J Ambient Intell Human Comput (2010) 1:211–220
DOI 10.1007/s12652-010-0015-1

ORIGINAL RESEARCH

Human motion detection using Markov random fields

Xiao-Qin Cao · Zhi-Qiang Liu

Received: 11 February 2010 / Accepted: 29 April 2010 / Published online: 14 May 2010

Abstract In this paper, we propose Markov random fields (MRFs) to automatically detect a moving human body through minimizing the joint energy of the MRF for the velocity and relative position of body parts. The relaxation labeling algorithm is employed to find the best body part labeling configuration between MRFs and observed data. We detect a walking motion viewed monocularly based on point features, where some points are from the unoccluded body parts and some belong to the background. The results show that MRFs can detect human motions robustly and accurately.

Keywords Human motion · Markov random fields · Relaxation labeling

1 Introduction

Automatically capturing human motion is very important and useful in computer vision with many potential applications such as human–computer interface (HCI) and real-time video surveillance. In the past two decades, the problem of human motion detection has received much more attention. Although the significant progress has been made (Song et al. 2003; Yu and Medioni 2009; Isard and Blake 1998; Lee and Nevatia 2009; Wu and Yu 2006; Khan et al. 2005), it is still a challenging open problem due to several reasons. First, the human body is composed of richly articulated parts such as head, hands, legs, and so on, which makes modeling motion a very complicated combinatorial optimization problem. Furthermore, the human motion encodes at least two types of information: spatial and temporal information. Spatially, human body has 78 joints with more than 200 degrees of freedom (DOF) (curse of dimensionality) that are governed by physiological and biomechanical factors (Magnenat-Thalmann and Thalmann 1991). Temporally, the motion is a dynamic time-series depending on many factors such as the laws of mechanics and external stimuli (e.g., music for dancing). This time-varying property makes detecting human parts more intractable. Second, it is hard to extract stable parts of human body from images and videos. For example, traditional image processors cannot clearly identify the limb boundaries in order to separate each body part from the background because of loose and textured clothes. In addition, the background objects in the natural scene may occlude parts of human body. Third, the background are not stable either. Rather than static and uniform background, when the human move in the outdoors, the background is changing all the time and nearly unpredictable. Also, there may be other moving objects in the scene such as cars and animals, leading to a more cluttered scene. All above reasons make detecting human motion a challenging problem in real world.

Two important psychological theories can guide us to build a salient model for human motion. The first lies in Johansson displays (Johansson 1973) that vivid human motions can be completely perceived through only the motion of light bulbs fixed on the main joints of human body. So, we can model the parts of human body and their intercalations in order to detect and recognize human activities in complex scenes. These body parts can be represented as feature points extracted from images. The second psychological evidence originates from Biederman's recognition-by-components (RBC) theory

X.-Q. Cao (✉) · Z.-Q. Liu
School of Creative Media, City University of Hong Kong,
Kowloon Tong, Hong Kong
e-mail: xiaoqcao@student.cityu.edu.hk

(Biederman 1987), which suggests that the visual input like moving bulbs of human motion is matched against the structural representation of a human pose in the brain. Therefore, we can represent statistical-structurally the human pose in terms of parts.

Traditional part-based representation of the human body divides into four forms: blob, silhouette, box and point (more details can be found in a recent survey (Moeslund and Granum 2001; Moeslund et al. 2006). Those extracted features are usually based on body segments according to some similarities such as coherent flow, similar colors, and cooccurrenced intensities of image pixels. Compared with the point features, the former three representations may be hard to detect in the situations of severe occlusion and time-varying background. Recently, inspired by Johansson's displays (1973), Song et al. (2003) used point features to represent different parts of a human body. In their work, the human body is represented by a decomposable triangulated graph (DTG). For each triplet of the graph a Gaussian model of relative positions and velocities of the body parts is learned. By maximizing the joint probability density function of the parts (triplets), the best labeling is found. Indeed, the task of human motion detection can be considered as a labeling problem or graph matching problem, in which the best labeling configuration is assigned to a set of input observed points including cluttered background points.

In this paper, we present an Markov random field-based (MRF-based) graph matching algorithm for the human motion detection using point features in an occluded and cluttered scene in Fig. 1. The MRFs have been widely employed to solve both low- and high-level computer vision problems (Li 2001). Most of the MRF-based applications are low-level processing such as image restoration and reconstruction, which directly process true image pixels. Recently, the use of MRFs in high-level vision such as handwritten Chinese character recognition (Zeng and Liu 2008a, b; Zeng et al. 2010) has emerged based on segmented character stroke features. In this paper, we also use the MRF model in high-level vision for segmented point features. The success of MRF models in computer vision no matter in low-level or high-level processing has been largely ascribed to its ability of modeling context-dependent observations (observation-based graph) through characterizing mutual influences among such observations using MRF-based probabilities (MRF-based graph) (Li 2001). In the observation-based graph, the node represents each observation $\mathbf{o}_i$ over sites $1 \leq i \leq I$, while the edge between observations imply potential mutual influences. We may define the neighborhood system $\mathcal{N}_i$ and clique potentials V_c to characterize the observation-based graph. The neighborhood system $\mathcal{N}_i$ is the set of neighbors of the site i, and clique c is the subset of sites that are all pairwise neighbors, consisting either of a single site $c = \{i\}$, or of a pair of neighboring sites $c = \{i, i'\}$. In the MRF, we also define the edge among all the labels, where the node represents each semantic label j, $1 \leq j \leq J$ to explain the observation, and the edge between two labels reflects some prior constraints of the labeling configuration. We obtain the best labeling configuration by matching both the constrained observation-based and MRF-based graphs.

To encourage or penalize various local graph matching possibilities, we assign the matching costs V_c to the cliques in the neighborhood system. According to the Hammersley–Clifford theorem (Li 2001), the joint probability of the labeling configuration at all sites in the MRF follows a Gibbs distribution associated with an energy function which is a sum of all clique potentials. Therefore, the desirable global graph matching is achieved by minimizing the corresponding joint energy function of observation- and MRF-based graphs. In our work we address the problem of defining and estimating an MRF model of human motion and use it for detecting the human body configurations (assign a feature point to each part of the human body) in Johansson displays (a number of markers are attached to the body). For each single-site clique representing a body part, a Gaussian model of its velocity is learned. Similarly, we learn a Gaussian model of its relative position between body parts for each pair-site clique. When these clique potentials are added together, they compose the joint energy function of the entire body. Therefore, human motion can be accounted for within the MRF framework. Compared with Song's DTG method (Song et al. 2003), our MRF strategy is a more general framework, and thereby can handle various cases. Better effects can be achieved through designing proper neighborhood system or clique potentials in MRFs for different human poses. More specifically, DTG is a simplified case of MRFs with triplet cliques. In this paper, we focus on designing specific clique potentials for characterizing clutters and employ partial labeling for occlusion.

In the next section, we introduce related work on human motion detection and tracking. In Sect. 3 we propose MRFs for the human motion detection as a labeling problem. In Sect. 4, we report and analyze the experimental results. Finally, we draw conclusions and envision future improvements in Sect. 5.

2 Related work

The contour (Fig. 2b) and articulated shape (Fig. 2c, d) are two descriptive and convenient object representations for detection and tracking nonrigid human body poses, where the contour is a closed region that define the boundary of the body pose shape and the inside region in a contour is called silhouette. Yilmaz et al. (2006) used contour-based body

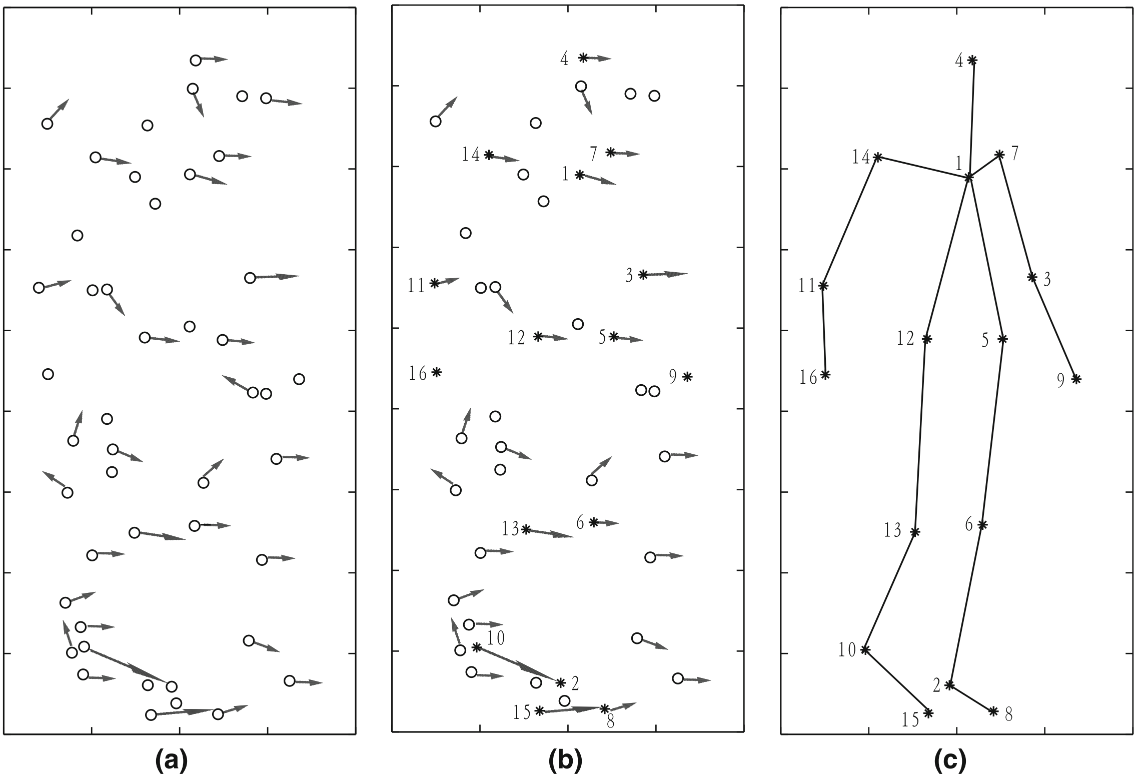

Fig. 1 **a** shows feature points including velocity and position extracted from consecutive frames of human motion, where the *arrows* denote the velocity of each point. **c** shows the MRF-based human motion model, where the *stars* denote the labels for human parts: *1* neck, *2* left ankle, *3* left elbow, *4* head, *5* left waist, *6* left knee, *7* left shoulder, *8* left toe, *9* left wrist, *10* right ankle, *11* right elbow, *12* right waist, *13* right knee, *14* right shoulder, *15* right toe, *16* right wrist. We aim to find the best labeling configuration or correspondence (**b**) between **a** and **c** to explain the observed feature points. The *star points* in **b** are feature points labeled by the corresponding human parts. Those *unlabeled points* in **b** are background clutters

representation to perform tracking by evolving the contour from frame to frame through minimizing the image energy and the shape energy in the gradient descent direction. The image energy is based on observations measured by color and textures using semiparametric models and those two types of features are fused using independent opinion polling. On the other side, the shape energy is based on the past contour observations, which is used to recover the missing object regions during occlusion. Kang et al. (2004) proposed a contour-based appearance descriptor which is characterized by multiple models representing spatial distribution of object's color and shape. This appearance descriptor is invariant to 2D rigid transformation and scale change over wide range of transformation within a large resolution. Chen et al. (2001) provided a parametric shape to model the body contours which also incorporate various image cues such as edge intensity, foreground color and background color. They used HMM states to represent the contour point locations along each normal lines, where the HMM transition probability is computed by a joint probability data association filter (JPDAF). Color is a widely used feature in tracking systems, however, the limitations of the color is that it is very sensitive to illumination and rely on the surface reflectance properties of the target (Yilmaz et al. 2006). To avoid this, Haritaoglu et al. (2000) designed a real time visual surveillance system W^4 for night-time or other low light level situations by using intensity feature. The foreground regions are extracted from background by grouping pixels according to their intensity which is described by each pixel's maximum and minimum intensity values, and the maximum intensity difference between consecutive frames. Optical flow is another commonly used feature in tracking. Mansouri et al. (2002) use the optical flow as the contour constraint to tracking, where the optical flow vector is computed for every pixel inside the contour region as a Bayesian estimation. Similarly, Bertalmio et al. (2000) also used optical flow constraint to evolve the contour in a given image to a desired position in a second image.

Contour-based body pose representation provides complete and accurate shape of the body pose. Alternatively, the part-based method can also represent articulated body pose shape. It includes two strategies: body parts based and body joints based pose representations in Fig. 2c, d.

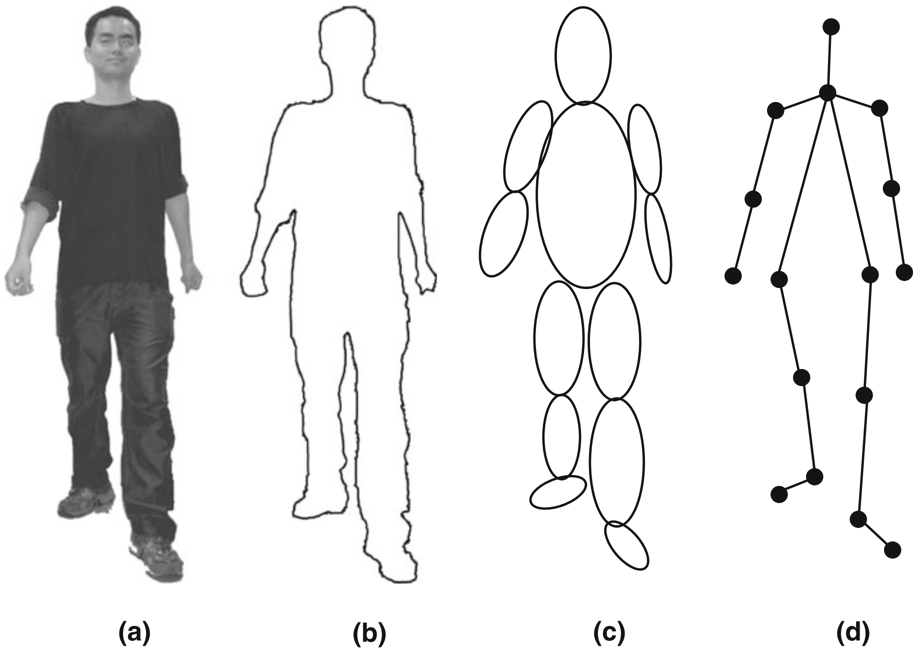

Fig. 2 **a** Human pose, **b** contour, **c** patch, and **d** point features

The former represents the body parts including head, legs, arms and torso by geometric shapes such as rectangle, ellipse, cylinder, etc. Lan et al. (2008) built a 3D human model with six body limbs based on some structural and kinematical constraints for tracking people walking in the side view. A main difficulty in human motion modeling comes from the high-dimensionality of the DOF which often makes the problem intractable. To reduce the high-dimensional parameters including locations and joint angles, in their work six independent particle filter (PF) based trackers are used for estimating some initial pose, then nonparametric belief propagation (NBP) is used as the post-processing of the initially estimated joint angles. Wu et al. (2003) used a dynamic Markov network to analyze subparts locally while reinforcing the structural constraints, in which mean field Monte Carlo algorithms are used for tracking different body parts. Experiments were carried out on different number of body parts without clutter and missing value. By compared with the results using multiple independent trackers (MiTs) which track without considering relationships between subparts, they conclude that MiTs are not suitable for tracking articulated body parts. Hua et al. (2005) used a Markove network to model 2D human pose from single images to track soccer players in cluttered background. They employed a quadrangular shape to represent body part and used principle component analysis (PCA) for shape parameter dimensionality reduction. They proposed a data driven belief propagation Monte Carlo algorithm to inference pose parameters from image cues including appearance, shape, edge and color, which carries out Bayesian inference in parallel. To avid slow sampling process in NBP, Han et al. (2006) propose an efficient NBP by sequentially mode propagation and kernel fitting. Their tracker achieves robust tracking for dancing, meeting and treadmill walking from frontal view and oblique view, but fails in severe self-occlusion and large 3D motions. All above works were initialized by hand. Sigal et al. (2004) presented a automatic method for tracking human bodies in 3D by using existing various body part detectors. They proposed a loose-limbed body model as a graph model in 3D in which the limbs are not rigidly connected by are rather attracted to each other. They solved human pose and motion estimation with NBP using a variation of PF that can be applied over a general loopy graph. Experiments on a walking person in video using four calibrated cameras prove that loose-limbed model is more robust than traditional kinematic model when the image quality is poor. Recently, some authors (Song et al. 2003) used a set of articulated points with each representing a main body joint to define a body pose for detection. Compared with body contour and patch segments, the most important advantage of point features is that they are more easier to obtain in severe occlusion and clutter. Unlike those works using connected geometric shapes to represent body pose, Song et al. (2003) used articulated points to represent a pose. While most works only consider pairwise relationships between adjacent body parts, they used a mixture of decomposable triangulated graphs to model human poses. They presented an unsupervised learning algorithm to obtain a probabilistic model of a moving human body automatically from unlabeled and cluttered image sequences. Their experiments on side walking person show that their method is able to detection motions with high accuracy.

3 Methods

3.1 Labeling problems

We first list some important notations for a reference purpose:

- $\mathcal{J} = \{1, \ldots, j, \ldots, J\}$: a set of J labels for human body parts.
- $\mathcal{I} = \{1, \ldots, i, \ldots, I\}$: a collection of I sites.
- $\mathbf{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_i, \ldots, \mathbf{o}_I\}$: a set of observed feature points.
- $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_I)^{\mathrm{T}}$: a labeling configuration of the sites in $\mathcal{I}$ in terms of the labels in $\mathcal{J}$.
- $\mathcal{N}$: the neighborhood system on labels $\mathcal{L}$.
- $\mathcal{N}'$: the neighborhood system on sites $\mathcal{I}$.
- c: a clique that all sites in it are pairwise neighbors.
- $\mathcal{C}_1$: a set of single-site cliques.
- $\mathcal{C}_2$: a set of pair-site cliques.
- V_1: a single-site clique potential.
- V_2: a pair-site clique potential.

The human motion detection can be formulated as a labeling problem. We assign a label to each point-based observation, where some observations are from the human body and others belong to the background (noisy points). Let $\mathcal{J} = \{1, \ldots, j, \ldots, J\}$ be a set of J labels for human parts including head, neck, hand, leg, etc., and $\mathcal{I} = \{1, \ldots, i, \ldots, I\}$ indexs a set of I observed points in an image plane. Usually, the number of labels J is not equal to the number of points I because of noise points. Correspondingly, let $\mathbf{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_i, \ldots, \mathbf{o}_I\}$ represent a vector of observed measurements at all points. The labeling problem is to find a matching, which is a set of semantic labels assigned to a set of points to explain the observations. Let $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_I)^{\mathrm{T}}$ be a labeling configuration at all points, where the labeling strength, $\mathbf{f}_i = f_i(j) \in [0, 1]$, measures the assignment probability of the label j to the point i. If label j is definitely assigned to point i, we denote $f_i = j$ or $f_i(j) = 1$. The null label is not assigned to any points denoted by $\sum_i f_i(j) = 0$, and the null point is not associated with any labels denoted by $\sum_j f_i(j) = 0$. Our goal is to find the best labeling configuration $\mathbf{F}^*$ for the observation $\mathbf{O}$, over all possible label vectors $\mathbf{F}$,

$$\mathbf{F}^* = \arg\max_{\mathbf{F}} P(\mathbf{F}|\mathbf{O}). \tag{1}$$

According to Bayes' law:

$$P(\mathbf{F}|\mathbf{O}) = \frac{p(\mathbf{O}|\mathbf{F})P(\mathbf{O})}{p(\mathbf{O})}, \tag{2}$$

because $p(\mathbf{O})$ is a constant for a given fixed $\mathbf{O}$, so, we turn to maximizing the joint probability $P(\mathbf{O}, \mathbf{F})$, which is factored as the likelihood function of $\mathbf{F}$ with respect to $\mathbf{O}$ and the prior probability of $\mathbf{F}$,

$$\mathbf{F}^* = \arg\max_{\mathbf{F}} \{p(\mathbf{O}|\mathbf{F})P(\mathbf{F})\}. \tag{3}$$

3.2 Neighborhood system and cliques

Two fundamental concepts in MRF are neighborhood system $\mathcal{N}$ and clique potential $\mathcal{V}_c$, which describes and measures the relations among different features (Li 2001). The pair $(\mathcal{J}, \mathcal{N}) = \mathcal{G}$ defines a graph for body pose, where $\mathcal{J}$ contains the nodes with each representing a body part and $\mathcal{N}$ determines the edges between the nodes according to the neighboring relationship. In our study, the neighboring set $\mathcal{N}_j$ of a node j is defined as the rest set of other nodes as

$$\mathcal{N}_j = \{j' | \forall j' \in \mathcal{L}, j' \neq j\}, \tag{4}$$

Similarly, the pair $(\mathcal{I}, \mathcal{N}') = \mathcal{G}'$ constitutes a graph in an usual image. In this case, the set of sites $\mathcal{I}$ replaces the set of labels $\mathcal{J}$. The neighborhood system $\mathcal{N}'$ on $\mathcal{I}$ consists of all the other sites $\mathcal{N}'_i = \{i' | \forall i' \in \mathcal{I}, i' \neq i\}$. The clique c is defined as a subset of nodes in $\mathcal{I}$ composed of both a single-site clique set $\mathcal{C}_1 = \{i | i \in \mathcal{I}\}$ and pair-site clique set $\mathcal{C}_2 = \{i, i' | i' \in \mathcal{N}_i\}$ to describe the dependencies between different parts of body. We choose to model the body pose and motion by the joint probability of the velocity and relative position of its parts. Correspondingly, the unary feature $\mathbf{o}_i$ is the velocity of the point i, $\mathbf{o}_i = (v_{ix}, v_{iy})^{\mathrm{T}}$, where x and y are horizontal and vertical axes of frames. We also use binary features $\mathbf{o}_{ii'}$ to represent the pairwise relationship between body parts. The relative position between two body parts is $\mathbf{o}_{ii'} = (p_{ix} - p_{i'x}, p_{iy} - p_{i'y})^{\mathrm{T}}$. The relative position $\mathbf{o}_{ii'}$ is used here so that we can deal with the translation invariance.

3.3 MRFs

According to the equivalence between MRFs and Gibb distribution established by Hammersley and Clifford (Li 2001), we can rewrite (3) in terms of the following Gibbs distribution form,

$$\frac{e^{-U(\mathbf{F}|\mathbf{O})}}{Z_1} = \frac{e^{-U(\mathbf{O}|\mathbf{F})}}{Z_2} \frac{e^{-U(\mathbf{F})}}{Z_3}. \tag{5}$$

Since it is a constant for the same MRF-based human pose model $Z_1 = Z_2 = Z_3$, we neglect the normalization factors in the Gibbs distribution. Thus, maximizing (3) is equivalent to minimizing the following joint energy function

$$U(\mathbf{F}|\mathbf{O}) = U(\mathbf{O}|\mathbf{F}) + U(\mathbf{F}). \tag{6}$$

For simplicity, if we assume that the unary and binary constraints are conditionally independent given the labeling configuration $\mathbf{F}$, thus, we can obtain the likelihood energy as

$$U(\mathbf{O}|\mathbf{F}) = \underbrace{\sum_{i \in \mathcal{C}_1} V(\mathbf{o}_i|\mathbf{f}_i) + \sum_{ii' \in \mathcal{C}_2} V(\mathbf{o}_{ii'}|\mathbf{f}_i, \mathbf{f}_{i'})}_{\text{likelihood clique potentials}}, \tag{7}$$

and the prior energy as

$$U(\mathbf{F}) = \underbrace{\sum_{i \in \mathcal{C}_1} V(\mathbf{f}_i) + \sum_{ii' \in \mathcal{C}_2} V(\mathbf{f}_i, \mathbf{f}_{i'})}_{\text{prior clique potentials}}, \tag{8}$$

Indeed, Eq. 6 is the log-likelihood of Eq. 7 if we derive the likelihood clique potentials from the likelihood functions. The energy function is a sum of clique potentials over all possible cliques c including both the single-site likelihood clique potential $V(\mathbf{o}_i|j)$[1] that encodes the statistical velocity information of observation $\mathbf{o}_i$ given the label j, and the pair-site likelihood clique potential $V(\mathbf{o}_{ii'}|j, j')$ which statistically describes the dependencies between $\mathbf{o}_i$ and $\mathbf{o}_{i'}$ given the labels j and j'. In our study, we assume that every body part is equally to appear, so the single-site prior clique potentials $V(j)$ are the same for all body parts. On the other side, the pair-site prior $V(j, j')$ encodes the prior dependency knowledge of neighboring body parts, which play the role in penalizing the inconsistent matching.

If we assume the likelihood function in (7) as Gaussian mixture models (GMMs) (Zeng and Liu 2008), we derive the single-site likelihood clique potential as

$$V(\mathbf{o}_i|j) = -\ln\left[\sum_{m=1}^{M_s} w_{jm} N(\mathbf{o}_i; \boldsymbol{\mu}_{jm}, \Sigma_{jm})\right], \tag{9}$$

and derive the pair-site likelihood clique potential as,

$$V(\mathbf{o}_{ii'}|j, j') = -\ln\left[\sum_{m=1}^{M_s} w_{jj'm} N(\mathbf{o}_{ii'}; \boldsymbol{\mu}_{jj'm}, \Sigma_{jj'm})\right], \tag{10}$$

where M_s is the number of mixture components, and w_{jm}, $w_{jj'm}$ are the weights of the mixture components. $N(\mathbf{o}; \boldsymbol{\mu}, \Sigma)$ is a multivariate Gaussian distribution,

$$N(\mathbf{o}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{o}-\boldsymbol{\mu})}, \tag{11}$$

where d is the dimensionality of $\mathbf{o}$. We use four mixtures in our experiments. All parameters can be estimated from training samples using the expectation-maximization (EM) algorithm (Zeng and Liu 2008).

3.4 Relaxation labeling (RL)

The human pose detection is equivalent to find the best matching between the observation set $\mathbf{O}$ and label set $\mathcal{J}$ according to the associated unary properties and binary relations, and the energy U is the matching cost. We use RL to search the best labeling configuration in order to minimize (6). First, we convert the minimization of the posterior energy into the maximization of a corresponded gain function

$$g(\mathbf{O}, \mathbf{F}) = \sum_{j=1}^{J} \sum_{i=1}^{I} \left[w_1 K_i(j) f_i(j) + w_2 \sum_{j'=1}^{J} \max_{i' \in \partial i} K_{ii'}(jj') f_i(j) f_{i'}(j') \right], \tag{12}$$

which is the sum of compatibility functions defined by the clique potentials,

$$K_i(j) = \mathrm{CONST}_1 - V(\mathbf{o}_i|j) - V(j), \tag{13}$$

$$K_{ii'}(jj') = \mathrm{CONST}_2 - V(\mathbf{o}_{ii'}|j, j') - V(j, j'), \tag{14}$$

where the constants CONST_1 and CONST_2 ensure that both compatibility functions are non-negative. Because the single-site (velocity) and pair-site clique potential (relative position) may play different roles in labeling body parts, we add weights w_1 and w_2 to them, satisfying $w_1 + w_2 = 1$. We first equally initialize each observation to each label. Subsequently, we update the labeling strength $f_i^t(j)$ until t reaches the fixed number T (T is set to 10 in our experiments) as shown in Algorithm 1. Finally, we use the winner-take-all strategy to assign the label j to the point i with the maximum $f_i^T(j)$.

```
input     : $\mathbf{O}, V$.
output    : $\mathbf{F}^*$.
initialize: $f_i^0(j) \leftarrow$ initial labeling,
            $K_i(j) \leftarrow CONST_1 - V(\mathbf{o}_i|j) - V(j)$,
            $K_{ii'}(jj') \leftarrow CONST_2 - V(\mathbf{o}_{ii'}|j, j') - V(j, j')$.
begin
    for $t \leftarrow 1$ to $T$ do
        for $j \leftarrow 1$ to $J$ do
            for $i \leftarrow 1$ to $I$ do
                $q_i(j) \leftarrow$
                $w_1 K_i(j) + w_2 \sum_{j'} \max_{i'} K_{ii'}(jj') f_{i'}(j')$;
                $f_i^{t+1}(j) = \frac{f_i^t(j) q_i^t(j)}{\sum_i f_i^t(j) q_i^t(j)}$;
            end
        end
    end
    for $j \leftarrow 1$ to $J$ do
        $i \leftarrow \arg\max_i f_i^T(j)$;
        $f_i^*(j) \leftarrow 1, \sum_i f_i^*(j) \leftarrow 1$;
    end
end
```

Algorithm 1: The relaxation labeling algorithm

The designed MRF has powerful descriptive ability for representing atomic pose by using both motion and structural constraints, thus can handle the occlusion and clutter well. To further enhance the performance, we design some heuristic methods. Instead of finding a label for each point, we assign a point to each label, so the unassigned points

[1] We denote $V(\mathbf{o}_i|f_i = j)$ as $V(\mathbf{o}_i|j)$.

belong to the background. We use the winner-take-all strategy to get the best configuration for each body part, but if the best candidate's $f_i^T(j)$ does not exceeds the predefined threshold 0.1, we consider it as a null label. As a result, each label is assigned to only one best candidate point; however, during matching, two labels may be assigned to the same point. In this case, we retain the label with the higher pair-point compatibility, and assign the other label to the second candidate point which does not overlap with other labeled body parts.

4 Experiments

4.1 Datasets and performance measures

To evaluate the proposed MRF's performance on human motion detection from different viewpoints, we use 3D motion capture data from ACCAD database (http://www.accad.osu.edu/research/mocap/mocap_data.htm). We convert 3D motion to 2D by orthographic projection to simulate a camera view from desired viewpoints. We define that a view angle 0° represents a left-side view and a view angle 90° responds to a frontal view of the subject. The walking sequences are around 4 seconds with a length of 581 frames at 120 samples per second. Originally, there are 42 markers fixed on the body parts. For simplicity, in our experiment, we use only 16 markers corresponding to the positions of main joints of human body including head, neck, left and right shoulders, left and right elbows, left and right wrists, left and right waists, left and right knees, left and right ankles, and left and right toes. The absolute position is meaningless because of time-varying scenes for different coordinates. Thus, we use the velocity and the relative position as the measurements for the point feature. For each frame, the velocity is set to the difference between two successive frames under the assumption that points could be tracked for two consecutive frames. In all our experiments, we split data into ten disjunct sets. From these sets, ten different training sets were built by joining nine of the sets in turn, while the tenth set was used as a test set for a tenfold crossvalidation. Finally, we report the average correct/error rate. A frame error occurs if at least one body part is labeled wrongly. A label error occurs in two situations: (1) a body part is present, but it is wrongly labeled to a other point or an empty point. (2) a body part is not present, but it is still assigned to a point. Thus, the label error is usually smaller than the frame error.

4.1.1 Balance between velocity and relative position

First, we investigate 11 combinations of weights w_1 (from 0 to 1) and w_2 (from 1 to 0) for the single-point clique potential and pair-point clique potential. The training and testing are done under the same view angle 45° which is between the side view and front view. Fig. 3 shows that the frame error is 2.56% and the label error is low at 0.32% when we only consider the relative position of body parts, but the frame error and label error are almost 100% when we only use the velocity to measure each body part. Consequently, we conclude that the relative position of body parts play the major role in accurate labeling performance. Also, we find that when we increase the weight w_1 for single-point clique potential, both the frame error and label error decrease and reach the lowest value 0.37 and 0.05% when w_1 = 0.8 and w_2 = 0.2. As a result, we may conclude that the velocity play the auxiliary role in decreasing the error rate in an uncluttered scene. So we use the best balance w_1 = 0.8 and w_2 = 0.2 in the rest of the experiments.

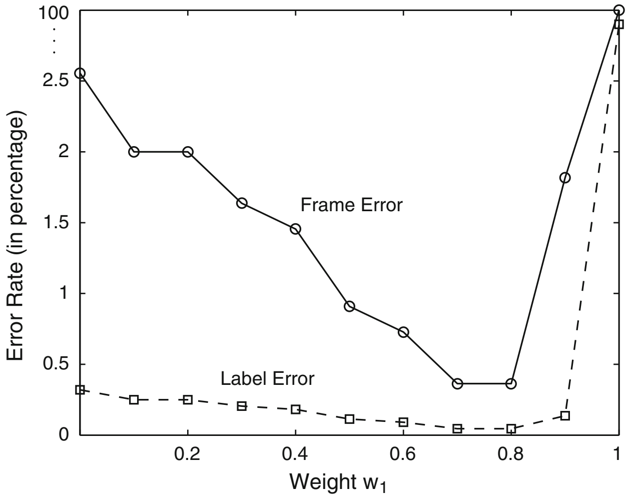

Fig. 3 Error rate for different weights w_1

Further investigation of the behavior of each body parts shows that the performance of some labels is much worse in Fig. 4. For example, the worst labeled body parts are left ankle and right ankle when they are very very close (almost in the same position) in some frames because we did not consider the self-occlusion from 3D to the 2D image plane. Furthermore, more detailed analysis reveals that the second candidate label for the wrongly labeled point is the correct choice. Therefore, our MRF model can encode the point statistical distribution quite well.

4.1.2 Performance with changing viewpoints

In the following experiment, we explore our model's performance when the testing viewing angle is different from that used during training. In the training phase, all the original 3D data are projected to 2D at a viewing angle of

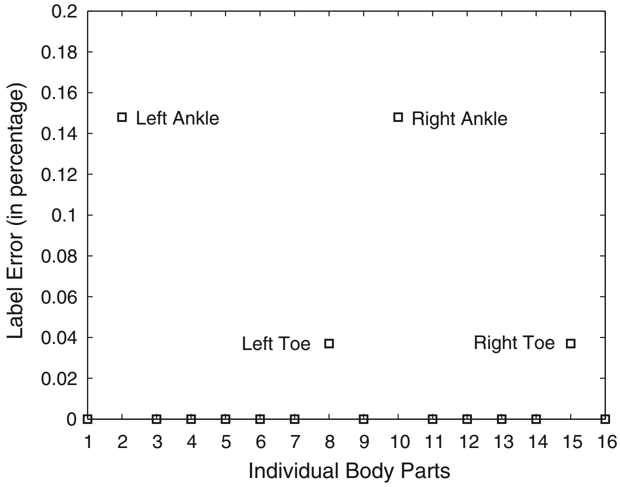

Fig. 4 Average error rate of body parts. *1* neck, *2* left ankle, *3* left elbow, *4* head, *5* left waist, *6* left knee, *7* left shoulder, *8* left toe, *9* left wrist, *10* right ankle, *11* right elbow, *12* right waist, *13* right knee, *14* right shoulder, *15* right toe, *16* right wrist. Most body parts can be correctly labeled in all frames when there are no clutter and missing body parts. The left ankle and right ankle are the body parts being worst labeled with a error rate of 0.35%

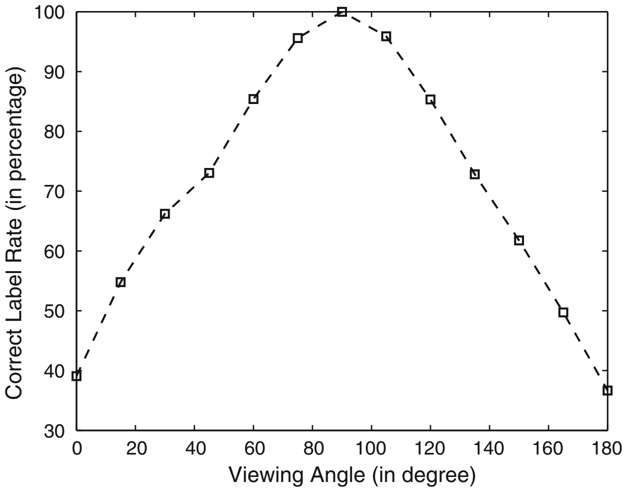

Fig. 5 Error rate from different view points

90° (frontal view), while in the testing phase, the viewing angle change from 0° to 180° with step increment of 15°. Figure 5 shows that when the viewing angle is between 60° and 120°, the correct label rate is higher than 85%, thereby demonstrating that our MRF model is robust to changes in viewpoint by up to 30° leaned at 90°. When the camera view is turned to the side view, there are errors by the mismatch of left and right ankles, left and right feet, left and right knees, etc. In most cases, the mismatch happens between the left and right joints. Since we use the orthographic projection without inter-body occlusions, this error is not surprising.

4.1.3 *Performance in cluttered scenes*

Furthermore, we examine if our model is robust in a sparse cluttered scene and a dense cluttered scene as shown in Fig. 6. The positions and velocities of cluttered points (totally 30 background points are added) are independently and randomly generated within image sizes by the Gaussian distribution. In a sparse cluttered scene, the whole sequence's leftmost, rightmost, upmost, downmost positions and velocities are used as the range of the four directions. On the other hand, in a dense cluttered scene, the current frame's horizontal and vertical maximum and minimum are used as the range for cluttered points' positions and velocities. Table 1 shows that the correct frame rate and label rate is very high in the sparse cluttered background. Even under the condition that the background points are very dense surrounding the human body, our result is also very encouraging. The correct frame rate is low, which is caused by one or two labels usually for the left or right ankle. In practice, we think that this kind of error can be ignored because two ankles are too close to differentiate. Note that the correct label rate remains at a very high 95.05% even in the dense cluttered background.

4.1.4 *Performance with missing body parts in cluttered scenes*

Finally, we evaluate our model's performance when some parts of the human body are occluded in cluttered scenes. For every frame, we randomly delete a few points as if they are missing, so in principle each body part has an equal chance to be occluded. We variate the number of missing body parts from 15 to 1. Figure 7 shows that the correct label rate decreases as the the number of features from body parts goes down because some occluded body parts are assigned to the near noisy points. We also see that even in the worst situation the correct label rate exceeds 50% when only one body part is present in the scene, which is much higher than the chance level (1/31, 31 is the number of candidate features detected including 1 body part point and 30 cluttered points).

5 Conclusions

We have presented the MRF to detect human motion, which is formulated as a labeling problem by minimizing the joint energy function of the MRF. Experimental results indicate that the proposed single clique potential and pair clique potential can describe each body part and their dependencies very well. The method is robust to different viewpoints, showing a very good performance with large variations. Furthermore, the MRF performs stably in both

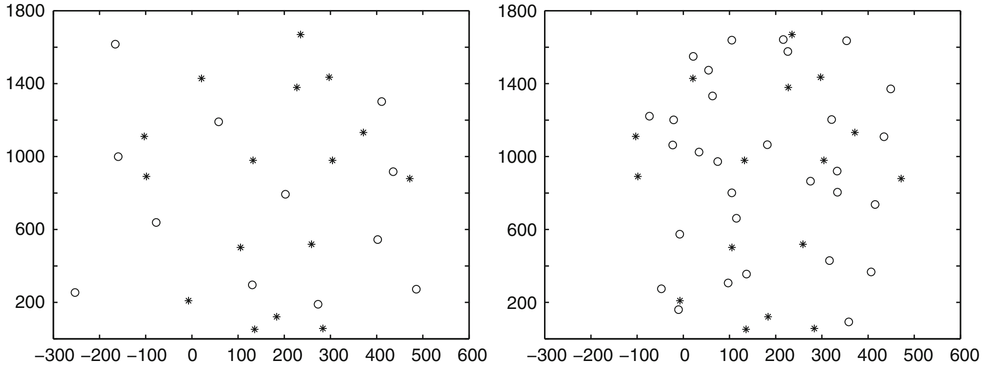

Fig. 6 Labeling and detection in a sparse cluttered (*left*) scene and a dense cluttered scene (*right*). *Circles* represent background points and *stars* represent body points

Table 1 Correct rates in sparse and dense cluttered scenes

Correct rate	Sparse (%)	Dense (%)
Label	99.55	95.05
Frame	94.55	44.03

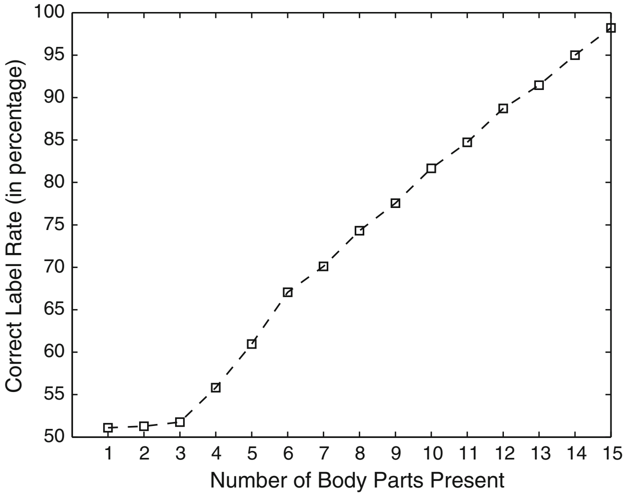

Fig. 7 Correct label rate for different number of body parts present

sparse and dense cluttered scenes. Finally, the results with missing body parts are also encouraging. The data used in this work are motion capture data, where markers are put on the main joints of the human body to facilitate the identification of important body part features. When real image data are considered in markerless condition, we can use some point detectors such as KLT (Tomasi and Kanade 1991) and SIFT (Lowe 2004) to obtain the velocity and position information.

Acknowledgments The work described in this paper was partial supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 118608).

References

Song Y, Goncalves L, Perona P (2003) Unsupervised learning of human motion. IEEE Trans Pattern Anal Mach Intell 25(7):1–14

Yu Q, Medioni G (2009) Multiple target tracking by spatio-temporal Monte Carlo Markov chain data association. IEEE Trans Pattern Anal Mach Intell 31(12):2196–2210

Isard M, Blake A (1998) Condensation-conditional density propagation for visual tracking. Int J Comput Vis 29(1):5–28

Lee MW, Nevatia R (2009) Human pose tracking in monocular sequence using multilevel structured models. IEEE Trans Pattern Anal Mach Intell 31(1):27–38

Wu Y, Yu T (2006) A field model for human detection and tracking. IEEE Trans Pattern Anal Mach Intell 28(5):753–765

Khan Z, Balch T, Dellaert F (2005) MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans Pattern Anal Mach Intell 27(11):1805–1819

Magnenat-Thalmann N, Thalmann D (1991) Complex models for animating synthetic actors. IEEE Comput Graph Appl 11(5):32–44

Johansson G (1973) Visual perception of biological motion and a model for its analysis. Percept Psychophys 14:201–211

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94(2):115–147

Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81:231–268

Moeslund TB, Hilton A, Kruger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104:90–126

Li SZ (2001) Markov random field modeling in image analysis. Springer, Tokyo

Zeng J, Liu Z-Q (2008a) Markov random field-based statistical character structure modeling for handwritten Chinese character recognition. IEEE Trans Pattern Anal Mach Intell 30(5):767–780

Zeng J, Liu ZQ (2008b) Type-2 fuzzy Markov random fields and their application to handwritten Chinese character recognition. IEEE Trans Fuzzy Syst 16(3):747–760

Zeng J, Feng W, Xie L, Liu Z-Q (2010) Cascade markov random fields for stroke extraction of chinese characters. Inf Sci 180:301–311

Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM Comput Surv 38(4):13

Kang J, Cohen I, Medioni G (2004) Object reacquisition using invariant appearance model. In: IEEE International Conference on Pattern Recognition (ICPR), pp 759–762

Chen Y, Rui Y, Huang TS (2001) JPDAF based HMM for real-time contour tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 543–550

Haritaoglu I, Harwood D, Davis LS (2000) w^4: Real-time surveillance of people and their activities. IEEE Trans Pattern Anal Mach Intell 22(8):809–830

Mansouri AR (2002) Region tracking via level set PDEs without motion computation. IEEE Trans Pattern Anal Mach Intell 24(7):947–961

Bertalmio M, Sapiro G, Randall G (2000) Morphing active contours. IEEE Trans Pattern Anal Mach Intell 22(7):733–737

Lan SF, Ho MF, Huang CL (2008) Human motion parameter capturing using particle filter and nonparametric belief propagation. In: IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), pp 37–40

Wu Y, Hua G, Yu T (2003) Tracking articulated body by dynamic markov network. In: IEEE International Conference on Computer Vision (ICCV), pp 1094–1101

Hua G, Yang MH, Wu Y (2005) Learning to estimate human pose with data driven belief propagation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 747–754

Han TX, Ning H, Huang TS (2006) Efficient nonparametric belief propagation with application to articulated body tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 214–221

Sigal L, Bhatia S, Roth S, Black MJ, Isard M (2004) Tracking loose-limbed people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 421–428

Tomasi C, Kanade T (1991) Detection and tracking of point features, Technical Report CMU-CS-91-132, Carnegie Mellon Univ

Lowe D (2004) Distinctive image features form scale-invariant keypoints. Int J Comput Vis 60(2):91–110