johansson displays. We anticipate using a simple feature/patch detector and tracker in order to provide the position-velocity measurements that are input in our system. Since our system can work with features that have a short life-span (in the limit 2-frame) this should be feasible without modifying the overall approach. A first set of experiments is described in [21], where features tracked by a Lucas-Tomasi-Kanade tracker [22] are used instead of the Johansson dots. The results are encouraging in that performance is comparable to the experiments conducted in this paper, even when training and testing were performed on different subjects. However, those results are only preliminary since they refer to only one type of motion (walking) imaged under one viewpoint.

Note that the body parts often appear/disappear in groups (for example a whole arm/leg may be occluded by the body). This correlation may be embedded in the model. Other extensions include training on a larger set of motions, using different probability density

functions that are more sophisticated than the Gaussian, dealing with different scales and extending viewpoint invariance to 360^0 .

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).

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] A. Blake and M. Isard, “3d position, attitude and shape input using video tracking of hands and lips”, In *Proc. ACM Siggraph*, pages 185–192, 1994.
- [3] C. Bregler and J. Malik, “Tracking people with twists and exponential maps”, In *Proc. IEEE CVPR*, pages 8–15, 1998.
- [4] J.E. Cutting and L.T. Kozlowski, “Recognizing friends by their walk: Gait perception without familiarity cues”, *Bulletin Psychonomic Society*, 9:353–356, 1977.
- [5] W. Dittrich, T. Troscianko, S. Lea, and D. Morgan, “Perception of emotion from dynamic point-light displays represented in dance”, *Perception*, 25:727–738, 1996.
- [6] K.H. Fielding and D.W. Ruck, “Recognition of moving light displays using hidden markov-models”, *Pattern Recognition*, 28:1415–1421, 1995.
- [7] D. Gavrilu, “The visual analysis of human movement: A survey”, *Computer Vision and Image Understanding*, 73:82–98, 1999.

- [8] L. Goncalves, E. Di Bernardo, E. Ursella, and P. Perona, “Monocular tracking of the human arm in 3d”, In *Proc. 5th Int. Conf. Computer Vision*, pages 764–770, Cambridge, Mass, June 1995.
- [9] 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.
- [10] D.D. Hoffman and B.E. Flinchbaugh, “The interpretation of biological motion”, *Biological Cybernetics*, 42:195–204, 1982.
- [11] 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.
- [12] A. Jepson and M.J. Black, “Mixture models for optical flow computation”, In *Proc. IEEE CVPR*, pages 760–761, 1993.
- [13] G. Johansson, “Visual perception of biological motion and a model for its analysis”, *Perception and Psychophysics*, 14:201–211, 1973.
- [14] H.M. Lakany and G.M. Hayes, “An algorithm for recognising walkers”, *LECT NOTES COMPUT SC: Audio- and Video-Based Biometric Person Authentication*, 1206:111–118, 1997.
- [15] G. Mather and L. Murdoch, “Gender discrimination in biological motion displays based on dynamic cues”, *Proc. R. Soc. Lond. B*, 259:273–279, 1994.
- [16] P. Neri, M.C. Morrone, and D.C. Burr, “Seeing biological motion”, *Nature*, 395:894–896, 1998.
- [17] R. Rashid, “Towards a system for the interpretation of moving light displays”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:574–581, 1980.

- [18] J.M. 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.
- [19] 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.
- [20] J. Shi and C. Tomasi, “Good features to track”, In *Proc. IEEE CVPR*, pages 593–600, 1994.
- [21] Y. Song, X. Feng, and P. Perona, “Towards detection of human motion”, In *Proc. IEEE CVPR*, volume 1, pages 810–817, June 2000.
- [22] C. Tomasi and T. Kanade, “Detection and tracking of point features”, *Tech. Rep. CMU-CS-91-132, Carnegie Mellon University*, 1991.
- [23] S. Wachter and H.-H. Nagel, “Tracking persons in monocular image sequences”, *Computer Vision and Image Understanding*, 74:174–192, 1999.
- [24] M. Weber, M. Welling, and P. Perona, “Unsupervised learning of models for recognition”, In *Proc. ECCV*, volume 1, pages 18–32, June/July 2000.
- [25] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:780–785, 1997.
- [26] M. Yang and N. Ahuja, “Extracting gestural motion trajectories”, In *International Conference on Face and Gesture Perception*, pages 10–15, Nara, Japan, April 1998.

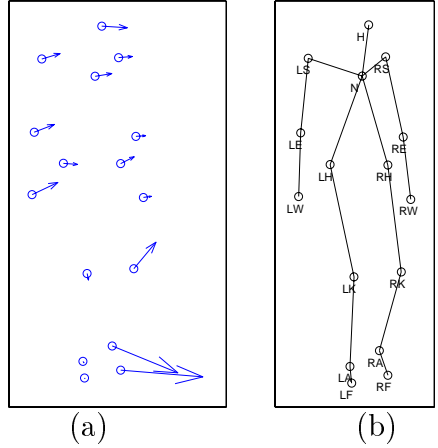


Figure 1: **The labeling problem (without clutter and missing points):** Given the position and velocity of body parts in the image plane (a), we use a probabilistic model to assign the correct labels to the body parts (b). 'L' and 'R' in label names indicate left and right. H:head, N:neck, S:shoulder, E:elbow, W:wrist, H:hip, K:knee, A:ankle and F:foot.

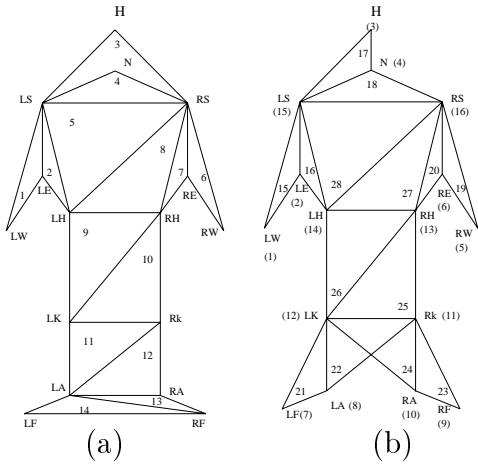


Figure 2: **Two decompositions of the human body into triangles.** 'L' and 'R' in label names indicate left and right. H:head, N:neck, S:shoulder, E:elbow, W:wrist, H:hip, K:knee, A:ankle and F:foot. The numbers inside triangles give the index of triangles used in the experiments. In (a) they are also the order in which the vertices are deleted. In (b) the numbers in brackets show the order.

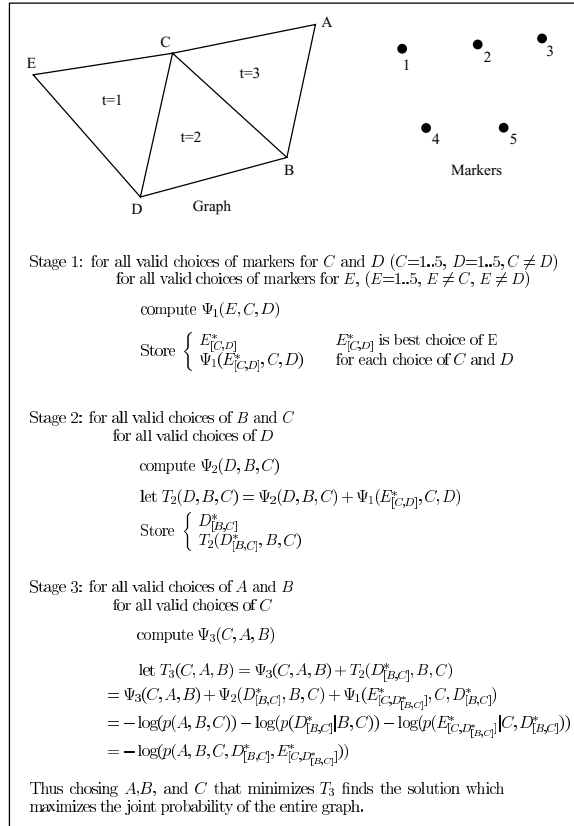


Figure 3: An example of dynamic programming algorithm applied to a simple graph

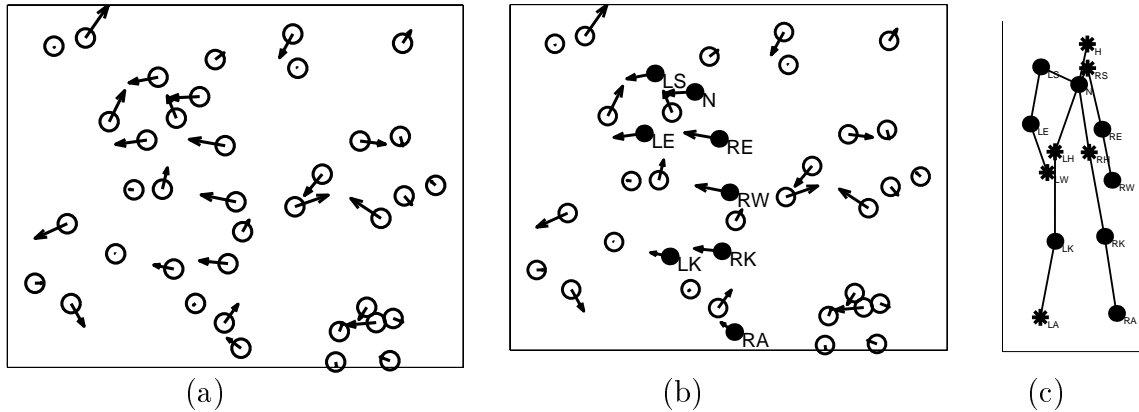


Figure 4: **Detection and labeling under the conditions of clutter and occlusion:** Given the position and velocity of dots in an image plane (a), we want to find the most possible human configuration: filled dots in (b) are body parts and circles are background points. Detection is done according to the likelihood of this best configuration. Arrows in (a) and (b) show the velocities. (c) is the full configuration of the body. Filled (blackened) dots representing those present in (b), and the '*'s are actually missing (not available to the program). The body part label names are the same as in Figure 1.

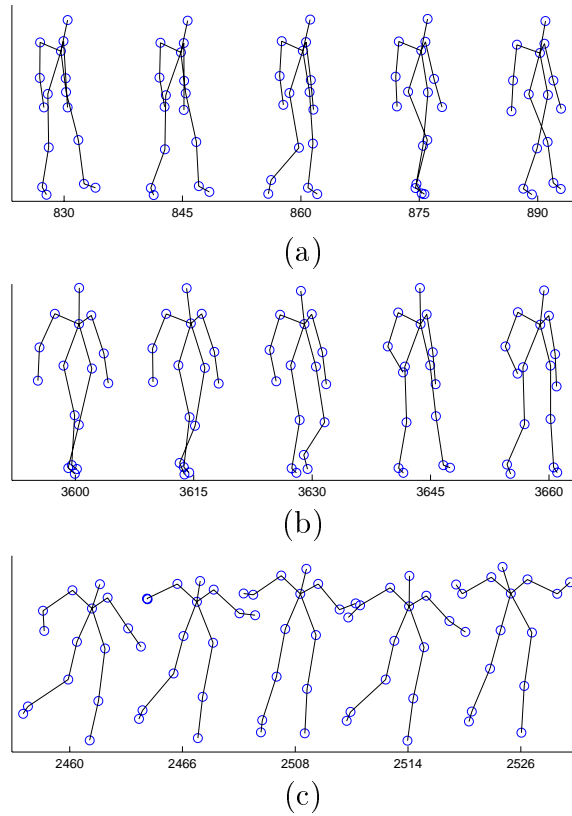


Figure 5: **Sample frames** for the (a) walking sequence W3; (b) happy walking sequence HW; (c) dancing sequence DA. The numbers on the horizontal axes are the frame numbers.

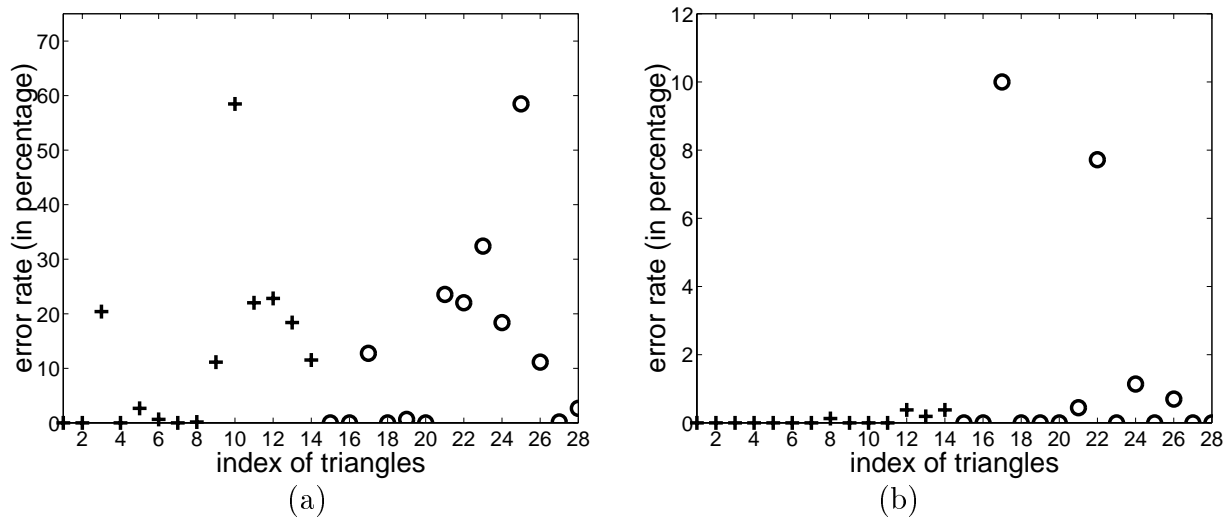


Figure 6: **Local model error rates** (percentage of frames for which the correct choice of markers did not maximize each individual triangle probability). Triangle indices are those of the two graph models of Figure 2. '+' : results for decomposition Figure 2(a); 'o' : results for decomposition Figure 2(b). (a) joint probability model (b) conditional probability model

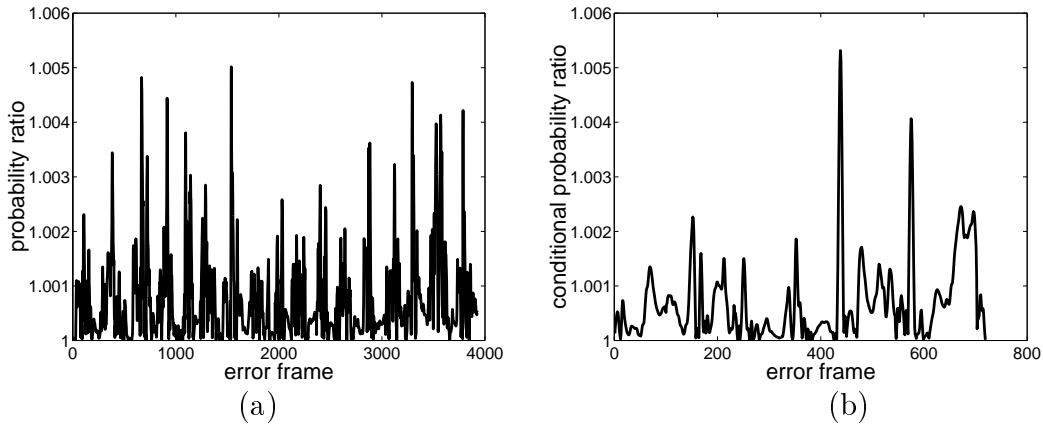


Figure 7: **probability ratio (correct markers vs. the solution with the highest probability when an error happens.)** The horizontal axis is the index of frames where error happens. **(a)** joint probability ratio for triangle 10 or 25 (RH, LK, RK) **(b)** conditional probability ratio for triangle 17 (H, N, LS)

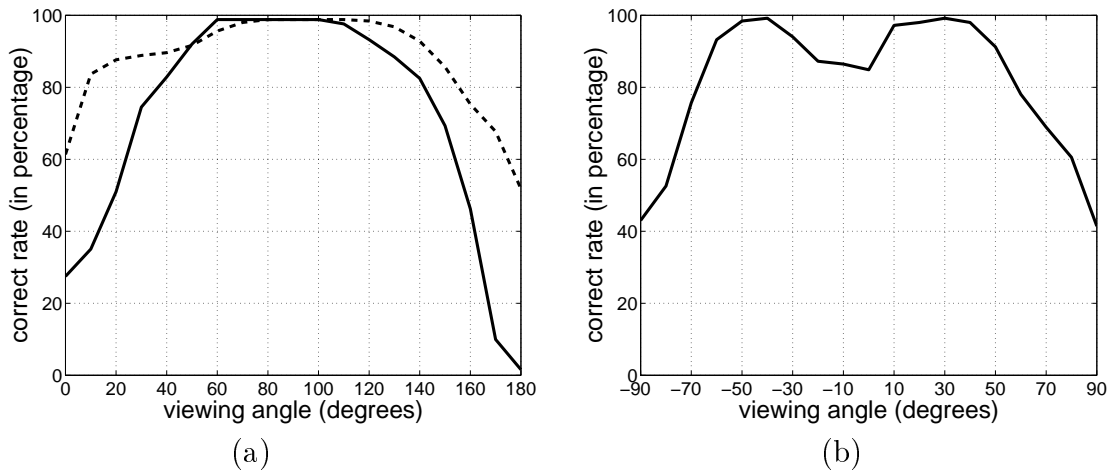


Figure 8: **Labeling performance as a function of viewing angle.** **(a)** Solid line : percentage of correctly labeled frames as a function of viewing angle, when the training was done at 90 degrees (frontal view). Dashed line: training was done by combining data from views at 30, 90, and 150 degrees. **(b)** Labeling performance when the training was done at 0 degrees (right side view of walker). The dip in performance near 0 degrees is due to the fact that from a side view orthographic projection without body self-occlusions it is almost impossible to distinguish left and right.

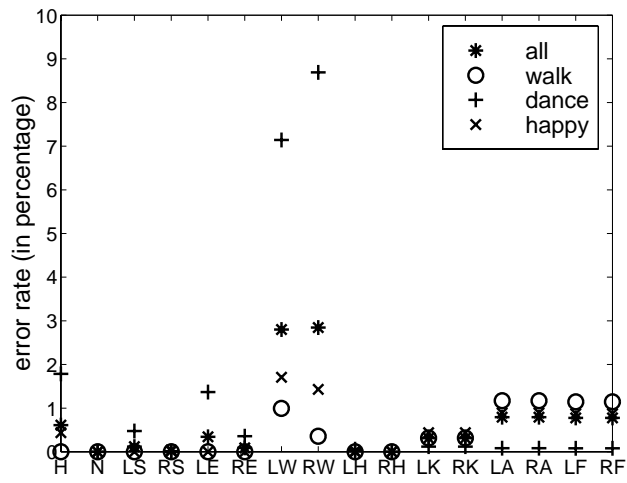


Figure 9: **Error rates for individual body parts.** 'L' and 'R' in label names indicate left and right. H:head, N:neck, S:shoulder, E:elbow, W:wrist, H:hip, K:knee,A:ankle and F:foot. See section 4.1.4.

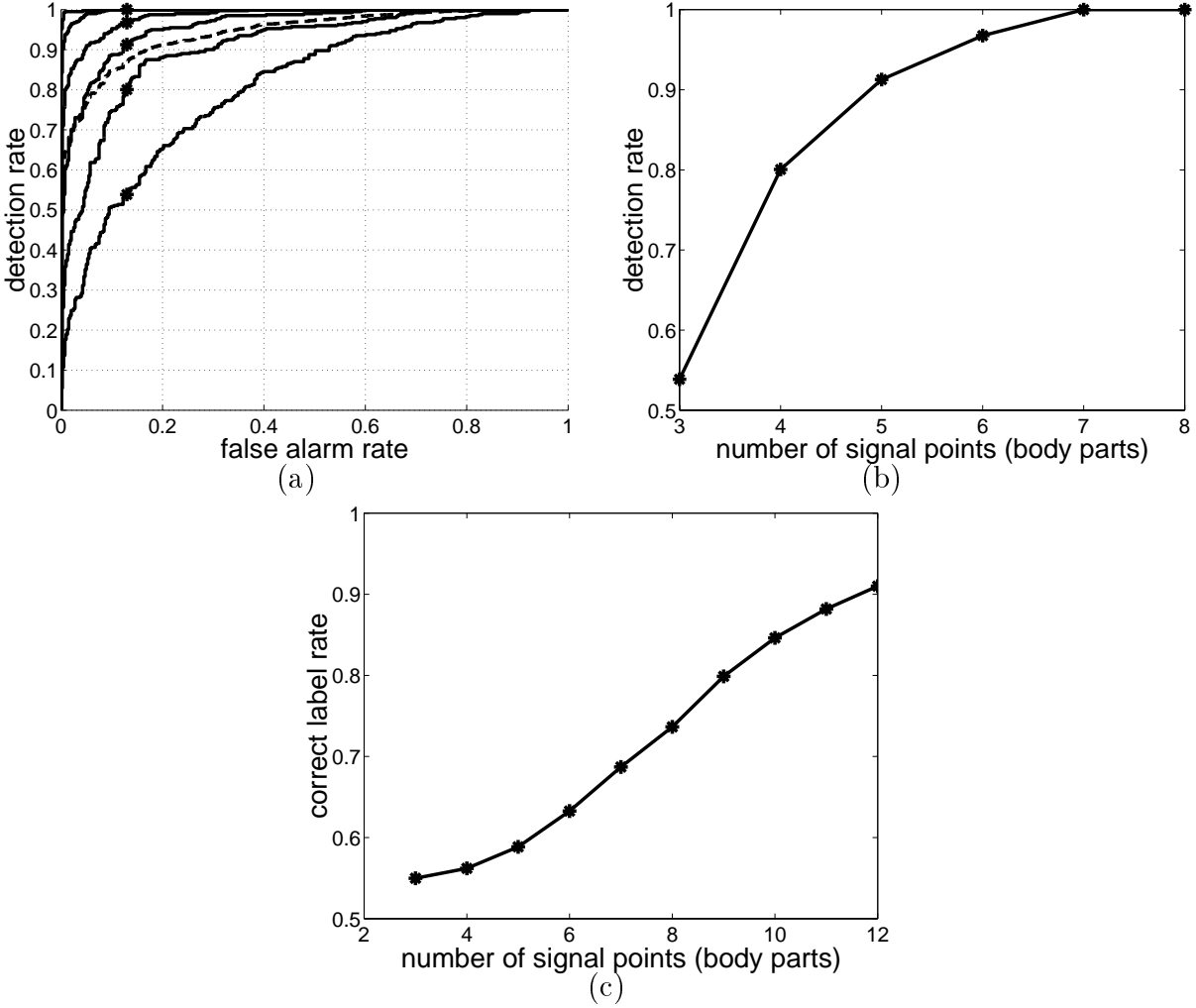


Figure 10: **Detection results** (under the conditions of clutter and occlusion). **(a)** ROC curves. Solid lines: 3 to 8 body parts with 30 background points vs. 30 background points only. The bigger the number of signal points is, the better the ROC is; dashed line: overall ROC considering all the frames used in six solid ROCs. The threshold corresponding to $P_D = 1 - P_{FA}$ on this curve was used for later experiments. The stars (*) on the solid curves are the points corresponding to that threshold. **(b)** detection rate vs. number of body parts displayed with regard to the fixed threshold at which $P_D = 1 - P_{FA}$ on the overall ROC curve in (a) (with false alarm rate 12.97%). **(c)** correct label rate (label-by-label rate) vs. number of body parts when a person is correctly detected (using the same threshold).

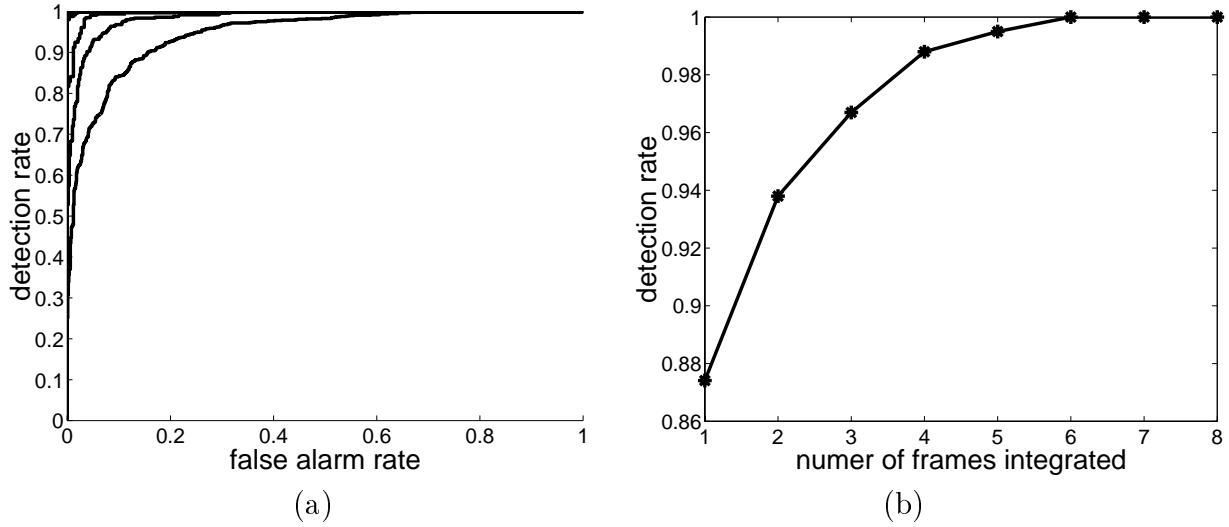


Figure 11: **Results of integrating multiple frames** (a) ROCs of integrating one to eight frames using only 5 body parts with 30 clutter points present. The more frames integrated, the better the ROC curve is. When more than five frames are used, the ROCs are almost perfect and overlapped with the axes. (b) detection rate (when $P_{detect} = 1 - P_{false-alarm}$) vs. number of frames used.

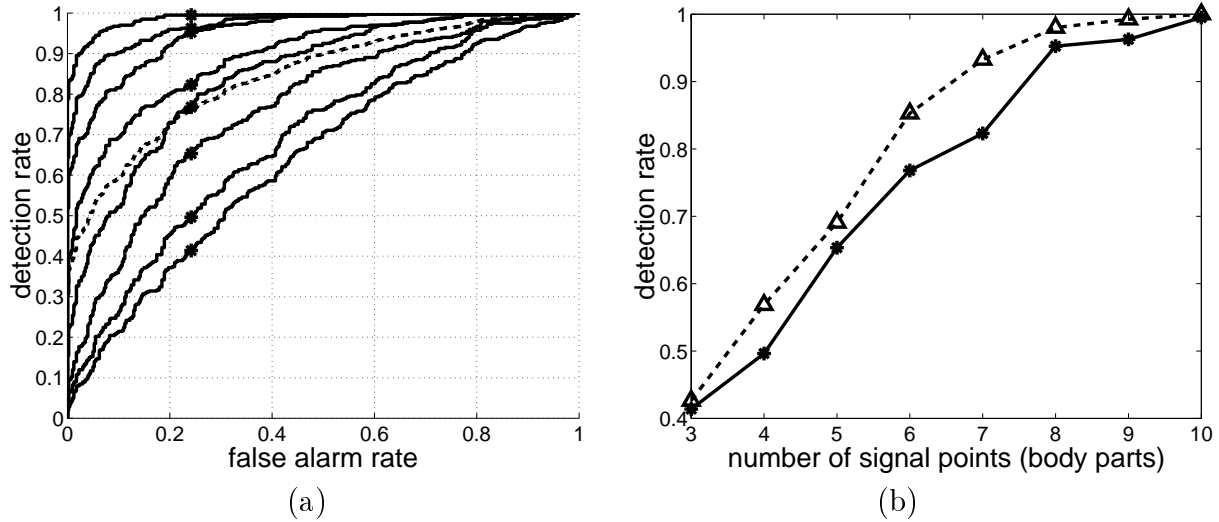


Figure 12: **Results of biological clutter** (a) Eight solid lines are ROCs for 3 to 10 body parts with 30 background points respectively. The bigger the number of signal points is, the better the ROC is; dashed line: overall ROC considering all the frames used in the eight solid ROCs. The threshold corresponding to $P_D = 1 - P_{FA}$ on this curve was used for (b). The stars (*) on the solid curves are the points corresponding to that threshold. (b) detection rate vs. number of signal points. Solid line (with stars): results of 30 added background points with false alarm rate 24.19%; Dashed line (with triangles): results of 20 added background points with false alarm rate 19.45%.

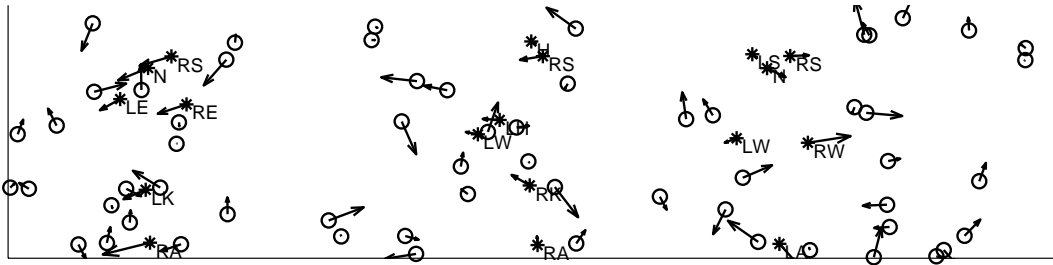


Figure 13: **One sample image of counting experiments.** '*'s denote body parts from a person and 'o's are background points. There are three persons (six body parts for each person) with sixty superimposed background points. Arrows are the velocities.

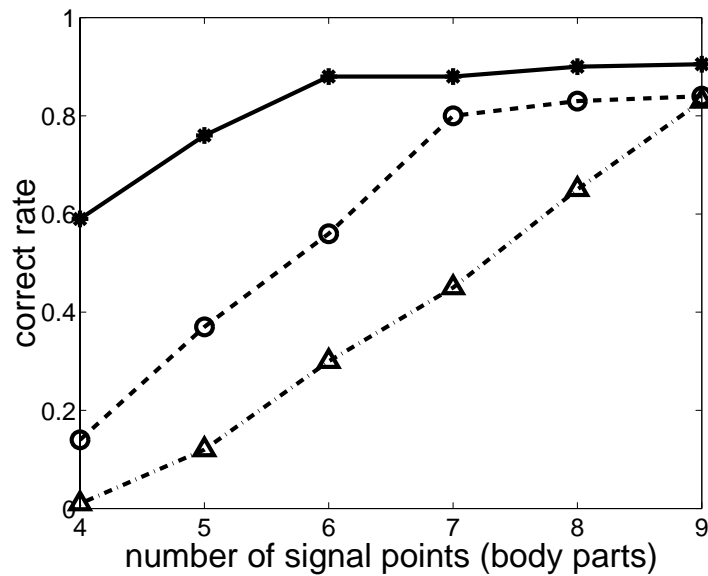


Figure 14: **Results of counting people.** Solid line (with *): one person; dashed line (with o): two persons; dashdot line (with triangles): three persons. Counting is done with regard to the threshold chosen from Figure 10 . For that threshold the correct rate for recognizing that there is no person in the scene is 95%.

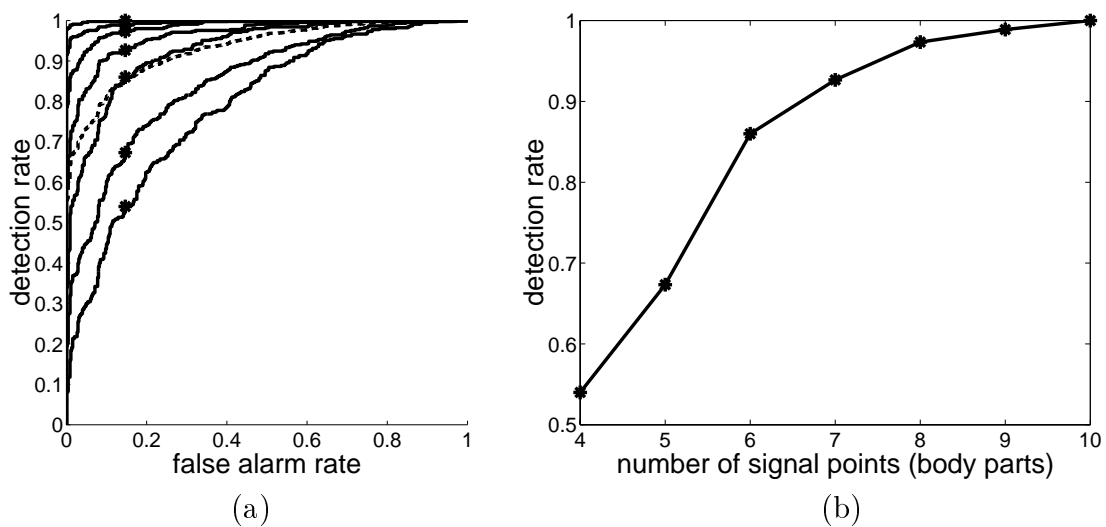


Figure 15: **Results of dancing sequences.** (a) Solid lines: ROC curves for 4 to 10 body parts with 30 added background points vs. 30 background points only. The bigger the number of signal points is, the better the ROC is. Dashed line: overall ROC considering all the frames used in seven solid ROCs. The threshold corresponding to $P_D = 1 - P_{FA}$ on this curve was used for (b). The stars (*) on the solid curves are the points corresponding to that threshold. (b) detection rate vs. the number of body parts displayed with regard to a fixed threshold at which $P_D = 1 - P_{FA}$ on the overall ROC curve in (a). The false alarm rate is 14.67%.