An Improved Scheme for Detection and Labeling in Johansson Displays

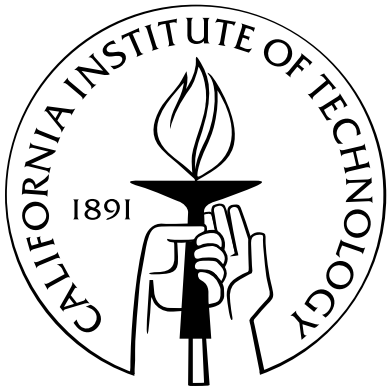
Thesis by

Claudio Fanti

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science



California Institute of Technology

Pasadena, California

2004

(Submitted May 28, 2004)

Acknowledgements

It is my great pleasure to thank my advisor Professor Pietro Perona for his help, support, and guidance throughout my research and studies. I would like to thank Marzia Polito with whom I had the pleasure to work on most of what I will present in this Thesis. A special thanks goes to my fellow members of the Computational Vision Lab and those at the Learning Group, for providing me with many valuable discussions and helpful suggestions for my work. Finally, I want to thank my family, for their endless love and support.

I am very grateful to Max Welling, who first proposed the idea of using Loopy Belief Propagation to solve for the optimal labeling in a 2001 Research Note, and who gave many useful suggestions. Sequences W1 and W2 used in the experiments were collected by Luis Goncalves and Enrico di Bernando. This work was partially funded by the NSF Center for Neuromorphic Systems Engineering grant EEC-9402726 and by the ONR MURI grant N00014-01-1-0890.

Abstract

Consider a number of moving points, where each point is attached to a joint of the human body and projected onto an image plane. Johannson showed that humans can effortlessly detect and recognize the presence of other humans from such displays. This is true even when some of the body points are missing (e.g. because of occlusion) and unrelated clutter points are added to the display. We are interested in replicating this ability in a machine. To this end, we present a labeling and detection scheme in a probabilistic framework. Our method is based on representing the joint probability density of positions and velocities of body points with a graphical model, and using Loopy Belief Propagation to calculate a likely interpretation of the scene. Furthermore, we introduce a global variable representing the body's centroid. Experiments on one motion-captured sequence suggest that our scheme improves on the accuracy of a previous approach based on triangulated graphical models, especially when very few parts are visible. The improvement is due both to the more general graph structure we use and, more significantly, to the introduction of the centroid variable.

Contents

List of Figures

Chapter 1

Introduction

This thesis presents an approach to the task of detecting the presence of a human being and labeling its body parts. This is a basic step in trying to solve the broader problem of automatically recognizing people's actions and behaviors.

Human motion analysis is a very important and hard problem in computer vision. When observing social interactions that take place in the surrounding environment, humans are, in general, the most important component. The interest is further justified by the number of application for which understanding people's actions and intentions is a central step. Among them, for instance, is monitoring people in airports or museums for security reasons. Detection of pedestrians is attractive to the automotive industry for safety and autonomous navigation systems. Even the daily interaction with computers and appliances could be greatly improved by a more user-friendly interface (in a sense, a more passive one, where it's the machine to autonomously infer what we expect it to do).

Motion provides a large amount of information about humans and is very useful for human social interactions. The goal of human motion analysis system is to extract information about human motion from video sequences, in an attempt to produce a concise description of the scene. The first question to be addressed should be whether or not there are humans in the scene. In case of positive answer, we would be interested in knowing both their location and which action they are doing.

Our visual system naturally perceives and analyzes human motion. Replicating this ability in machines is one of the most challenging and ambitious goals of ma-

Figure 1.1: People and Actions : we show an image of people running and superimposed some features with an arrow depicting their velocity. The set of positions and velocities constitutes the basic input for our algorithm.

chine vision. Johannson's experiments [4] show that humans find the instantaneous information on the position and velocity of a few features (such as the joints of the body, for instance) a sufficient cue to detect human presence and understand the gist of their activity. This still holds true even when clutter features are present in the scene, or some body parts features are occluded (see figure 1.1 for an example of features).

A side task that we will not discuss here, is that of identifying features in video sequences and computing their velocity across frames by tracking their position in time. It is a fairly well-understood problem for which reasonably good solutions are available from the literature (see [5], for instance).

In our work we will therefore assume that a number of features that are associated to the body have been detected and their velocity has been computed. We will allow some of such features to be missing and admit that some are not at all associated to

the body but rather originated from the background. Given such input (which we call *Generalized Johannson display*) we investigate ways of detecting the presence of a human in the scene and the labeling of the point features as body parts or clutter.

1.1 Graphical Models

The approach presented here is a generalization of [3] where the pattern of point positions and velocities associated to human motion was modelled with a triangulated graphical model for which a simple inference schema in the form of a dynamic programming algorithm was developed.

Given the probabilistic nature of our work, graphical models are a natural way for describing interaction among random quantities and an excellent conceptual representation to guide the implementation of inference schemas. In their most general form, graphical models can be thought of as a machine that can answer queries regarding the values of a set of random variables. The beauty stands in the fact that this machinery is built combining information locally, and propagating it in agreement with the theory of probabilities in order to reach global consistency.

This is the very basic idea that justifies our point of view in relation to the problem of detection and labeling. We describe a human being (the "global entity") as a collection of parts (the "local entities") and their mutual relationships. Starting out in a bottom-up fashion, we detect features and produce a local description of the information they convey about the human presence. The propagation of such information across local entities, while enforcing global consistency, is what ultimately provides the most likely interpretation of the scene.

1.2 Overview

The contribution of our work is twofold. We explore the benefits of representing more general and complex dependencies among the local entities in the model, at the price of an increased computational complexity. This is achieved by removing

the decomposability constraint and hence allowing a higher degree of connectivity in the graphical model. Furthermore, we develop an EM-like approximate Belief Propagation schema that allows for the simultaneous determination (in a maximum likelihood sense) of both local and global quantities. The novelty stands in the co-presence of (discrete) local variables as well as (continuous) global ones. While the discrete labels carry information that is primarily of local interest, global variables such as translation of the body, or its scale and orientation, are quantities that are common to the person as a whole. We recall how in [3] translation invariance is obtained at the level of individual cliques by computing relative positions locally. A major drop in performance can be observed with high levels of occlusion, due to the impossibility of computing such relative quantities. To circumvent this problem, we represent the location of the body as a single hidden global variable that is never observed but always present.

In Chapter 2 we will introduce the Labeling problem in generalized Johannson displays. In Chapter 3 and 3.1 we derive the probabilistic model and the learning of parameters and structure from data. The Expectation Maximization (EM) and Loopy Belief Propagation (LBP) based inference algorithm is detailed in Chapter 4, while in Chapter 5 and 6 we present the experimental results and conclusions.

Chapter 2

The Labeling Problem in the generalized Johannson displays

We start by introducing a few notational conventions that we will use throughout this thesis. We then formally introduce the detection and labeling problem for the generalized Johannson display.

2.1 Notations

In what follows, we will use bold-face letters $\mathbf{x}$ for random vectors, and italic letters x for their sample values.

A probability density function (or probability mass function, in the discrete case), defined over the variables $\mathbf{x}$, is denoted by $f_{\mathbf{x}}(x)$ where the subscript remind us of which variables are in the domain of the function, while the argument is a place-holder for the actual sample value of those variables.

When $q(\mathbf{x})$ is a deterministic function of a random quantity we write its expectation as $E_{f_{\mathbf{x}}}[q(x)]$ where the subscript denotes the distribution with respect to which we are taking the expectation.

An ordered set $\mathcal{I} = [i_1 \dots i_K]$ used as a vector's subscript has the intuitive meaning of the sub-vector whose entries are the elements that have indices in the set $\mathcal{I}$ in that order, i.e. $\mathbf{y}_{\mathcal{I}} = [\mathbf{y}_{i_1} \dots \mathbf{y}_{i_K}]$. When the ordered set is enclosed in squared brackets with a subscript $[\mathcal{I}]_s$ and it is applied to a dimension of a matrix $V = [v_{ij}]$, it selects the s-dimensional members of the matrix along that dimension, i.e. $V_{[1,2]_4[1,2]_4}$ is the

8×8 matrix obtained by selecting the first two 4-dimensional rows and columns.

According to the graphical models conventions, we will represent a probability density function $f_{\mathbf{x}}(x)$ as a Bayesian Network $\mathcal{B} = (G, \Psi)$, i.e. a directed acyclic graph $G = (V, E)$ with nodes $V = [\mathbf{x}]$ and oriented edges E, and a set of positive potentials (or densities, when normalized) $\Psi_i(x_{[C_i]})$ each defined on some subset $\mathbf{x}_{C_i}$ of the variables.

Each node $i \in V$ represents a random variable $\mathbf{x}_i$ in the domain of f. $\mathbf{x}_i$ and its parents $\mathbf{x}_{[\pi_i]}$ in the graph are called a family. Potentials are in general (but not necessarily) defined as $\Psi_i(x_{[i,\pi_i]}) = f_{\mathbf{x}_i|\mathbf{x}_{\pi_i}}(x_i|x_{\pi_i})$. If that is the case, then each potential is associated to one family in G and vice versa. The original probability distribution f can then be written as

$$f_{\mathbf{x}}(x) = \frac{1}{Z}\prod_{i=1}^{N}\Psi_i(x_{C_i}) = \prod_{i=1}^{N} f_{\mathbf{x}_i|\mathbf{x}_{\pi_i}}(x_i|x_{\pi_i}) \tag{2.1}$$

where Z is a normalization constant (which in this case equals 1 due to the potentials being the normalized conditionals).

2.2 Problem Definition

Similarly to a Johannson's setup, we identify a set of M relevant body *parts*, which for instance can be thought of as being in correspondence with the main joints and/or limbs of the body, although in general they need not carry any physical meaning. We identify each part i with a continuous random variable $\mathbf{x}_i \in \mathbb{R}^4$, representing its position and velocity.

Given a display on which a set of points (referred to as a *detection* or *observation*) have been marked, our goal is to find the most probable assignment of parts to detections, allowing for some parts to go undetected and for some detections to be assigned to a "generic" part which we think of as the background.

Each detection is denoted by $y_i \in \mathbb{R}^4$ and (like the parts) is endowed with four values, i.e. $y_i = [y_{i,a}, y_{i,b}, y_{i,v_a}, y_{i,v_b}]^T$ corresponding to its horizontal and vertical

positions and velocities. We model each single observation as a 4×1 random vector $\mathbf{y}_i$. For each display we call $y = [y_1^T \dots y_N^T]^T$ the $4N \times 1$ vector of all observations. As mentioned above, in general $N \geq M$ and some or all of the M parts might not be present in a given display.

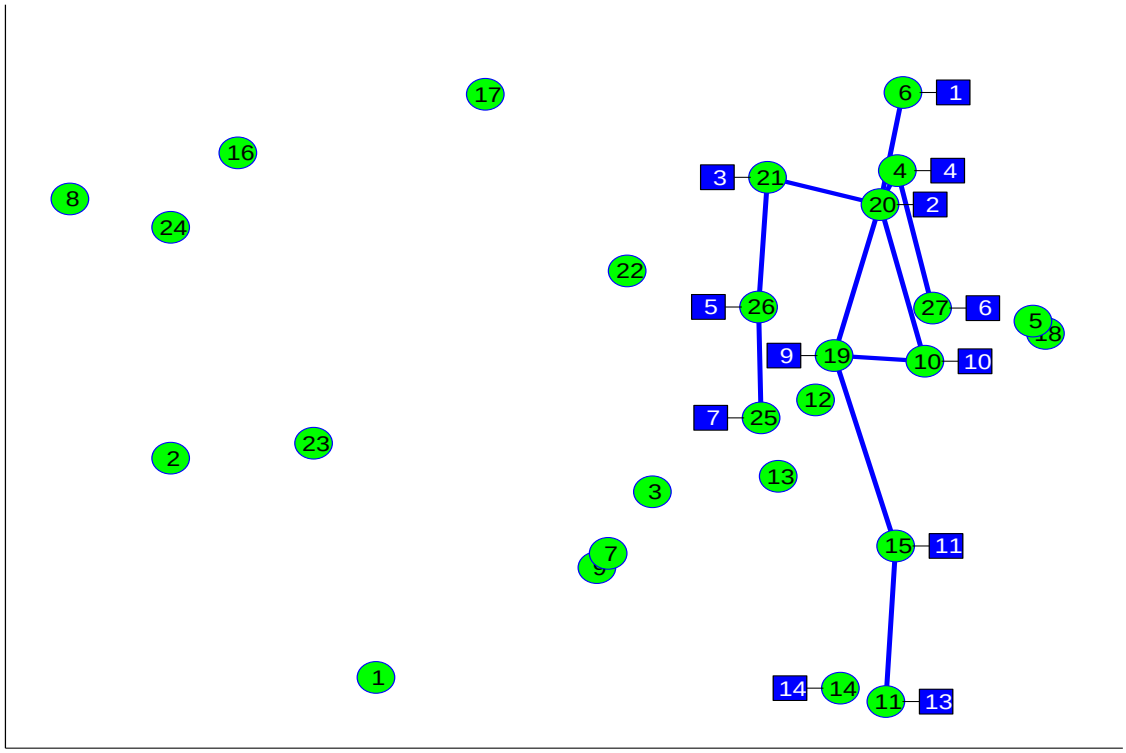


Figure 2.1: Sample display from sequence W2 : To illustrate the labeling process we show a generalized Johannson display with $N = 27$ detections (light-shaded dots). Superimposed (dark squares and sticks) is the labeling result that assigns 12 of the $M = 14$ parts found in the display to 12 detections and declares 2 parts to be missing. For instance, we can easily see that $\delta_1 = 1$ and $\lambda_7 = 25$ (i.e., body part number 1 has been detected and the number 7 has been assigned to detection number 25), while $\delta_{12} = 0$ (i.e., part number 12 was missing on this frame). We do not show which background part in the model is assigned to which detection in the display as it is of very little interest.

To account for missing parts we introduce a binary random variable $\boldsymbol{\delta}_i$, $i \in \{1 \dots M\}$ for each of the parts. $\boldsymbol{\delta}_i$ indicates whether or not the i^{th} part has been detected.

For $i \in \{1 \dots M\}$, a discrete random variable $\boldsymbol{\lambda}_i$ taking values on $\{1 \dots N\}$ is used to specify the correspondence of a body part i to a particular detection λ_i. Since this

makes sense only if the body part is detected, we assume by convention that $\lambda_i = 0$ if $\delta_i = 0$. A pair $\mathbf{h} = [\boldsymbol{\lambda}, \boldsymbol{\delta}]$ is called a labeling *hypothesis* (see Figure 2.1 for an example of labeled display).

In the rest of this thesis we will illustrate how a maximum likelihood estimate of $\mathbf{h}$ can be obtained, and we will show experimental results on detecting and labeling human body parts in presence of clutter and occlusions.

Chapter 3

Probabilistic Model

For a given display, let $y = [y_1^T \dots y_N^T]^T$ be the $4N \times 1$ vector of all observations. Any particular labeling hypothesis determines a partition of its set of indices into foreground ($\mathcal{F}$) and background ($\mathcal{B}$) detections, i.e. $[1 \dots N]^T = \mathcal{F} \cup \mathcal{B}$, where $\mathcal{F} = [\lambda_i : \delta_i = 1, i = 1 \dots M]^T$ and $\mathcal{B} = [1 \dots N]^T \setminus \mathcal{F}$. We say that $m = |\mathcal{F}|$ parts have been detected and $M - m$ are missing. Based on the partition induced on $\boldsymbol{\lambda}$ by $\boldsymbol{\delta}$, we can define two vectors $\boldsymbol{\lambda}^f = \boldsymbol{\lambda}_{\mathcal{F}}$ and $\boldsymbol{\lambda}^b = \boldsymbol{\lambda}_{\mathcal{B}}$, identifying the detections that were assigned to the foreground and those assigned to the background respectively. Finally, the set of detections $\mathbf{y}$ remains partitioned into the vectors $\mathbf{y}_{\boldsymbol{\lambda}^f}$ and $\mathbf{y}_{\boldsymbol{\lambda}^b}$ of the foreground and background detections respectively.

The foreground and background detections are assumed to be (conditionally) independent (given $\mathbf{h}$) meaning that their joint distribution factorizes as follows

$$f_{\mathbf{y}|\boldsymbol{\lambda}\boldsymbol{\delta}}(y|\lambda\delta) = f_{\mathbf{y}_{\boldsymbol{\lambda}^f}|\boldsymbol{\lambda}\boldsymbol{\delta}}(y_{\lambda^f}|\lambda\delta) \cdot f_{\mathbf{y}_{\boldsymbol{\lambda}^b}|\boldsymbol{\lambda}\boldsymbol{\delta}}(y_{\lambda^b}|\lambda\delta).$$

We model the foreground density $f_{\mathbf{y}_{\boldsymbol{\lambda}^f}|\boldsymbol{\lambda}\boldsymbol{\delta}}(y_{\lambda^f}|\lambda\delta)$ as a gaussian, while the background one $f_{\mathbf{y}_{\boldsymbol{\lambda}^b}|\boldsymbol{\lambda}\boldsymbol{\delta}}(y_{\lambda^b}|\lambda\delta)$ is assumed to be uniform $\mathcal{U}_{N-m}(A)$, with A determining the area of the position and velocity hyperplane for each of the $N - m$ background parts. The independence assumption is justified by the fact that influences of the body on the positions and velocities of the background-originated detections are (in principle) negligible. However, people do not appear against any background and in any possible location. We are more likely to find people in some settings than others

so that some consistency in the background could potentially be observed. Therefore, our assumption of independence, although reasonable, is not completely correct. Similarly, the assumption of uniform distribution for the background detections may be violated due to a particular background or implementation of the detector. Finally, modelling the collection of body parts as a gaussian is a further simplifying assumption. In what follows we will just ignore these aspects.

When all M parts are observed ($\boldsymbol{\delta} = [1 \dots 1]^T$) we have that $f_{\mathbf{y}_{\boldsymbol{\lambda}_{[1:M]_1}}|\boldsymbol{\lambda\delta}}(y_{\lambda_{[1:M]_1}}|\lambda\delta)$ is $\mathcal{N}(\mu, \Sigma)$. In general, if $m \leq M$ is the number of body parts that are present in a display, then $\mathcal{N}(\mu^f, \Sigma^f)$ is the marginalized (over the $M - m$ missing parts) version $\mathcal{N}(\mu^f, \Sigma^f)$ of the complete model $\mathcal{N}(\mu, \Sigma)$.

In what follows, our goal is to find a maximum likelihood hypothesis estimate $\hat{h} = [\hat{\lambda}, \hat{\delta}]$ such that

$$[\hat{\lambda}, \hat{\delta}] = \arg\max_{\lambda\delta} \{f_{\mathbf{y}_{\boldsymbol{\lambda}}|\boldsymbol{\lambda\delta}}(y_\lambda|\lambda\delta)\}. \tag{3.1}$$

3.1 Learning the Model's Parameters and Structure

In the following sections we will assume some familiarity with the connections between probability density functions and graphical models. Let us initially assume that the moving human being we want to detect is centrally positioned in the frame. We will then enhance the model in order to accommodate for horizontal and vertical translations.

Let us start by noticing how computing an optimal labeling hypothesis is in general equivalent to solving an very high dimensional combinatorial problem. In fact, if N is the number of detections and M is the number of parts in the model, we are after a maximum likelihood estimate of a mapping from M of the N detections into the M parts. Unfortunately there are $\binom{N}{M}$ such mappings, which prevents us from carrying out an exhaustive search. On the other hand, if each part could be assigned to a

detection independently of all the others, we would be facing a much simpler problem for which a brute force approach would suffice.

In general, there exists a trade off between the computational complexity of determining a solution, and the accuracy of the data description by the model. One of the purposes of this work is to automatically identify an optimal compromise.

3.2 Conditional Independencies and Computational Complexity

It turns out that given a finite number of realizations of a set of random variables, no independencies are in general observed even when the variables themselves are independent by construction. However, as the sample we are considering grows in size, it is reasonable to expect that a noticeable weaker level of dependency will be observed whenever the underlying variables are indeed independent.

In our work, we will assume that conditional independencies hold among positions and velocities of some set of body parts. As we already noted, this can greatly simplify (in fact render computationally feasible) the problem of determining the optimal labeling.

This idea has been explored in the literature. In [3] each of the body parts is assumed to be conditionally independent of all the others, given its parents-set of at most two other parts. As anticipated in the introduction, an additional constraint of decomposability is imposed on the structure of the graphical model that depicts such independencies. Furthermore, the connectivity among variables in the graph is manually determined according to the authors' experience and intuition.

As a first generalization, we will not require that the resulting graph be decomposable, so that all pairs of variables can be selected as parents-set for any part. Such a model provides in principle a more accurate description of the data, as well as allowing longer range connection among variables. Additionally (and more importantly), we implement a greedy schema that automatically determines what are the conditional

independencies observable in the training data. At the same computational cost, this has the benefit of increasing the accuracy in the description of the data by the model, in that it sets a bias in the learning process to ignore those dependencies that are weaker (therefore discarding "less information").

3.3 Bayesian Information Criterion for Structure Learning

In the learning process we want to estimate the parameters of $f_{\mathbf{y}_{\boldsymbol{\lambda}^f}|\boldsymbol{\lambda\delta}}(y_{\lambda^f}|\lambda\delta)$, where the labeling of the training set is known, $N = M$ and $\boldsymbol{\delta} = [1 \dots 1]^T$ (i.e., no clutter is present and all parts are visible). Additionally we would like to determine the best connectivity or structure among the variables (see Figure 3.1, for instance) so to minimize the computational complexity and simultaneously maximize the likelihood of the data in the model.

The problem of learning the optimal structure from data is known to be NP-Hard [11], however heuristics exists that do perform well in practice. In order to make the trade off between complexity and likelihood explicit, we adopt what is known as the *Bayesian Information Criterion* (BIC) score (see also [8]).

We recall that the BIC score is consistent, i.e., if the data we are using for learning were actually i.i.d. samples drawn from a probability distribution among the ones we admit, the corresponding graph would achieve the highest possible score. Moreover, since the probability distribution factorizes family-wise due to the conditional independencies among its variables, the score decomposes additively. Let G be a graph representing the conditional independencies in a gaussian model $\mathcal{N}(\mu, \Sigma)$, and let us denote by π_i the set of parents for node i. The BIC score is a function $J : G \longrightarrow \mathbb{R}$ defined as

$$J(G) = \sum_{i=1}^{M} J(i, \pi_i) \tag{3.2}$$

with

$$J(i, \pi_i) = \frac{N}{2} \log \frac{|\Sigma_{[i,\pi_i][i,\pi_i]}|}{|\Sigma_{ii}||\Sigma_{\pi_i\pi_i}|} + \frac{d_{\pi_i} d_i}{2} \log N. \tag{3.3}$$

and where d_i is the dimension of the i-th gaussian and $d_{\pi_i} = \sum_{j \in \pi_i} d_j$.

Being an exhaustive search among graphs infeasible, we attempt to determine the highest scoring graph by mean of a greedy hill-climbing algorithm, with random restarts. Starting from a random structure, the algorithm prunes the space of directed acyclic graphs by locally optimal changes of the structure. At each step an elementary operation among adding, removing or inverting an edge is performed, so that the resulting graph has the highest increase of score. Every time we update the graph, we enforce a maximum number of two parents for each vertex (fan-in). As mentioned in the introduction, this causes the maximum size of each family to be three and therefore puts a cubic upper bound on the order of complexity during inference. To prevent getting stuck in local maxima, we randomly restart a number of times once we cannot get any score improvements, and then we pick the graph achieving the highest score overall. As a final step, we obtain the full model by retaining the best structure found, together with the associated maximum likelihood parameters.

Removing the decomposability constraint on the graph structure has a major consequence in that exact belief propagation methods that pass through the construction of a junction tree are not applicable. When the junction property is satisfied, the maximum-weight spanning tree algorithm allows an efficient (linear time) construction of the junction tree as the one with the most populated separators between cliques. Here, we propose instead a construction of the junction graph that (greedily) attempts to minimize the complexity of the induced subgraph associated with each variable. The aim is to minimize the number of cycles in the final junction graph, hoping that it improves the chances of the Belief Propagation algorithm to converge to the globally optimal labeling hypothesis.

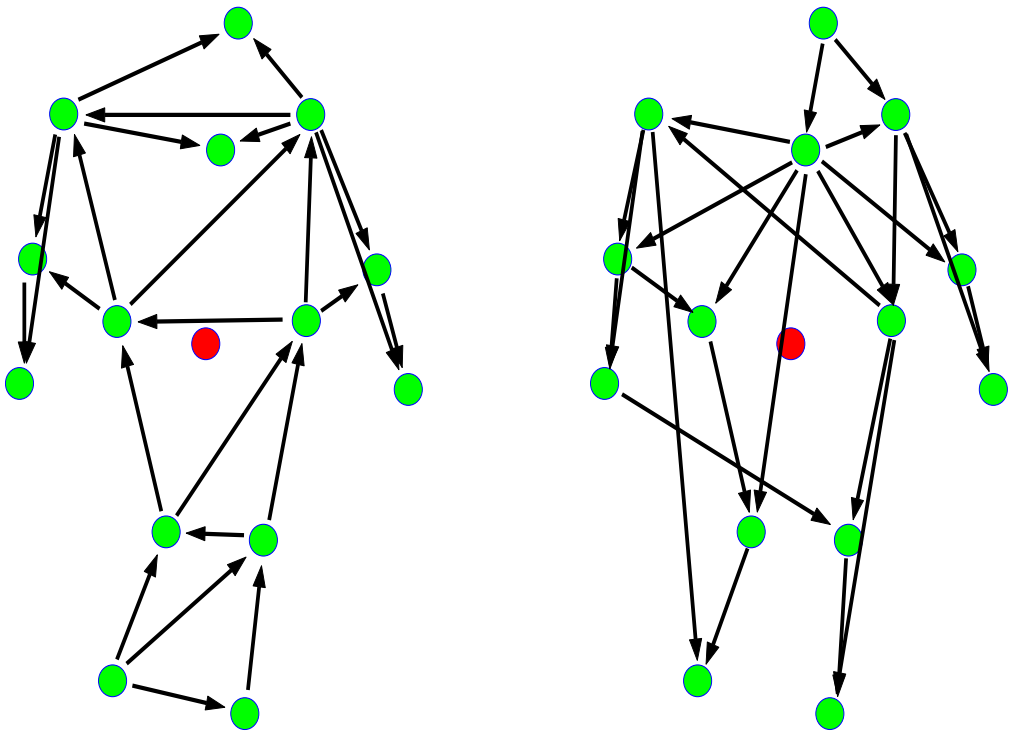


Figure 3.1: Graphical Models. Light shaded vertices represent variables associated to different body parts, edges indicate conditional (in)dependencies, following the standard Graphical Models conventions. [Left] Hand made decomposable graph from [3], used for comparison. [Right] Model learned from data (sequence W1, see chapter 5), with max fan-in constraint of 2.

Chapter 4

Detection and Labeling with Expectation Maximization (EM) and Loopy Belief Propagation (LBP)

As we mentioned in the introduction, the coexistence of continuous and discrete variables in the same graphical model has been proven to make the problem of inference intractable. In this chapter we will see how we can combine Belief Propagation with an Expectation-Maximization type of algorithm to simultaneously solve for the maximum likelihood labeling and location of the person.

4.1 Expectation Maximization

In order to solve the maximization problem (3.1) one could simply take advantage of the conditional independencies among variables and apply standard Belief Propagation techniques to compute the maximizing hypothesis. In most practical situation, however, we require our system to be invariant with respect to translations in position, that is, we would like to detect the presence of a human being regardless of the actual location in which it appears. Our approach in dealing with this issue is to use relative positions for the first two coordinates (position) of the observations. The following sections will cover the details of the revised probabilistic model and estimation procedure.

4.1.1 Translation Invariant Probabilistic Model

So far the reference system's origin for the parts' position was implicitly zero. We introduce a new parameter $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_b]^T$ to represent such reference system's origin, which we now allow to be different than zero. One can think of it as a continuous quantity which we model as a non-informative Gaussian. We define a new set of *centered observations*

$$\bar{\mathbf{y}}_i = \mathbf{y}_i - J_4^T \boldsymbol{\gamma} \tag{4.1}$$

with

$$J_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{4.2}$$

so that our model becomes

$$f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}|\gamma h) = f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}^f}|\boldsymbol{\gamma}\boldsymbol{\lambda}\boldsymbol{\delta}}(\bar{y}_{\lambda^f}|\gamma\lambda\delta) \cdot f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}^b}|\boldsymbol{\lambda}\boldsymbol{\delta}}(\bar{y}_{\lambda^b}|\lambda\delta).$$

where, in the second member, the first factor is now $\mathcal{N}(\bar{\mu}^f, \bar{\Sigma}^f)$ while the second factor becomes $\mathcal{U}_{N-m}(\bar{A})$. We will now turn to the description of an EM-based estimation procedure for $\boldsymbol{\gamma}$, while obtaining (as a by-product) the maximizing hypothesis h we are after.

4.1.2 E-Step

We start by assuming that the hypothesis h is unobservable and we therefore treat it as a hidden variable. In doing so, we have no access to the so called complete-data log-likelihood, which we instead replace with its expected value

$$\hat{L}_c(\tilde{f}, \gamma) = E_{\tilde{f}_{\mathbf{h}|\bar{\mathbf{y}}}}[\log f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}\mathbf{h}|\boldsymbol{\gamma}}(\bar{y}_\lambda h|\gamma)], \tag{4.3}$$

where the expectation is taken with respect to a generic distribution $\tilde{f}_{\mathbf{h}|\bar{\mathbf{y}}}(h|\bar{y})$. It is known that the E-step maximizing solution is

$$\begin{aligned} \tilde{f}^{(k)}_{\mathbf{h}|\bar{\mathbf{y}}}(h|\bar{y}) &= f_{\mathbf{h}|\bar{\mathbf{y}}\boldsymbol{\gamma}}(h|\bar{y}\gamma^{(k-1)}) \\ &\propto f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}_\lambda|\gamma^{(k-1)}h). \end{aligned}$$

Since we will not be able to compute such distribution for all the labelings h, we will make a so-called *hard assignment*, i.e., we will approximate $f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}_\lambda|\gamma^{(k-1)}h)$ with $\mathbf{1}(h - h^{(k)})$, where

$$h^{(k)} = \arg\max_h \{f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}_\lambda|\gamma^{(k-1)}h)\}.$$

Given the current estimate $\gamma^{(k-1)}$ of $\boldsymbol{\gamma}$, the hypothesis $h^{(k)}$ can be determined by maximizing the (discrete) potential function $\Psi(h) = \log f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}^f}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}_{\lambda^f}|\gamma^{(k-1)}h) \cdot f_{\mathbf{y}_{\boldsymbol{\lambda}^b}|\mathbf{h}}(y_{\lambda^b}|h)$ with a Max-Sum Loopy Belief Propagation (LBP) on the associated junction graph. Thanks to the conditional independencies among the variables, $\Psi(h)$ decomposes into a number of factors. Each factor's potential function is initialized to the family's conditional probability mass function (pmf). For a root node, we multiply its marginal pmf into one of its children's potential.

If LBP converges and the determined $h^{(k)}$ maximizes the expected log-likelihood $\hat{L}_c(\tilde{f}^{(k)}, \gamma^{(k-1)})$, then we are guaranteed (otherwise there is just reasonable[1] hope) that EM will converge to the sought-after ML estimate of $\boldsymbol{\gamma}$.

4.1.3 M-Step

In the M-Step we maximize (4.3) with respect to $\boldsymbol{\gamma}$, holding $h = h^{(k)}$, i.e. we compute

$$\gamma^{(k+1)} = \arg\max_\gamma \{\log f_{\bar{\mathbf{y}}_{\boldsymbol{\lambda}}|\boldsymbol{\gamma}\mathbf{h}}(\bar{y}_{\boldsymbol{\lambda}^{(k)}}|\gamma h^{(k)})\} \tag{4.4}$$

[1]Experimentally it is observed that when LBP converges, the determined maximum is either global or, although local, the potential's value is very close to its global optimum. If the potential is increased (not necessarily maximized) by LBP, that suffices for EM to converge

The maximizing γ can be obtained from

$$0 = \nabla_{\boldsymbol{\gamma}}[(y_{\lambda^{f^{(k)}}} - \bar{\mu}^f - J\gamma^{(k+1)})^T(\bar{\Sigma}^f)^{-1}(y_{\lambda^{f^{(k)}}} - \bar{\mu}^f - J\gamma^{(k+1)})] \tag{4.5}$$

where

$$J = \underbrace{[\ J_4 \quad J_4 \quad \cdots \quad J_4\ ]}_{m}^{T}. \tag{4.6}$$

The solution involves the inversion of the matrix $\bar{\Sigma}^f$ as a whole which we have observed to be a numerically instable operation, given the minimal variance in the vertical component of the motion we have examined. We therefore approximate the solution with the following

$$\gamma^{(k+1)} = J_4 \sum_{\delta_i^{(k)}=1}^{M} \left[\alpha_i(y_{\lambda_i^{(k)}} - \bar{\mu}_i)\right] \tag{4.7}$$

where the α_i's are defined as,

$$\alpha_i = \frac{\det(\bar{\Sigma})}{\det(\bar{\Sigma}_{[i]_4[i]_4})}. \tag{4.8}$$

Although not optimal, the (4.8) attempts to produce a smooth estimate by giving a higher weight (in the average (4.7)) to those observations assigned to parts with low variance.

4.2 Detection Criteria

Let $\boldsymbol{\sigma}$ be a (discrete) indicator random variable for the event that the Johansson's display represents a scene with a human body. So far, in our discussion we have implicitly assumed that $\sigma = 1$. In the following section we will describe a method for determining whether a human body is actually present (*detection*). We start by defining the likelihood ratio

$$R(y) = \frac{f_{\boldsymbol{\sigma}|\mathbf{y}}(1|y)}{f_{\boldsymbol{\sigma}|\mathbf{y}}(0|y)}. \tag{4.9}$$

Whenever $R(y) > 1$ we claim that a human body is present. By Bayes rule, $R(y)$ can be rewritten as

$$R(y) = \frac{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|1)}{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|0)} \cdot \frac{f_{\boldsymbol{\sigma}}(1)}{f_{\boldsymbol{\sigma}}(0)} = \frac{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|1)}{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|0)} \cdot R_p \tag{4.10}$$

where

$$R_p = \frac{P[\sigma = 1]}{P[\sigma = 0]} \tag{4.11}$$

is the contribution to $R(y)$ due to the prior on $\boldsymbol{\sigma}$, i.e., it's an indication of how likely it is for a human to be present in a scene before we actually look at it.

In order to compute the $R(y)$ we marginalize over the labeling hypothesis $\mathbf{h}$, i.e.,

$$f_{\mathbf{y}|\boldsymbol{\sigma}}(y|\sigma) = \sum_{\lambda,\delta} \left[f_{\mathbf{y}|\boldsymbol{\lambda}\boldsymbol{\delta}\boldsymbol{\sigma}}(y|\lambda\delta\sigma) f_{\boldsymbol{\lambda}|\boldsymbol{\delta}\boldsymbol{\sigma}}(\lambda|\delta\sigma) f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\delta|\sigma) \right].$$

When $\sigma = 0$, the only admissible hypotheses must have $\delta = 0^T$ (no body parts are present) which translates into $f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\delta|\sigma) = P[\delta|\sigma = 0] = 1_k(\delta - 0^T)$. Also, $f_{\boldsymbol{\lambda}|\boldsymbol{\delta}\boldsymbol{\sigma}}(\lambda|\delta 0) = N^{-N}$ as no labeling is more likely than any other before we have seen the detections. All N detections are labeled by $\boldsymbol{\lambda}$ as background and their conditional density is $\mathcal{U}_N(\bar{A})$. Therefore, we have

$$f_{\mathbf{y}|\boldsymbol{\sigma}}(y|0) = \frac{1}{\bar{A}^N} \frac{1}{N^N}. \tag{4.12}$$

When $\sigma = 1$, we have

$$f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\delta|1) = P[\delta|\sigma = 1] = \prod_{i=1}^{M} q_i^{\delta_i} (1 - q_i)^{1-\delta_i} \tag{4.13}$$

as we assume that any body part of index i appears in a given display with some fixed probability q_i, independently of all other parts. Similarly to the case of $\delta = 0$,

we have $f_{\boldsymbol{\lambda}|\boldsymbol{\delta}\boldsymbol{\sigma}}(\lambda|\delta 1) = N^{-N}$ and therefore we can write

$$f_{\mathbf{y}|\boldsymbol{\sigma}}(y|1) = \sum_{\lambda,\delta} \left[f_{\mathbf{y}|\boldsymbol{\lambda}\boldsymbol{\delta}\boldsymbol{\sigma}}(y|\lambda\delta 1) \frac{1}{N^N} f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\delta|1) \right]. \tag{4.14}$$

We conclude that

$$R(y) = R_p \frac{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|1)}{f_{\mathbf{y}|\boldsymbol{\sigma}}(y|0)} = R_p \bar{A}^N \sum_{\lambda,\delta} \left[f_{\mathbf{y}|\boldsymbol{\lambda}\boldsymbol{\delta}\boldsymbol{\sigma}}(y|\lambda\delta 1) f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\delta|1) \right]$$

In most practical situation the marginalization above is intractable. Additionally, the use of Gaussian models often requires working with log-probabilities for numerical accuracy. To circumvent these issues, in our implementation of Loopy Belief Propagation, we have assumed that the ML labeling $\hat{h}$ is predominant over all other labeling, so that in the estimate of $\boldsymbol{\sigma}$ we can approximate marginalization with maximization Therefore, we can write

$$R(y) \approx R_p \frac{f_{\mathbf{y}|\boldsymbol{\lambda}\boldsymbol{\delta}\boldsymbol{\sigma}}(y|\hat{\lambda}\hat{\delta}1) f_{\boldsymbol{\delta}|\boldsymbol{\sigma}}(\hat{\delta}|1)}{\bar{A}^{-N}}$$

where $\hat{\lambda}, \hat{\delta}$ is the maximizing hypothesis when $\sigma = 1$.

Chapter 5

Experimental Results

In our experiment we use two sequences W1 and W2 of about 7,000 frames each, representing a human subject walking back and forth along a straight line. The sequences were acquired and labeled with a motion capture system that recorded the positions of a set of markers attached to the main joints of the body (see Figure 5.1 for a graphical illustration of the data sets used).

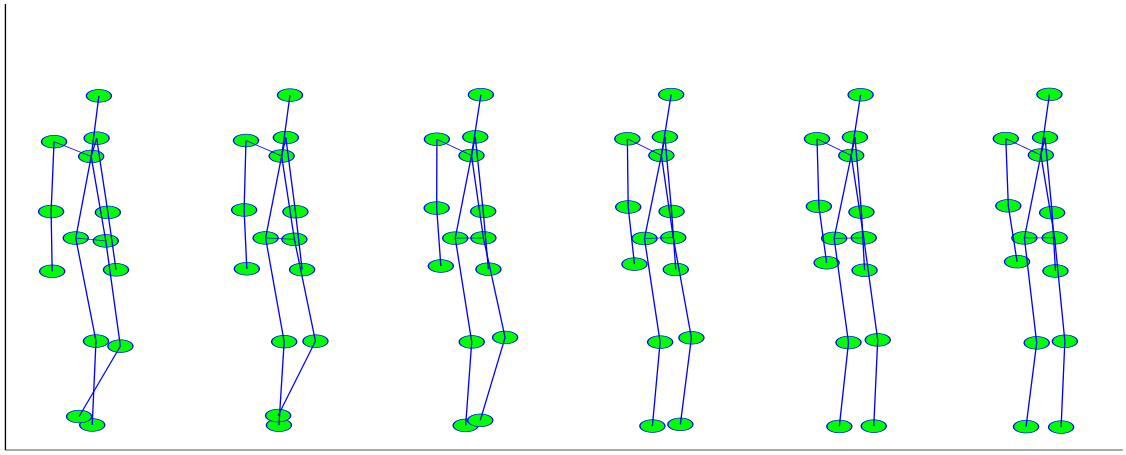


Figure 5.1: Sample frames from sequence W2: To illustrate the data set used in our experiments we took sequence W2 (a motion capture recording of a person walking) seen at 45° with respect to the direction of motion. Each dot represent the (x, y) positions of the 14 body parts appearing in frames 1001, 1006, 1011, 1016, 1021 and 1026. Although not present in the dataset, a few lines are drawn between some of the body parts for ease of interpretation of the display.

From each pair of consecutive frames a Johannson display is produced containing positions and velocities (difference in positions between the two frames) of the

parts. W1 is used to learn the probabilistic model's parameter and structure. A 700 frames random sample from W2 is then used to test our algorithm. Occlusions are synthetically generated by removing randomly chosen parts from the display. Clutter detections are obtained by sampling. The position is sampled from a uniform distribution over the size of the display, while the velocity's modulus and angle are obtained by sampling uniform distributions over the range of velocities observed in the whole training sequence over all parts, and over the unit circle respectively.

We evaluate the performance of our technique and compare it with the hand-made, decomposable graphical model of [3]. There, translation invariance is achieved by using relative positions within each clique. We refer to it as to the *local* version of translation invariance (as opposed to the *global* version we propose).

We first explore the benefits of just relaxing the decomposability constraint, still implementing the translation invariance locally. The lower two dashed curves of Figures 5.2 and 5.3 already show a noticeable improvement, especially when fewer body parts are visible. However, the biggest increase in performance is brought by global translation invariance as it is evident from the upper two curves of the same figures.

As for the dynamical programming algorithm of [3], the Loopy Belief Propagation algorithm runs in $O(MN^3)$, however 4 or 5 more iterations are needed (on average) for it to converge. Furthermore, to avoid local maxima, we restart the algorithm at most 10 times using a randomly generated schedule to pass the messages. Finally, when global invariance is used, we re-initialize $\boldsymbol{\gamma}$ up to 10 times. Each time we randomly pick a value within a different region of the display. On average, about 5 restarts for $\boldsymbol{\gamma}$, 5 different scheduling and 3 iterations of EM suffice to achieve a labeling with a likelihood comparable with the one of the ground truth labeling.

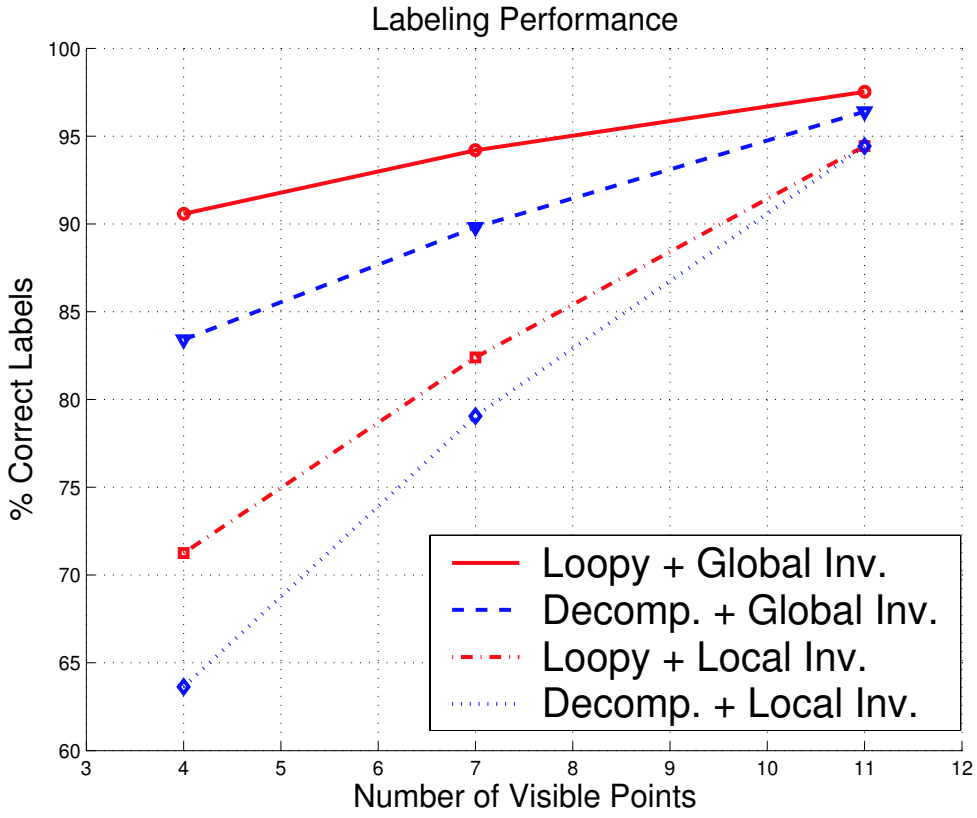


Figure 5.2: Labeling Performance. On each display from the sequence W2, we randomly occlude between 3 and 10 parts and superimpose 30 randomly positioned clutter points. For any given number of visible parts, the four curves represent the percentage of correctly labeled parts out of the total labels in all 700 displays of W2. Each curve reflects a combination of either Local or Global translation invariance and Decomposable or Loopy graph.

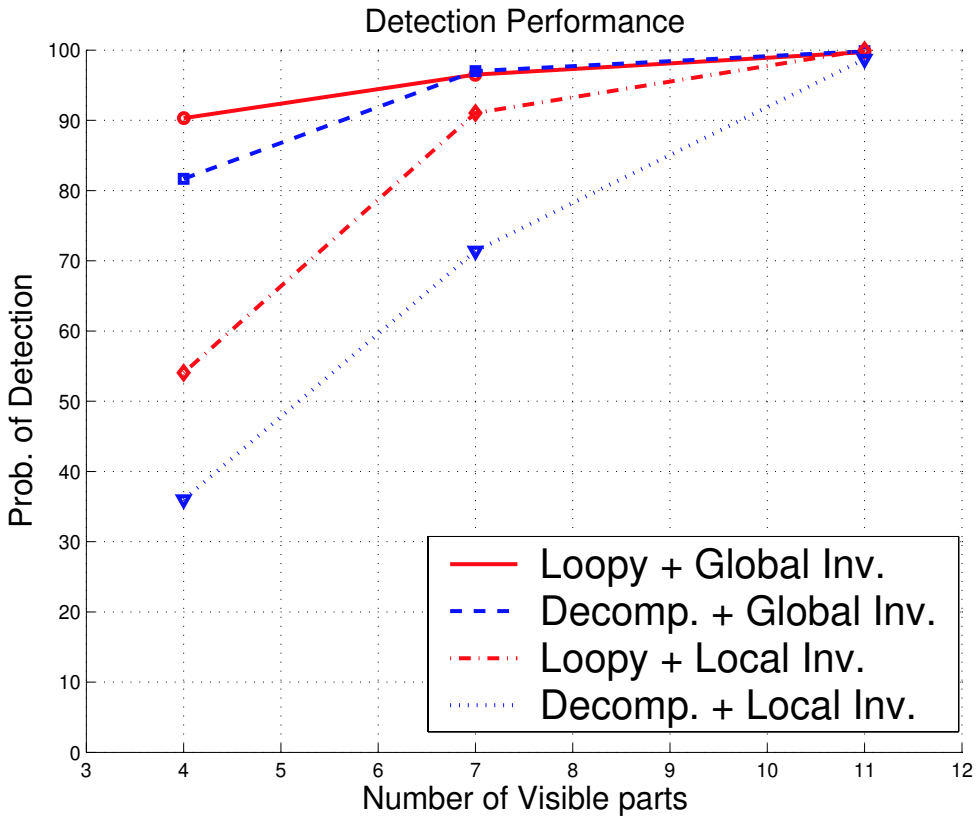


Figure 5.3: Detection Performance. On each display from the sequence W2, we randomly occlude between 3 and 10 parts and superimpose 30 randomly positioned clutter points. For any given number of visible parts, the four curves represent the probability of detecting a person when the display shows one, for a fixed $P_{false-alarm} = 10\%$. Here, the probability of false-alarm is that of stating that a person is present when only 30 points of clutters are presented. The number of visible points varies between 4, 7 and 11.

Chapter 6

Discussion, Conclusions and Future Work

The generalization of the decomposable graphical model to loopy here introduced, produced a gain in performance compared to [3]. Further improvement would be expected when allowing larger cliques in the junction graph, at a considerable computational cost. A more sensible improvement was obtained by adding a global variable modeling the centroid of the figure.

Taking [3] as a reference, there is about a 10x increase in computational cost when we either allow a loopy graph or account for translations with the centroid. When both enhancement are present the cost increase is between 100x and 1,000x.

We believe that the combination of these two techniques points in the right direction. The local translation invariance model required the computation of relative positions within the same clique. These could not be computed in the majority of cliques when a large number of body parts were occluded, even with the more accurate loopy graphical model. Moreover, the introduction of the centroid variable is also valuable in light of a possible extension of the algorithm to multi-frame tracking.

We should also note that the structure learning technique is sub-optimal due to the greediness of the algorithm. In addition, the model parameters and structure are estimated under the hypothesis of no occlusion or clutter. An algorithm that considers these two phenomena in the learning phase could likely achieve better results in realistic situations, when clutter and occlusion are significant.

Finally, the step towards using displays directly obtained from gray-level image sequences remains a challenge that will be the goal of future work.

Bibliography

[1] Y. Song, L. Goncalves and P. Perona, "Learning Probabilistic Structure for Human Motion Detection", *Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol II, pages 771-777, Kauai, Hawaii, December 2001.*

[2] Y. Song, L. Goncalves and P. Perona, "Unsupervised Learning of Human Motion Models", *Advances in Neural Information Processing Systems 14, Vancouver, Canada, December 2001.*

[3] Y. Song, L. Goncalves, and P. Perona, "Monocular perception of biological motion - clutter and partial occlusion", *Proc. of 6th European Conferences on Computer Vision, vol II, pages 719-733, Dublin, Ireland, June/July, 2000.*

[4] G. Johansson, "Visual Perception of Biological Motion and a Model For Its Analysis", *Perception and Psychophysics 14, 201-211, 1973.*

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

[6] S.M. Aji and R.J. McEliece, "The generalized distributive law", *IEEE Trans. Info. Theory, 46:325-343, March 2000.*

[7] P. Giudici and R Castelo, "Improving Markov Chain Monte Carlo Model Search for Data Mining", *Machine Learning 50(1-2), 127-158, 2003.*

[8] F. R. Bach, M. I. Jordan, "Learning graphical models with Mercer kernels", *Advances in Neural Information Processing Systems 15, Vancouver, Canada 2003.*

[9] W.T.Freeman and Y. Weiss, "On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs", *IEEE Transactions on Information Theory 47:2 pages 723-735. (2001).*

[10] J.S. Yedidia, W.T.Freeman and Y. Weiss, "Bethe free energy, Kikuchi approximations and belief propagation algorithms", *Advances in Neural Information Processing Systems 13, Vancouver, Canada, December 2000.*

[11] D. Chickering, "Learning bayesian networks is NP-Complete", *Learning from Data: Artificial Intelligence and Statistics 5, Springer-Verlag, 1996.*

[12] D. Chickering, "Optimal Structure Identification with Greedy Search", *Journal of Machine Learning Research 3, pages 507-554 (2002).*