

Learning to Estimate Human Pose with Data Driven Belief Propagation

Gang Hua[†]

Ming-Hsuan Yang[‡]

Ying Wu[†]

[†]ECE Department, Northwestern University
Evanston, IL 60208, U.S.A.

{ganghua, yingwu}@ece.northwestern.edu

[‡]Honda Research Institute
Mountain View, CA 94041, U.S.A.

myang@honda-ri.com

Abstract

We propose a statistical formulation for 2-D human pose estimation from single images. The human body configuration is modeled by a Markov network and the estimation problem is to infer pose parameters from image cues such as appearance, shape, edge, and color. From a set of hand labeled images, we accumulate prior knowledge of 2-D body shapes by learning their low-dimensional representations for inference of pose parameters. A data driven belief propagation Monte Carlo algorithm, utilizing importance sampling functions built from bottom-up visual cues, is proposed for efficient probabilistic inference. Contrasted to the few sequential statistical formulations in the literature, our algorithm integrates both top-down as well as bottom-up reasoning mechanisms, and can carry out the inference tasks in parallel. Experimental results demonstrate the potency and effectiveness of the proposed algorithm in estimating 2-D human pose from single images.

1. Introduction

Inferring human pose from single images is arguably one of the most difficult problems in computer vision and finds numerous applications from motion analysis to visual tracking. In this paper, we posit the 2-D human pose estimation problem within a probabilistic framework and develop an inference algorithm on a rigorous statistical footing. A human body pose is modeled by a Markov network where the nodes denote body parts and the edges encode constraints among them. An efficient data driven belief propagation Monte Carlo algorithm with importance sampling functions, built from low-level visual cues, is proposed to infer the 2-D human pose from a single image snapshot.

From a set of hand labeled images, we apply principal component analysis to learn the 2-D shape models of each body part which serve as prior knowledge in predicting potential candidates. Each body part is represented by a state variable describing its shape and location parameters. The data driven importance sampling for the head pose is built using a computationally efficient AdaBoost-based face detector [19]. Constrained by the head location from face de-

tection, a probabilistic hough transform [8] is adopted to extract salient line segments in the image and they are assembled to form good candidates for constructing an importance sampling function for the human torso. A skin color model pertaining to the specific subject in the image is built based on the face detection result, and is utilized in sampling functions to predict potential body part candidates such as arms and legs.

The data driven importance functions for body parts are incorporated in the belief propagation Monte Carlo framework for efficient Bayesian inference of the human pose. For human pose estimation, the observation models are built based on the steered edge response of the predicted body parts. Diametric to the sequential data driven Markov chain Monte Carlo algorithm, the proposed algorithm integrates both top-down as well as bottom-up reasoning mechanism with visual cues, and carries out the inference tasks in parallel within a sound statistical framework. For concreteness, we apply the developed method to estimate pose of soccer players in single images with cluttered backgrounds. Experimental results demonstrate the potency and effectiveness of the proposed method in estimating human pose from single images. We conclude with discussions on limitations of the current work and future plan to tackle these problems.

2. Prior work and context

While there exist numerous works on human body tracking [11], only a few of them address the initialization problem, i.e., estimating the human pose from single or multiple views. We observe the emergence of research work on this topic with impressive results [2, 9, 10] in the last few years. These algorithms are categorized into deterministic and statistical methods for ease of presentation.

Deterministic methods either approach this problem by applying deterministic optimization methods where the objective function is the matching error between the model and the image data [2, 1] or between the image data and the exemplar set [13]. An alternative is to build detectors for different body parts and rank the assembled configuration based on a set of human coded criteria [10]. Notwith-

standing the demonstrated success, there exist many challenging issues to be resolved for robust and efficient pose estimation. First, it entails solving an optimization problem of high dimensionality and thus the computation is inevitably intractable unless certain assumptions are explicitly made. Consequently, the application domains are limited to uncluttered backgrounds [1, 2] or the human body of fixed scale [10]. Second, the set of exemplars must be large enough to cover the parameter space to achieve satisfactory estimation results at the expense of growing computational complexity [13]. Third, it is difficult to build robust body part detectors except faces [19] due to the large appearance variation caused by clothing [10].

One salient merit of statistical formulation for posture estimation is that prior knowledge of human body parts (e.g., appearance, shape, edge and color) can all be exploited and integrated in a rigorous probabilistic framework for efficient inference. Ioffe and Forsyth [5] propose an algorithm to sequentially draw samples of body parts and make the best prediction by matching the assembled configurations with image observations. However, it is best applied to estimate poses of humans in images without clothing or cluttered background since their method relies solely on edge cues. Sigal et al. [15] resort to a non-parametric belief propagation algorithm [6] for inferring the 3-D human pose as the first step of their human tracking algorithm. Background subtraction and images from multiple views are employed to facilitate the human pose estimation and tracking problems. Lee and Cohen [9] apply the data driven Markov Chain Monte Carlo (DDMCMC) algorithm [18] to estimate 3-D human pose from single images, in which the MCMC algorithm is utilized to traverse the pose parameter space. Nevertheless it is not clear how the detailed balance condition and convergence in the Markov chain are ensured. Most importantly, the problem of inferring 3-D body pose from single 2-D images is intrinsically ill-posed as a result of depth ambiguity.

In this work, we propose a statistical formulation to infer 2-D body pose from single images. Different from the previous works, the proposed algorithm integrates the top-down and bottom-up inference with visual cues through a data driven belief propagation Monte Carlo algorithm for Bayesian reasoning. The algorithm is intrinsically parallel which is in direct contrast to the sequential sampling algorithm [5] and the sequential DDMCMC approach [9]. Furthermore we explicitly learn the shape models of body parts using quadrangles rather than rectangular templates [2, 5, 10], thereby facilitating inference of pose parameters.

3. Bayesian formulation

We posit the human pose estimation problem within a Bayesian framework and the task is to recover the hidden states, i.e., pose parameters, from image observations.

3.1 Markov network

A human body configuration is represented by a Markov network as shown in Figure 1. Each random variable \mathbf{x}_i represents the pose parameter (i.e., hidden state) of body part i , e.g., \mathbf{x}_h describes the pose of *head*, \mathbf{x}_t describes the pose of *torso*, and \mathbf{x}_{rul} describes the pose of the *right-upper-leg*. Each undirected link models the constraints between two adjacent body parts by a potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$. Each directed link depicts the image observation \mathbf{z}_i of body part i with an observation likelihood function $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$. Let \mathcal{S} be the set of all subscripts, we denote the set of pose parameters $\mathbf{X} = \{\mathbf{x}_i, i \in \mathcal{S}\}$ and the set of observations $\mathbf{Z} = \{\mathbf{z}_i, i \in \mathcal{S}\}$, respectively. The joint posterior distribution of this Markov network is

$$P(\mathbf{X}|\mathbf{Z}) \propto \prod_{(i,j) \in \mathcal{E}} \psi_{i,j}(\mathbf{x}_i, \mathbf{x}_j) \prod_{i \in \mathcal{V}} \phi_i(\mathbf{z}_i|\mathbf{x}_i), \quad (1)$$

where \mathcal{E} is the set of all undirected links and \mathcal{V} is the set of all directed links [7]. Consequently, the pose estimation problem is formulated as a Bayesian inference problem of estimating the marginal posterior distribution $P(\mathbf{x}_i|\mathbf{Z})$.

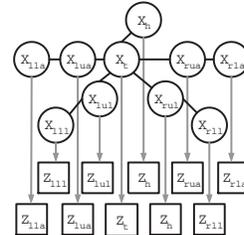


Figure 1. Markov network for human body pose.

A brute force approach to computing (1) is intractable since it involves numerous integrations of real valued random variables in every $P(\mathbf{x}_i|\mathbf{Z})$. The belief propagation algorithms, facilitated by local message passing (i.e., local computation), offer an efficient solution to such inference problems. Recently a Monte Carlo approach for belief propagation is proposed to deal with graphical models with non-Gaussian distributions [4]. In Section 4, we present a novel data driven belief propagation algorithm, which naturally integrate the bottom-up reasoning with the belief propagation Monte Carlo algorithm in a principled way.

3.2 Pose parametrization

We represent each body part by a quadrangular shape in a way similar to the existing works [2, 10]. However, we do not model them with rectangles or trapezoids since the body contours usually do not form parallel lines in images. From a set of 50 images, we manually labeled the quadrangular shapes and poses of human parts which best match human perception. A few examples of the labeled images are illustrated in Figure 2.

For each of the labeled quadrangular shape, we define the lines along the body outer contour as the *left* and the



Figure 2. Examples of labeled images.

right lines, and the other two lines as the *top* and the *bottom* lines, respectively. We define the local coordinate system of each body part by choosing the centroid of the quadrangular shape as its origin. The *Y* axis is pointed from the middle point of the *top* line to the middle point of the *bottom* line, and the *X* axis is perpendicular to the *Y* axis such that the local coordinate system is only subject to a rotation and a translation of the image coordinate system. Each labeled shape is then rotated with respect to a reference frame and then normalized in both *X* and *Y* directions, i.e., the length (width) along the *X* axis between the *left* and the *right* lines is normalized to 40 pixels, and the length (height) along the *Y* axis between the *top* and the *bottom* lines is normalized to 50 pixels, as depicted in Figure 3. Each normalized shape is then represented by a 8-dimensional vector by clockwise enumerating the coordinates of the four vertices.

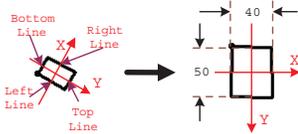


Figure 3. Normalization of the labeled shape.

We apply probabilistic principle component analysis (PCA) [17] to each set of the 8-dimensional normalized body part shapes for dimensionality reduction. In Section 4.2, we show how we use the learned shape model with probabilistic PCA to construct good importance sampling functions for body parts. In our experiments, 99% of the shape variation can be retained with the top 3 principal components. We denote the shape representation with reduced dimensionality for each body parts $i \in \mathcal{S}$ as \mathbf{ps}_i . Consequently, the 2-D pose of body part i can be represented by the rotation θ , scaling s_x, s_y , and translation t_x, t_y , in both *X* and *Y* directions of \mathbf{ps}_i , i.e.,

$$\mathbf{x}_i = \{\mathbf{ps}_i, s_x, s_y, \theta, t_x, t_y\}. \quad (2)$$

where we call \mathbf{ps}_i the intrinsic pose parameter and the rest the extrinsic pose parameters. By learning a low-dimensional shape representation, we reduce the originally 13-dimensional state space to 8 dimensions which in turns facilitates efficient sampling process. Figure 4 shows some of the original labeled shapes, the normalized shapes, as

well as the reconstructed shapes from the probabilistic PCA for the *right-upper-arm*. It is clear that the reconstructed shapes match well with the original labeled shapes.

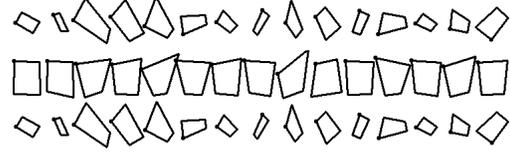


Figure 4. The original shapes (first row), the normalized (second row), and the reconstructed (third row) shapes of the *right-upper-arm* using probabilistic PCA.

3.3 Potential function and likelihood model

As mentioned earlier, a potential function ψ_{ij} models the pose constraints between two adjacent body parts. For pose estimation, the natural constraints entail any two adjacent body parts should be *loosely connected* [15], and we use a Gaussian distribution to model the Euclidean distance between the link points of two adjacent body parts, i.e.,

$$\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \propto \exp\left(-\frac{\|\tilde{\mathbf{P}}\mathbf{t}_{ij} - \tilde{\mathbf{P}}\mathbf{t}_{ji}\|^2}{2\sigma_{ij}^2}\right). \quad (3)$$

where $\|\cdot\|$ is the Euclidean distance, σ_{ij}^2 is the variance learned from the manually labeled images, and $\tilde{\mathbf{P}}\mathbf{t}_{ij}$ is the link point of the i^{th} to j^{th} body part while $\tilde{\mathbf{P}}\mathbf{t}_{ji}$ is the link point of the j^{th} to i^{th} body part. Figure 5 shows all the link points of the body parts. In our model, the link points are either corner points or middle points of either *bottom* or *top* line of the shape. For example, the link point of the *left-upper-arm* to the torso is defined as the corner point of the *left* line and the *bottom* line of the left-upper-arm shape, and the link point of the torso to the left-upper-arm is also specified by the corner point of the *left-bottom* corner of the torso shape. Whereas the link point of the *left-upper-arm* to the *left-lower-arm* is delineated by the middle point of the *top* line of the *left-upper-arm* shape, the link point of the *left-lower-arm* to the *left-upper-arm* is defined as the middle point of the *bottom* line of the *left-lower-arm* shape.

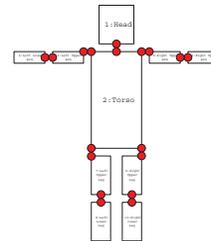


Figure 5. Each pair of red circle points represents the link point pair of two adjacent body parts. The link points are either corner points or middle points of bottom or top lines.

Although object appearance or texture has been successfully utilized in tasks such as face detection, the body contour information may be the only salient cue at our disposal

as clothing causes large visual variation. In this work, the likelihood function ϕ_i is constructed based on the average steered edge response [14] along the boundaries of the pose hypothesis of an body part. For example, let the rotation angle of one line segment l be α and the total number of points on the line is N_l , then the average steered edge response is

$$\bar{\mathcal{E}}_{l,\alpha} = \frac{1}{N_l \mathcal{E}_m} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in l} |\mathcal{E}_x(\mathbf{x}_i, \mathbf{y}_i) \sin \alpha - \mathcal{E}_y(\mathbf{x}_i, \mathbf{y}_i) \cos \alpha|, \quad (4)$$

where \mathcal{E}_m is the maximum value of the steered edge response. Unlike [14], we do not compute the steered edge response at different scales because the average steered edge responses across scales may make the steered edge response less discriminant. Instead, we compute the steered edge response in the RGB channels, i.e., $\mathcal{E}_\alpha^{(R)}(\mathbf{x}_i)$, $\mathcal{E}_\alpha^{(G)}(\mathbf{x}_i)$ and $\mathcal{E}_\alpha^{(B)}(\mathbf{x}_i)$ for each hypothesized body part \mathbf{x}_i . For *head* and *torso*, the average steered edge response is computed using all the four line segments of the shape pose hypothesis, whereas the average steered edge response is only calculated on the *left* and *right* line segments for the other body parts. Since all the steered edge responses have been normalized between 0 and 1, the likelihood function is defined based on the maximum steered edge response, i.e.,

$$\phi_i(\mathbf{z}_i | \mathbf{x}_i) = \max(\mathcal{E}_\alpha^{(R)}(\mathbf{x}_i), \mathcal{E}_\alpha^{(G)}(\mathbf{x}_i), \mathcal{E}_\alpha^{(B)}(\mathbf{x}_i)). \quad (5)$$

The reason for using the maximum steered edge response from different color channels is based on our empirical studies in which more discriminant likelihood functions can be obtained using the maximum rather than average edge response. We have experimented with the Gibbs likelihood model proposed in [12] but the performance is less satisfactory. One explanation is that background subtraction is utilized so that the body contours can be better extracted before learning a Gibbs model for likelihood estimation [12]. Nevertheless, background subtraction is inapplicable in this work as we aim to estimate human pose from single images.

4. Data driven belief propagation

With the Bayesian formulation described in Section 3, the pose estimation problem is to infer the marginal posterior distribution. In this Section, we propose a data driven belief propagation Monte Carlo algorithm (DDBPMC) for Bayesian inference with real valued graphical model.

4.1 Belief propagation Monte Carlo

Belief propagation is an efficient algorithm to computing posterior, $P(\mathbf{x}_i | Z)$, through a local message passing process where the message from \mathbf{x}_j to \mathbf{x}_i is computed by [7, 3]:

$$\mathbf{m}_{ij}(\mathbf{x}_i) \leftarrow \int_{\mathbf{x}_j} \phi_j(\mathbf{z}_j | \mathbf{x}_j) \psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \prod_{k \in \mathcal{N}(\mathbf{x}_j) \setminus i} \mathbf{m}_{jk}(\mathbf{x}_j), \quad (6)$$

where $\mathcal{N}(\mathbf{x}_j) \setminus i$ is the set of neighboring nodes of \mathbf{x}_j except \mathbf{x}_i . The belief propagation algorithm iteratively updates the messages passed among the connected nodes until it converges, and the marginal posterior distribution $P(\mathbf{x}_i | Z)$ on node \mathbf{x}_i can be efficiently computed by

$$P(\mathbf{x}_i | \mathbf{Z}) \propto \phi_i(\mathbf{z}_i | \mathbf{x}_i) \prod_{j \in \mathcal{N}(\mathbf{x}_i)} \mathbf{m}_{ij}(\mathbf{x}_i). \quad (7)$$

When both the potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and the observation likelihood $\phi_i(\mathbf{z}_i | \mathbf{x}_i)$ are Gaussian distributions, (6) can be evaluated analytically and thus (7) can be analytically computed. However, situations arise where the observation likelihood functions $\phi_i(\mathbf{z}_i | \mathbf{x}_i)$ can only be modeled with non-Gaussian distributions. In such cases, the messages $\mathbf{m}_{ij}(\mathbf{x}_i)$ are also non-Gaussians, thereby making the computation intractable.

To cope with this problem and allow greater flexibility, we resort to Monte Carlo approximation within the belief propagation formulation, and thereby a belief propagation Monte Carlo (BPMC) algorithm. We represent both the message $\mathbf{m}_{ij}(\mathbf{x}_i)$ and the marginal posterior distribution $P(\mathbf{x}_i | \mathbf{Z})$ as weighted sample sets by

$$\mathbf{m}_{ij}(\mathbf{x}_i) \sim \{\mathbf{s}_i^{(n)}, \omega_i^{(j,n)}\}_{n=1}^N, j \in \mathcal{N}(\mathbf{x}_i) \quad (8)$$

$$P(\mathbf{x}_i | \mathbf{Z}) \sim \{\mathbf{s}_i^{(n)}, \pi_i^{(n)}\}_{n=1}^N. \quad (9)$$

The iterative computation in the belief propagation can be implemented based on these weighted sample sets as summarized in Figure 6.

Note that in both the non-parametric belief propagation [16] and PAMPAS [6] algorithms, the messages as well as the marginal distributions are modeled with Gaussian mixtures, and the message passing process is carried out by a Markov chain Monte Carlo (MCMC) algorithm. In contrast, the BPMC algorithm models both the messages and marginal distributions with weighted samples, and the message passing process is computed efficiently based on the samples drawn from an importance sampling. It is worth emphasizing that good importance functions leads to efficient computation in the BPMC algorithm and better inference results. In Section 4.2, we show how we construct good importance functions with bottom-up visual cues for human pose estimation.

4.2 Data driven importance sampling

In this section, we describe the importance functions for drawing samples of body parts using visual cues. For concreteness, we present our algorithm with an application to estimate pose of soccer players in images. In such cases, we can exploit certain image cues for computational efficiency.

4.2.1 Importance function for head pose

With the demonstrated success in detecting faces efficiently, we utilize a variant of the AdaBoost-based face detector

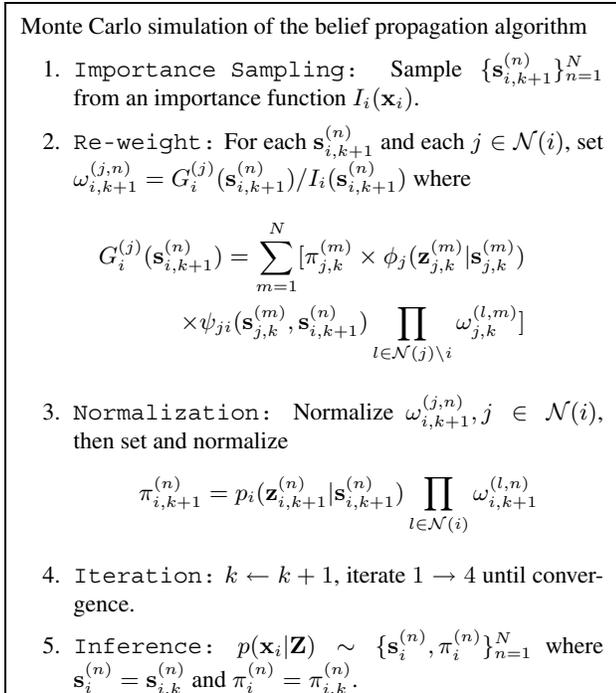


Figure 6. Belief Propagation Monte Carlo algorithm.

[19] to locate the face of a human in an image. However, this view-based detector performs best in detecting faces in upright frontal views although this problem can be alleviated by utilizing a multi-view extension. Figure 7(a) shows one face detected by the AdaBoost-based detector.

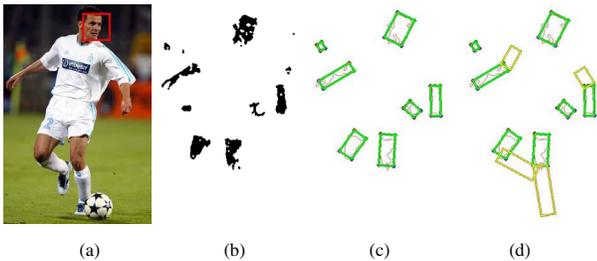


Figure 7. (a). Face detected by a AdaBoost-based face detector. (b). Image specific skin color segmentation. (c). Fitted lower-arm and upper-leg hypotheses. (d) Upper-arm and lower-leg hypotheses (yellow quadrangular shape).

One common problem with this view-based face detector is that the raw detection results are usually not very accurate (i.e., the returned rectangles do not precisely lock on faces in the correct pose and often enclose background pixels), and thus more efforts are required to better estimate head pose. Since skin color pixels account for the majority portion of a rectangular area enclosing a face, we use a k -means algorithm ($k = 2$) to group the pixels within the rectangle into skin/non-skin clusters. The center of the face rectangle is repositioned to centroid of the cluster of skin color pixels.

We then project the rectangular shape onto the learned PCA subspace of the head shape, thereby obtaining its intrinsic pose parameters as defined in (2). Along with the extrinsic rotation, scaling and translation parameters extracted from the face rectangle, we obtain an approximated head pose $\mathbf{I}\mathbf{x}_h$, and thereby an importance sampling function:

$$\mathbf{I}_h(\mathbf{x}_h) \sim \mathcal{N}(\mathbf{x}_h|\mathbf{I}\mathbf{x}_h, \Sigma_h) \quad (10)$$

where Σ_h is a diagonal covariance matrix.

4.2.2 Importance functions for arm and leg pose

For the human pose estimation problem considered in this paper, the soccer players often wear short sleeve shirts and short trunks, and consequently skin color is a salient cue for locating *lower-arm* and *upper-leg* regions.

A skin color model is constructed from the pixels of skin color cluster obtained from the k -means algorithm within the detected face region as discussed in Section 4.2.1. Specifically, a 2-D color histogram is computed from the normalized RGB pixel values of the skin color cluster. Although it is difficult and time consuming to develop a generic skin color model to account for all variations (as a result of lighting and race factors), it is relatively easy and effective to construct a skin color model specific to the human subject considered for pose estimation, and consequently skin color regions can be extracted effectively with thresholds. Figure 7(b) shows some segmentation results using the learned skin color histogram, and Figure 7(c) shows the results with best fit rectangles after discarding small blobs. Note that the number of skin tone blobs do not necessarily match the number of body parts.

Geometric cues such as shape, size, position, and orientation with respect to the head position of a human can be exploited to generate good pose hypotheses for the *lower-arm* and the *upper-leg* body parts from these fitted rectangles. The hypotheses for the *upper-arm* and the *lower-leg* are then generated by first rotating the shape with respect to the link point of the corresponding *lower-arm* and the *upper-leg* hypotheses respectively, and then evaluating the image likelihoods based on edge response using (4) and (5) for each rotation angle. The hypotheses with maximum likelihoods for *upper-arm* and *lower-leg* parts are selected for importance functions. Figure 7(d) shows one hypothesis for each of the upper-arm and lower-leg. The importance sampling function is modeled by a Gaussian mixture of these hypotheses. That is, after obtaining \mathcal{K} good pose hypothesis $\mathbf{I}\mathbf{x}_i^{(n)}$, $n = 1 \dots \mathcal{K}$ for body part i , we draw samples from the importance function

$$\mathbf{I}_i(\mathbf{x}_i) \sim \sum_{n=1}^{\mathcal{K}} \frac{1}{\mathcal{K}} \mathcal{N}(\mathbf{x}_i|\mathbf{I}\mathbf{x}_i^{(n)}, \Sigma_i), i \in \mathcal{S} \setminus \{h, t\}, \quad (11)$$

where Σ_i is a diagonal covariance matrix. Note that a small number of \mathcal{K} good hypotheses facilitate efficient sampling

and inference process although it may have adverse effects if the value is too small.

4.2.3 Importance function for torso pose

Locating the torso region may be the most important task in human pose estimation since it is connected to most of the other body parts. However, detecting a torso part is difficult as it is usually clothed and thereby has a large variation in appearance. Without salient image cues (e.g., color and texture) to facilitate the detection process, we utilize line segments extracted from the probabilistic Hough transform [8] to assemble good shape hypotheses of the torso part.

A Canny edge detector is first applied to build the edge map, and then a probabilistic Hough transform is performed to detect those near-horizontal and near-vertical line segments. For each combination of a pair of vertical line segments, l_{v1} , l_{v2} and a pair of horizontal line segments l_{h1} , l_{h2} , let their corner points of the assembled shape be $p_{v1,h1}$, $p_{v1,h2}$, $p_{v2,h1}$, and $p_{v2,h2}$ respectively. Torso hypotheses are obtained by solving an optimization problem with an objective function specified by

1. The normalized shape of a good hypothesis should be reconstructed by the learned PCA subspace of the torso with minimum error.
2. The distance between a good hypothesized torso part should be as close to the detected face as possible.
3. The two vertical lines, l_{v1} , l_{v2} should be as symmetric as possible in the assembled shape.

subject to the constraints that $p_{v1,h1}$, $p_{v1,h2}$, $p_{v2,h1}$, and $p_{v2,h2}$ are within the range of image.

For each of the \mathcal{M} torso hypotheses $\mathbf{I}\mathbf{x}_t^{(n)}$ obtained by solving the above-mentioned optimization problem ($n = 1, \dots, \mathcal{M}$ and usually $\mathcal{M} < 10$), we compute the response of edges extracted by the Canny detector with likelihood $\beta_t^{(n)}$ using functions similar to (4) and (5). The importance sampling function for the torso pose is specified by a Gaussian mixture, i.e.,

$$\mathbf{I}_t(\mathbf{x}_t) \sim \sum_{n=1}^{\mathcal{M}} \beta_t^{(n)} \mathcal{N}(\mathbf{x}_t | \mathbf{I}\mathbf{x}_t^{(n)}, \Sigma_t). \quad (12)$$

where Σ_t is the diagonal covariance matrix. Figure 8 shows one example of the detected near-horizontal and near-vertical line segments from the probabilistic Hough transform, and the corresponding torso hypotheses. Although the number of combinations using horizontal and vertical lines is large, solving the above-mentioned optimization problem significantly prunes the number of torso hypotheses (i.e., $\mathcal{M} < 10$), thereby facilitating efficient and effective inference.

5. Experiments

For concreteness, we apply our algorithm to estimate pose of soccer players in images. The proposed algorithm can be extended to estimate human pose in other domains.

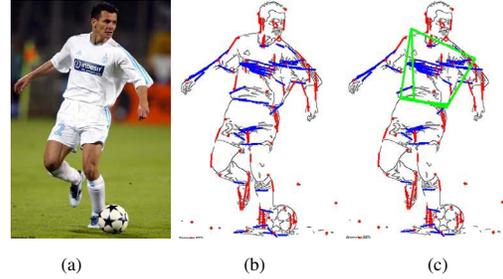


Figure 8. (a). Original image. (b). Line segments extracted by probabilistic Hough transform (red for near-vertical and blue for near-horizontal lines). (c). Torso hypotheses assembled from the line segments shown in (b).

5.1 Validation of the likelihood model

To demonstrate the effectiveness of the likelihood function proposed in Section 3.3, we generate a number of *left-lower-leg* hypotheses by translating the correctly labeled body part horizontally as shown in Figure 9(a), and their likelihoods are shown in Figure 9(b).

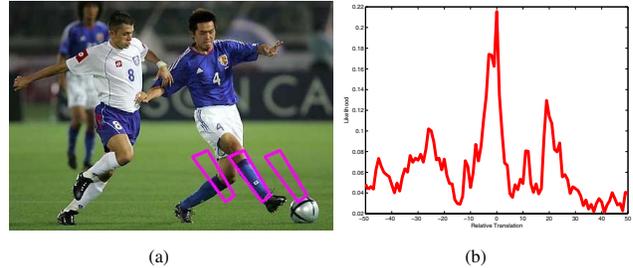


Figure 9. (a) Translation of the *left-lower-leg* part with respect to the correct location horizontally. (b) Likelihoods of the translated *left-lower-leg* hypotheses from the correct location.

As exemplified in Figure 9(b) the maximum likelihood occurs at the correct labeled location (i.e., 0 translation horizontally). The two small peaks correspond to the cases when one of the *left* and *right* lines of the shape pose is aligned with the boundary of the *left-lower-leg* in the image. The likelihood plots for the other body parts are similar to Figure 9(b) except the likelihood model for the torso may not peak at the correct labeled location and may have more local peaks (due to noisy edge response). This observation indicates that the difficulty of constructing a likelihood model of the torso part using only edge cues.

5.2 Pose estimation results

To learn the PCA subspace for each body part, we collected a set of 50 images and manually labeled the quadrangular shapes and poses of human body parts which best match human perception (Please see the accompanied video “720.wmv” for all the training images.). For pose estimation experiments, we gathered another set of 30 images and

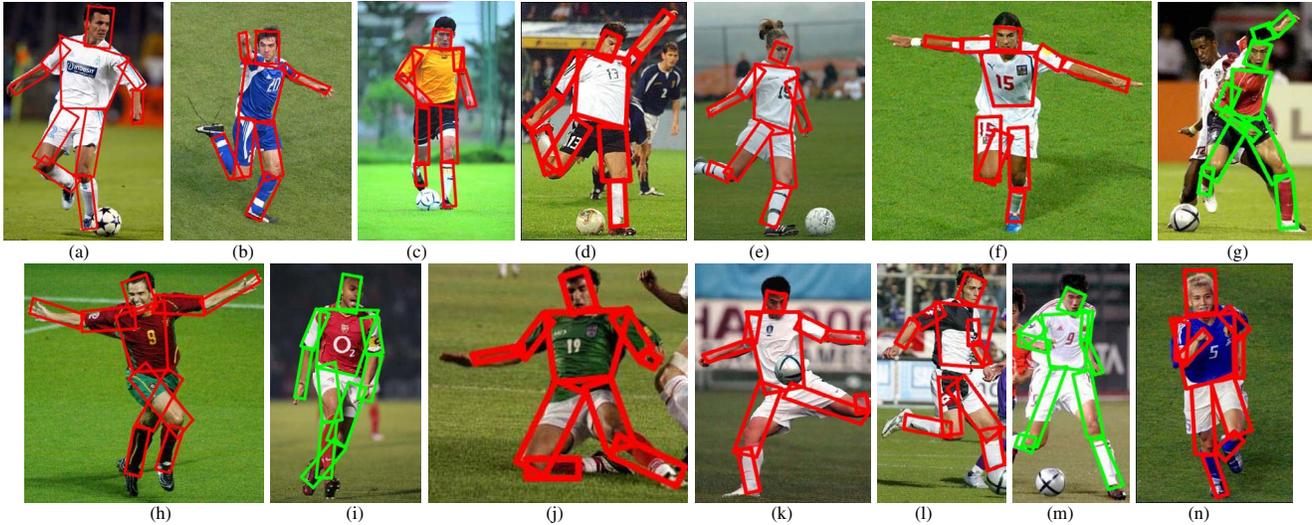


Figure 10. Experimental results of human pose estimation. More results can be found in the accompanied video “720.wmv”.

manually located the body parts as ground truth (We will make the test image set publicly available at appropriate time.). These images contain humans with large variation in pose and backgrounds, as well as occlusions either due to clothing or view angles. The values of the diagonal covariance matrices in importance functions (10)-(12) are empirical learned from the training image set.

Empirical results on estimating pose of soccer players in single images are illustrated in Figure 10 where the best estimated shapes and locations of body parts are enclosed with quadrangles (Please note that more test results can be found in the accompanied video “720.wmv”). The experimental results show that our algorithm is able to locate the body parts and estimate their pose well even though they appear in different posture, background, view angles and lighting conditions. Our algorithm is able to infer pose which are heavily occluded in Figure 10(f)-(g) as a result of data driven importance sampling from the visual cues. Specifically, the left lower leg of the player in Figure 10(f) is located as a result of the best pose estimation using image likelihoods and importance function (11). Similarly, the occluded body parts and their poses in Figure 10(h)-(j) are inferred using the proposed DDBPMC algorithm.

		Head	Torso	LUA	LLA	RUA	
RMSE		14.32	18.96	14.62	11.85	19.52	
		RLA	LUL	LLL	RUL	RLL	Overall
RMSE		19.01	23.75	18.19	20.48	18.98	17.96

Table 1. Average root mean square error (RMSE) of the estimated 2-D pose for each body part and for the whole body pose (e.g., LUA refers to left-upper-arm).

We evaluate the accuracy of our pose estimation algorithm by computing the root mean square errors (RMSE)

between the estimated pose enclosed by quadrangles and the ground truth, i.e., the RMSE between the four corner points of the two quadrangles. The average RMSE of each body part as well as that of the overall full body pose estimation over the 30 test images are presented in Table 1. At first glance, it seems that the RMSE of our algorithm is larger than the result of 20 test images reported in [9] even though the test sets are different. Nevertheless, we compute the accuracy of four points for each body parts while they just evaluated the accuracy of the joint locations, and thus the RMSE comparison is not justified. Further, the number of points set we compute is larger than that in [9]. Another complicating factor is the difficulty of determining what the “ground truth” of body pose is, as a result of covered clothing and difference of human perception in labeling body parts as well as pose precisely. Finally, the average RMSE of each image is presented in in Figure 11 to show the distribution of the overall RMSE among the 30 test images.

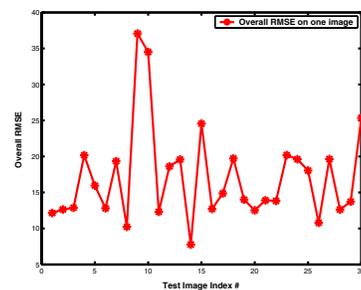


Figure 11. Overall RMSE of each of the test images.

The current implementation of the proposed algorithm draws 500 samples for each of the body parts, and the message passing process of the DDBPMC algorithm is iterated 6 times. Without code optimization, it takes about 2 to 3 minutes to process an image on a Pentium IV 1.7 GHz ma-

chine with 256 MB memory.

5.3 Discussions

Compared with the most relevant work [9], the problem we address in this paper is well posed rather than inferring 3-D pose from single 2-D images. Furthermore, the test images in our work are more challenging since they contain complex poses with occlusions in textured background. Finally, we have done a larger scale experiment and present all results in the accompanied video.

Although the experimental results demonstrate success of our algorithm in pose estimation from single images, there are a few research issues to be explored. The body postures such as torso may be more accurately estimated with more complicated body shapes. However, the inference problem will be more complicated due to the increasing degree of freedom in body shape. The proposed algorithm sometimes fails when long line segments are observed near the torso region. This is not surprising since long line segments often cause problems in generating good hypotheses of the torso region.

6. Concluding remarks

We propose a rigorous statistical formulation for 2-D human pose estimation from single images. The theoretic foundation of this work is based on a Markov network, and the estimation problem is to infer pose parameters from observable cues such as appearance, shape, edge, and color. A novel data driven belief propagation Monte Carlo (DDBPMC) algorithm, which combines both top-down and bottom-up reasoning within a rigorous statistical framework, is proposed for efficient Bayesian inference. This is in contrast to the data driven Markov chain Monte Carlo (DDMCMC) algorithm in that DDBPMC carries out the Bayesian inference in parallel while the DDMCMC algorithm performs sequentially. Experimental results demonstrate the potency and effectiveness of the proposed method in estimating human pose from single images.

The proposed algorithm can be easily extended to better estimate human pose in situations where contour or motion cues abound. Our future work will focus on integrating visual cues to build better data driven importance functions for a more efficient pose estimation algorithm.

Acknowledgment

The majority of this work was carried out while GH was an intern at HRI. It was also supported in part by NSF grants IIS-0347877, IIS-0308222, Northwestern faculty startup funds for YW and Murphy Fellowship for GH.

References

[1] C. Barrn and I. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284, 3 2001.

[2] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2066–2073, 2000.

[3] W. T. Freeman and E. C. Pasztor. Learning low-level vision. In *Proc. IEEE International Conference on Computer Vision*, pages 1182–1189, 1999.

[4] G. Hua and Y. Wu. Multi-scale visual tracking by sequential belief propagation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 826–833, 2004.

[5] S. Ioffe and D. Forsyth. Finding people by sampling. In *Proc. IEEE International Conference on Computer Vision*, pages 1092–1097, 1999.

[6] M. Isard. PAMPAS: Real-valued graphical models for computer vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 613–620, 2003.

[7] M. Jordan and Y. Weiss. Graphical models: Probabilistic inference. In *The Handbook of Brain Theory and Neural Network*, pages 243–266. MIT Press, second edition, 2002.

[8] N. Kiryati, Y. Eldar, and A. M. Bruckstein. A probabilistic Hough transform. *Pattern Recognition*, 24(4):303–316, 1991.

[9] M. W. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 334–341, 2004.

[10] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 326–333, 2004.

[11] T. Moselund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[12] S. Roth, L. Sigal, and M. Black. Gibbs likelihoods for Bayesian tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2004.

[13] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. IEEE International Conference on Computer Vision*, volume 2, pages 750–757, 2003.

[14] H. Sidenbladh and M. Black. Learning the statistics of people in image and video. *International Journal of Computer Vision*, 54(1-3):183–209, 2003.

[15] L. Sigal, M. Isard, B. Sigelman, and M. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Advances in Neural Information Processing System 16*. MIT Press, 2004.

[16] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 605–612, 2003.

[17] M. E. Tipping and C. M. Bishop. Probabilistic principle component analysis. *Journal of Royal Statistical Society, Series B*, 61(3):611–622, 1999.

[18] Z. Tu and S.-C. Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[19] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.