PAPER

Efficient Human Body Tracking by Quick Shift Belief Propagation

Kittiya KHONGKRAPHAN[†], *Nonmember and* **Pakorn KAEWTRAKULPONG**[†a)], *Member*

SUMMARY We propose a novel and efficient approach for tracking 2D articulated human body parts. In our approach, the human body is modeled by a graphical model where each part is represented by a node and the relationship between a pair of adjacent parts is indicated by an edge in the graph. Various approaches have been proposed to solve such problems, but efficiency is still a vital problem. We present a new Quick Shift Belief Propagation (QSBP) based approach which benefits from Quick Shift, a simple and efficient mode seeking method, in a part based belief propagation model. The unique aspect of this model is its ability to efficiently discover modes of the underlying marginal probability distribution while preserving the accuracy. This gives QSBP a significant advantage over approaches like Belief Propagation (BP) and Mean Shift Belief Propagation (MSBP). Moreover, we demonstrate the use of QSBP with an action based model; this provides additional advantages of handling self-occlusion and further reducing the search space. We present qualitative and quantitative analysis of the proposed approach with encouraging results.

key words: *human body tracking, belief propagation, quick shift*

1. Introduction

Tracking human body parts or body pose has recently attracted increased attention from computer vision researchers. These approaches typically focus on the body parts in more detail than tracking human location as a whole. This is an important problem to solve because it can serve as a front end for higher-level processes to understand human activities in several applications, such as human-computer interaction, virtual reality, and character animation. A number of algorithms have been proposed to address human body pose tracking. One of the initial approaches was based on the use of *markers*; however, this method requires a great deal of user intervention. The system using markers is uncomfortable in many applications due to the use of special equipment which must be installed on the body. On the other hand, the marker-less system can run without special equipment.

Among the marker-less methods, several researchers turn to image-based approaches. In the image-based techniques, features from image (s) are served as its input to estimate human pose. There are two main approaches: discriminative and generative approaches. In the discriminative approach (or bottom up approach), candidate body parts are first extracted from the image and then used to generate possible configurations of human body. The performance of this approach is based on results of feature extraction since it is used to represent body parts.

In the generative approach (or top down approach), human is normally modeled as a skeleton model and the main concept is an effort to fit a predefined human model into image data. A large number of projected samples are generated and their similarity with respect to the input image is measured and compared. Such approaches are generally computationally intensive because of the complex search over the high dimension state space. The computational cost of these approaches is $O(N^m)$, where N is the number of samples for each body part and m is the number of body parts. Several approaches have been proposed to reduce the computation load. Constraints are normally introduced to limit the search space for some specific applications, e.g. walking parallel to a plane [1] or golf swing movements [2]. Motion models and prior knowledge are also employed to predict the state space trajectory or to reduce the search space [1]. Lee et al. [3] detected some body parts separately and used them to update subsets of the state space. Alternative approaches attempted to reduce the number of samples using support vector machines [4], annealed particle filtering [5], hybrid Monte Carlo filtering [6] and kernel particle filtering [7].

Several researchers turned to part-based approach that employs Bayesian inference concept to estimate human pose. It has been popular due to reduction in computational cost from $O(N^m)$ to $O(mN^2)$ [8]–[14]. In this approach, the human body is represented by a graphical model where each part is represented by a node and the relationship between the adjacent parts is indicated by an edge of the graph. Each body part is considered separately and then is combined into the global solution later. However, the computational time is still quite substantial, which limits its application from real-time human body tracking. Various approaches [15], [16] have been proposed to alleviate this problem, but efficiency is still a vital problem. The high-dimensional space is generally associated with exponential increase of computation time, and that is the motivation of our paper.

The main contribution of our paper is the introduction of a novel method for fast tracking of articulated body while maintaining high precision. The proposed QSBP model is based on the key idea of efficiently refining the estimate of the mode of marginal probability in each iteration of belief propagation. This leads to the reduction in compu-

Manuscript received May 13, 2010.
Manuscript revised November 8, 2010.

[†]The authors are with the Department of Control System and Instrumentation Engineering, King Mongkut University of Technology Thonburi 126 Prachautid, Bangmod, Toongkru, Bangkok, 10140, Thailand.
a) E-mail: pakorn.kae@kmutt.ac.th

DOI: 10.1587/transinf.E94.D.905

tation time since it requires fewer iterations than MSBP (Mean Shift Belief Propagation) [16] and require fewer samples than NBP (Non-parameter belief propagation) [17]–[19], [26]–[29]. The samples only in the close proximity of the current estimate of mode are used. We utilize Quick Shift [20] for mode seeking method in each iteration to set initial samples. In addition, we also demonstrate the use of a model video, when possible; to further reduce the search space for QSBP (Quick Shift Belief Propagation). Using similar action based models have been shown to be useful for addressing self-occlusion problem [21] as a motion model.

The paper is organized as follows. After the general introduction, related work is presented in this section. Our proposed method is described in Sects. 2 and 3. In Sect. 4, we present some experimental results. Finally, the conclusion is discussed in Sect. 5.

1.1 Related Work

The part-based is an approach that is more powerful than others since it can reduce computational cost from $O(N^m)$ to $O(mN^2)$. This approach represents parts of the human body with a graphical model. The main idea is to pass messages iteratively between adjacent nodes of the graph. Two popular techniques used to calculate the messages are the belief propagation and mean field Monte Carlo methods [8]. Their difference lies in the pattern of messages. Shen at el. [9] compared and reported that belief propagation offered better performance than that of mean field Monte Carlo.

The tracking is performed by generating a number of candidate samples for each node and then calculating their beliefs. The beliefs are computed from observation functions and messages. Ramanan et al. [10], [11] proposed 2D human tracking by considering candidate parts first and then combining those to find the optimal solution by Belief Propagation (BP). Gao and Shi [12] detected human faces and hands by color cues to guide the sample generation for face and hand parts in order to achieve better efficiency in tracking. In some papers, both messages and beliefs (or marginal distributions) are represented by weighted samples [13], [14], whereas several other researchers represented messages with a more complex continuous distribution such as mixtures of Gaussians [17]–[19] or NBP. The samples in NBP are drawn by the Gibbs sampler for all iterations, which leads to a huge increase in computational time. To alleviate this problem of NBP, Han et al. [15] proposed to approximate the mixture of Gaussians by mode propagation and kernel filtering. They report that their method is 80 times faster than BP for tracking a 2D articulated body model; however, it is still far from adequate for real-time tracking requirements. Park et al. [16] proposed MSBP. They further reduce time complexity by computing only samples that moves toward the best solution. It is 30-50 times faster in 2D state vector tracking, and 300 times faster in 3D state vector tracking than BP.

To handle self-occlusion problem, some approaches utilize multiple cameras [5], [27]. The occlusion constraint is proposed in likelihood computation [28]. Wang and Mori proposed occlusion and spatial constraints by representation of human with multiple tree models [30]. Some approaches make use of the prior 2D and/or 3D information about the structure or kinematics of human body. Some shape based [22], motion-based [4], [23], and a combination of both approaches [24], [25] have been proposed in the literature. Their availability of large databases of shapes and motion patterns increases robustness to viewpoint change.

However, tracking 2D articulated human body parts is still a difficult problem with high computational cost. In this paper, we focus on this problem. We introduce a novel and efficient approach for tracking 2D articulated human body parts. It extends the belief propagation by applying mode seeking. As we will show in Sect. 2.3, the computational complexity of our proposed approach is approximately 36 and 99 times faster than those of MSBP and BP in case of 4-state (2D position plus body part length and angle) tracking. We also incorporate model video to limit our search space and to enhance tracking accuracy especially in case of occlusions.

1.2 Quick Shift

Quick Shift, proposed by Vedaldi and Soatto [20], is a simple and extremely efficient mode seeking method. Like Mean Shift, Quick Shift is a local optimization algorithm. Mean Shift can be regarded as a gradient ascent method [31] while the Quick Shift does not require gradient information. Quick Shift is a quick Euclidean version of medoid shift that is guaranteed to converge for all starting locations [32].

In mode seeking techniques, it is started by defining the multivariate kernel density estimate. Like these techniques, Quick Shift also starts by computing the kernel density estimate

$$f(\mathbf{a}) = \frac{1}{M}\sum_{i=1}^{M}\varphi(\mathbf{a}-\mathbf{a}_i), \tag{1}$$

where $\mathbf{a}_i$ is the i^{th} data point and $\mathbf{a}_i \in \mathcal{X} = \mathbb{R}^d$, $\varphi(\mathbf{a})$ is a kernel function [32] (e.g. Gaussian) and M is the number of data points. The main concept of Quick Shift is the iterative movement of each mode estimate from its current position to a new position which is the nearest neighbor with higher probability until a mode is reached. The updated position of data point $\mathbf{a}_i$ at iteration $k+1$ is computed by

$$\begin{aligned}\mathbf{y}_i^{k+1} &= \underset{\mathbf{a}_j\in\{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_M\}:P(\mathbf{a}_j)>P(\mathbf{y}_i^k)}{\arg\min} D(\mathbf{y}_i^k,\mathbf{a}_j) \quad ,\\ P(\mathbf{b}) &= \frac{1}{M}\sum_{j=1}^{M}\varphi(D(\mathbf{b},\mathbf{a}_j)),\end{aligned} \tag{2}$$

where $D(\mathbf{y}_i,\mathbf{a}_j)$ is the distance between current positions of y_i^k and data point $\mathbf{a}_j$, $P(\mathbf{a}_i)$ is probability value of data point $\mathbf{a}_i$. The mode seeking is repeated on $\{\mathbf{y}_i^k\}$ until no further change in labelling occurs and then the modes are obtained

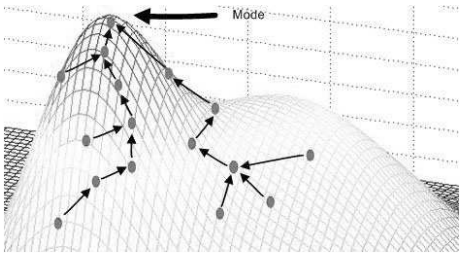


Fig. 1 Probability surface and motion of data points toward mode value of Quick Shift. Motion directions and data points are plotted by arrows and dots, respectively.

as the set of unique locations

$$mode = \{\mathbf{y}_i^t\}, \tag{3}$$

where $\mathbf{y}_i^t$ is the final position of $\mathbf{y}_i$.

All data points (initial values of mode estimates) move toward a single mode as shown in Fig. 1. To balance under- and over-fragmentation of the modes, a threshold parameter, κ, is introduced into Eq. (2)

$$y_i^{k+1} = \underset{\mathbf{a}_j \in \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M\}: P_j(\mathbf{a}_j) > P_i(y_i^k), D(y_i^k, \mathbf{a}_j) < \kappa}{\arg\min} D(y_i^k, \mathbf{a}_j). \tag{4}$$

2. Proposed Method

At each frame, a silhouette of the human body is obtained by a background subtraction. The silhouette as well as its original image are served as the input in our tracking method. The main concept of our approach is applying mode seeking to reduce computational time. A model video concept [21] is used for initializing the motion model in our proposed method. The initial state is served as an initial configuration for our QSBP. For the next step, samples are generated around the initial state and their probabilities are measured by belief propagation. Only the best solution of each node is selected to be a part of the optimal pose solution. We applied Quick Shift with belief propagation for mode seeking in each body part so that all samples (initial values of the mode estimates) are not necessary moved to a single point. In this sense, it reduces risks of getting into spurious solutions which have highest probability. From our approach, the modes are selected as initial samples in the next iteration of belief propagation.

2.1 Human Model

We model a 2D view of a human body as a graphical model with N hidden nodes and pair-wise potentials as shown in Fig. 2 (a). The hidden nodes are represented by $\mathbf{X} = \{\mathbf{x}_i | i \in [1, N]\}$, where $\mathbf{x}_i$ is the i^{th} body part consisting of 4 states, i.e., position coordinate, orientation and length. A corresponding observation set is denoted by $\mathbf{Z} = \{\mathbf{z}_i | i \in [1, N]\}$, where $\mathbf{z}_i$ is the image observation node for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by an observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$. In addition, every pair of adjacent body parts $\mathbf{x}_i$ and $\mathbf{x}_j$ (as defined by the structure), is connected and encoded by a potential function, $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$. Human body parts represented in the model are shown in Fig. 2 (b).

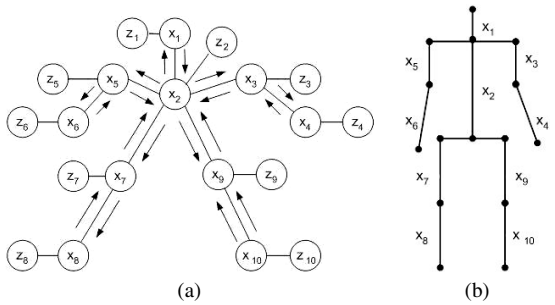


Fig. 2 The skeleton human model in our approach: (a) hidden nodes and pair-wise relationship and (b) joints and segments.

2.2 Model Video

A model video [21] is utilized for initializing the motion model in our approach. The main concept of the model video is to estimate the joint locations in the test video automatically using two geometric constraints. Given the locations of joints in the model video and the first frame of test video, the affine constraint is used to estimate initial positions based on invariance ratio between model video and test video. Moreover, the epipolar constraint is used to reduce estimation error of different actors and view points in the model and test videos.

One advantage of this work is the avoidance of error propagation from frame to frame in the estimation process, because each joint estimate is computed based on correspondence between the first frame of the model and the test videos. Another advantage is robustness to variations in anthropometry, execution rate, viewpoint and execution style. Moreover, it is not computationally expensive and does not require extensive training. We have found from empirical studies that a good tracking performance can be obtained if the model video gives an initial state within 1.5 times of the grid size. Figure 3 (a) shows the input frame. The candidate joints from the model video [21] are generated and overlaid on the input image as displayed in Fig. 3 (b). The best match with the silhouette is selected to be the initial samples as can be seen in Fig. 3 (c).

2.3 Quick Shift Belief Propagation

The belief propagation algorithm is an iterative method to infer the hidden state until it converges to the optimal solution. The marginal probability of $\mathbf{x}_i$ at iteration n, $p^n(\mathbf{x}_i|\mathbf{Z})$ can be computed by taking the product of incoming messages and local observations, as shown in Eq. (5). The incoming messages also contain prior knowledge of the node obtained from the neighboring nodes, as shown in Eq. (6).

$$p^n(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha \phi_i(\mathbf{x}_i, \mathbf{z}_i) \prod_{j \in \Gamma(i)} m_{ji}^n(\mathbf{x}_i). \tag{5}$$

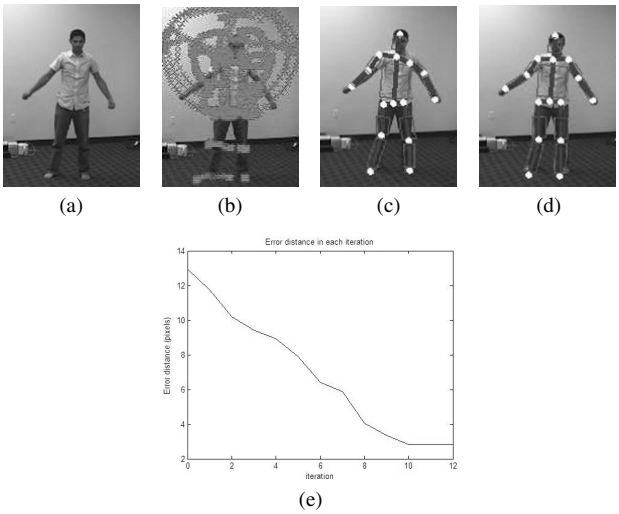


Fig. 3 QSBP tracking (a) input image, (b) predicted points from model video, (c) initial configuration, (d) final tracking result, (e) error distance in each iteration.

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i^{th} hidden and corresponding observation nodes, respectively. $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$ is the observation function of node i. α is a normalizing factor. For graphical models with continuous hidden state, $m_{ji}^n(\mathbf{x}_i)$ is a message sent from node j to i at iteration n and can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \int_{x_j} (\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)\phi_j(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j))d\mathbf{x}_j. \quad (6)$$

where $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$ is the potential function between nodes i and j, $\phi_j(\mathbf{x}_j, \mathbf{z}_j)$ is the observation function of node j and $\Gamma(j)\backslash i$ represents all the neighboring nodes of j except node i.

In our approach, we define parts of human body by nodes of graph that the optimal hidden node is computed by maximizing the marginal probability of each node given the current observation. Instead of considering all the possible states of our nodes, we work on a grid around initial samples (modes found in the previous iteration). A new discrete grid is generated around the mode. The grid is 5x5x3x3 of 4 states (x, y positions, length and angle of body part). The marginal probability of $\mathbf{x}_i$, $p(\mathbf{x}_i|\mathbf{Z})$ can be computed by taking the product of incoming messages and local observations, as shown in Eq. (5). The incoming messages also contain prior knowledge of the node obtained from the neighboring nodes. The message sent from node j to node i at iteration n, $m_{ji}^n(\mathbf{x}_i)$, can be calculated by

$$m_{ji}^n(\mathbf{x}_i) \leftarrow \sum_{x_j} \left(\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)\phi_j(\mathbf{x}_j, \mathbf{z}_j) \prod_{k\in\Gamma(j)\backslash i} m_{kj}^{n-1}(\mathbf{x}_j) \right). \quad (7)$$

The message at the first iteration, $m_{ji}^0(\mathbf{x}_i)$, is defined to be 1. From the iterative concept of belief propagation, the best solution is obtained in the final iteration by

$$\mathbf{x} = \arg\max_{\mathbf{s}_i\in\{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_N\}} p^n(\mathbf{s}_i|\mathbf{Z}), \quad (8)$$

where N is the number of samples in the grid. Figure 3 (d) shows the best solution in the final iteration of tracking and Fig. 3 (e) illustrates error distance in each iteration of tracking using our approach. It can be seen that the error distance is reduced until reaching the best solution.

2.4 Mode Seeking in Belief Propagation

In mode seeking, we use marginal probability of belief propagation in movement of mode estimates. Only samples around the modes are computed, not the entire surface of belief propagation. This makes the convergence to an optimal solution very fast and computation time is reduced. The updated position of sample $\mathbf{s}_i$ at iteration $k+1$ of the mode seeking is computed by

$$\mathbf{y}_i^{k+1} = \arg\min_{\mathbf{s}_j\in\{\mathbf{S}_1,\mathbf{S}_2,...,\mathbf{S}_N\}:P(\mathbf{S}_j|\mathbf{Z})>P(\mathbf{y}_i^k|\mathbf{Z}),D(\mathbf{y}_i^k,\mathbf{S}_j)<\kappa} D(\mathbf{y}_i^k, \mathbf{s}_j), \quad (9)$$

where $D(\mathbf{y}_i^k, \mathbf{x}_j)$ is the distance between current positions of $\mathbf{y}_i^k$ and sample $\mathbf{x}_j$ which is less than a distance threshold κ and $P(\mathbf{x}_i|\mathbf{Z})$ is the marginal probability value of sample i. From Quick Shift mode seeking concept, the position of $\mathbf{y}_i$ is updated until no further change in labelling occurs and then modes of samples are obtained as a unique set

$$mode = \{\mathbf{y}_i^t\}, \quad (10)$$

where y_i^t is the final position of $\mathbf{y}_i$.

We find the mode of marginal probability by our approach which is faster than by MSBP [16]. Because moving toward a mode of Quick Shift is based on the locally maximum probability value, while the Mean Shift is based on weighted mean value. In this sense, the number of iterations in moving of Quick Shift approach is less than that of Mean Shift which makes its converge to the optimal solution much faster than Mean Shift. Figures 4 (a) and (b) show convergence to the optimal solution by QSBP and MSBP, respectively. From these figures, each initial sample is plotted by a solid square marker. The only grid members (plotted by markers inside of a grid window) around the initial sample are considered in moving toward a mode of the sample. The new position of the mode estimate in each iteration until they reach a mode are shown by a circle marker.

Main steps of proposed approach

1. Generate initial joint predictions from model video [21] as shown in Fig. 3 (b).
2. Select the best match to be initial samples as shown in Fig. 3 (c).
3. Iterate steps 4-7 until convergence.

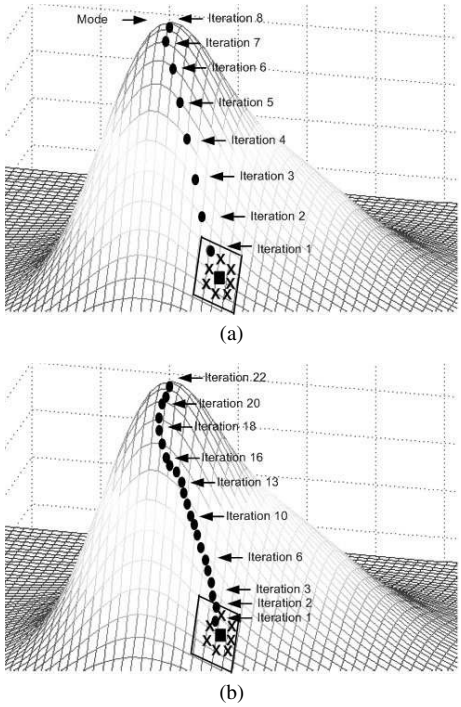


Fig. 4 A comparison of moving toward a mode value by (a) Quick Shift, and (b) Mean Shift.

4. Generate a local grid of samples around each initial sample.
5. Compute message by Eq. (7).

$$m_{ji}^{n}(\mathbf{x}_i) \leftarrow \sum_{x_j} \left(\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j) \phi_j(\mathbf{x}_j, \mathbf{z}_j) \prod_{k \in \Gamma(j) \setminus i} m_{kj}^{n-1}(\mathbf{x}_j) \right).$$

6. Compute marginal probability using in Eq. (5).

$$p^{n}(\mathbf{x}_i|\mathbf{Z}) \leftarrow \alpha \phi_i(\mathbf{x}_i, \mathbf{z}_i) \prod_{j \in \Gamma(i)} m_{ji}^{n}(\mathbf{x}_i).$$

7. Seek mode using Quick Shift, set each mode as initial sample for the next iteration.
8. Select the best solution as

$$\mathbf{x} = \arg\max_{x_i \in \{x_1, x_2, \ldots, x_N\}} p^{n}(\mathbf{x}_i|\mathbf{Z}),$$

2.5 Observation Function and Potential Function

The observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$ is used to measure the joint likelihood of $\mathbf{z}_i$ and $\mathbf{x}_i$. In order to measure the likelihood, each body part $\mathbf{x}_i$ is modelled by a planar patch and projected onto the input image, and then its likelihood (or similarity) is computed. In this work, region overlapping [3], [5] and RGB color are the features used for similarity measurement. The region overlapping feature of node $\mathbf{x}_i$ is computed by

$$w_i = \frac{1}{2}\left(\frac{n_{i,o}}{n_{i,p}} + \frac{n_{i,o}}{n_m}\right), \quad n_m = \arg\max_i \; n_{i,o} \tag{11}$$

where $n_{i,o}$ is the number of pixels in overlapping region between projected region of i^{th} sample and the silhouette. $n_{i,p}$ is the number of pixels in the i^{th} projected region. This measure is very simple and efficient; however, its reliability reduces greatly in case of occlusion. This is because overlapped parts form a larger foreground region which provides high values of this feature in several false locations.

In case of occlusion, we therefore switch to using a more detailed color feature (Sect. 3 explains how to detect the beginning and the end of occlusion). For the color feature, each human body part is formed by a color histogram based on RGB components. From the color models of the node $c(\mathbf{x}_i)$ and the template, t, the similarity measure is defined by the histogram intersection between the color shape matching model of the sample and the template as

$$w_i = \sum_{j=1}^{n} c_{j,i} \cap t_j, \tag{12}$$

where $c_{j,i}$ is the normalized number of pixels in the j^{th} bin of the i^{th} node and t_j is the normalized number of pixels in the j^{th} bin of template in RGB color histogram and n is the number of bins in each histogram. The observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$ of node $\mathbf{x}_i$ is computed by

$$\phi_i(\mathbf{x}_i, \mathbf{z}_i) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(1-w_i)^2}{2\nu^2}}, \tag{13}$$

where w_i is the similarity of node $\mathbf{x}_i$ obtained by Eq. (11) or Eq. (12) and ν is its standard deviation. For a low value of ν, a more weight is given to the appearance similarity.

The $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$ potential function is used to show the relationship between body parts i and j. We model the potential function by a Gaussian to represent the distribution of the Euclidean distance between the two adjacent body parts.

$$\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{D(\mathbf{X}_i, \mathbf{X}_j)^2}{2\sigma^2}}, \tag{14}$$

where $\mathbf{x}_i$ is the i^{th} body part, $\psi_{ji}(\mathbf{x}_i, \mathbf{x}_j)$ is a potential function of body parts i and j, $D(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between the connected points of body parts i and j, and σ is its standard deviation. In the same way as ν, σ specifies insensitivity to displacement of adjacent body parts.

3. Occlusion Handling

A common problem in human body tracking is self-occlusion problem. To reduce the computational complexity and handle self-occlusion problem of tracking in our method, we used prediction information from a model video [21] as the motion model. It is used for sample generation in the first iteration. We use two measures for detecting start and end of self-occlusion as in [21]. The first measurement is α_j^t, which represents the area of the foreground silhouette, corresponding to the j^{th} segment in the t^{th} frame.

The second measure is β_j^t, which represents the proportion of the detected segment j that is occluded by the other segments of the cardboard model. The condition for occlusion is based on the normalized change over time τ,

$$\frac{\sum_{i=t-\tau}^{t-1} \alpha_j^{i+1} - \alpha_j^i}{\tau \alpha_j^{t-\tau}} < T \ , \ \frac{\sum_{i=t-\tau}^{t-1} \beta_j^{i+1} - \beta_j^i}{\tau \beta_j^{t-\tau}} > T. \qquad (15)$$

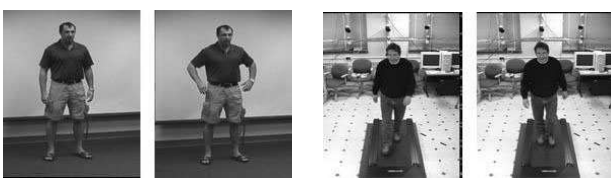
where T is the percentage threshold and τ is the number of previous frames that are used to consider occlusion. Positive T value signifies occlusion entering, while the negative T value indicates occlusion termination.

4. Experiments

To evaluate the performance of our proposed method, we tested it on several videos and compared it with the BP and MSBP methods in both simple and occlusion cases. We tested our method on 6 sequences of UCF dataset, containing 400, 162, 400, 560, 500 and 500 frames. These videos included aerobics style activities that were also used in [21]. Moreover, we experimented on 2 sequences of CMU dataset (also used in [16]), containing 199 and 190 frames including walking action in both front and side views. These videos include both simple and self-occlusion cases. To present qualitative and quantitative results, we compared our approach with [16]. Note that the joint locations were manually initialized in the first frame, and then an RGB template of each human part is automatically generated from the initialized joint locations. The sample prediction was then performed automatically for the remaining frames like the method presented in [21].

In our experiments, we generated samples with a size of 8,000 samples for BP, and a 5×5×3×3 grid for both MSBP and QSBP, respectively. Those methods were run until convergence (joint position movement is less than 2 pixels or 50 iterations are reached.) The threshold parameter in mode seeking of Quick Shift and the kernel size of MSBP were chosen as half of the grid size. For occlusion handling, we used 70 percentage in our experiments and the number of pervious frames was chosen as 15 to consider occlusion. The bin size of RGB histogram was 16×16×16. In observation function, the standard derivation of observation function and potential function were chosen as 0.4 and 3, respectively.

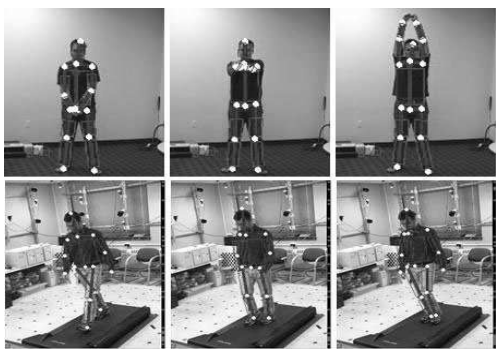
We applied model video to the sample prediction in the first iteration of belief propagation. The samples of model video on UCF and CMU datasets are shown in Figs. 5 (a) and (b), respectively. Samples of tracking results are displayed in Fig. 6. The figure shows samples of human model fitting on 4 sequences. The fitting results of QSBP, BP and MSBP approaches are shown in Figs. 6 (a), (b) and (c), re-



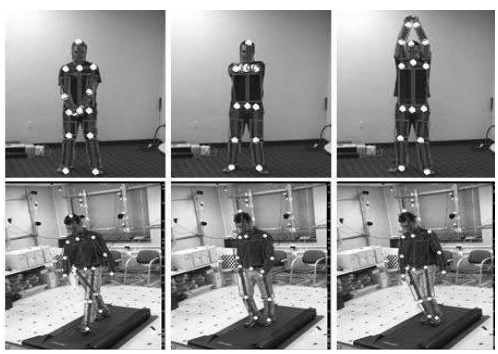
(a) (b)

Fig. 5 The samples of model video (a) UCF dataset (b) CMU dataset.

(a)



(b)



(c)

Fig. 6 Some results of human body tracking by using model video in sample prediction. Similar results from (a) Quick Shift belief propagation, (b) Belief propagation and (c) Mean Shift propagation.

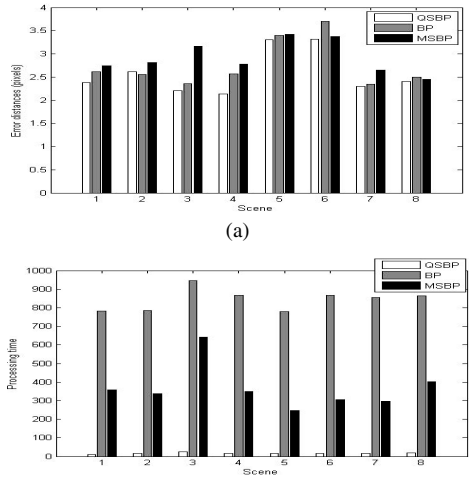


Fig. 7 A performance comparison between the proposed method and BP and MSBP are shown by white, gray and black bars, respectively (a) Accuracy (b) Efficiency.

spectively. Since similar results among all methods were obtained, only two results were included for BP and MSBP in Figs. 6 (b) and (c). It can be seen that the model fits well to the images for each approach. The results show that the accuracy of QSBP is comparable to those of BP and MSBP as illustrated in Fig. 7 (a). However, the computational time of QSBP is significantly less than those of BP or MSBP (Fig. 7 (b)). In particular, for our case of 4-state (2D position plus body part length and angle) tracking, our proposed technique is respectively 36 and 99 times faster than those of MSBP and BP. On average, the numbers of iterations to get the best solution are 9, 30 and 42 for QSBP, MSBP and BP, respectively. The 2D tracking results are compared with ground truth which is manually obtained. The average distance errors from the corresponding ground truth are shown in Fig. 7 (a). It can be seen that all methods provide accuracy; however, our approach is far more efficiency than the others as shown in Fig. 7 (b).

5. Conclusion

We propose a part-based tracking algorithm by integrating Quick Shift, a simple and efficient mode seeking method, into the belief propagation framework. The main idea is to find the mode of marginal probability of belief propagation to be used to predict points in the next iteration, and that way only samples around modes are computed in each iteration of belief propagation. Therefore, our proposed method needs fewer samples than NBP or MSBP. In addition, it converges to the best solution faster than the other methods. This approach can reduce the computational complexity dramatically due to the reduction of search space while preserving accuracy. In addition, we apply model video in the first iteration of belief propagation for predicting and resolving occlusion problems. The method was experimented on several videos and the results showed very good performance and robustness in both accuracy and efficiency.

Acknowledgement

This work is supported by Office of the Higher Education Commission (OHEC) and the Nation University Research (NRU) program. This work was conducted at UCF Computer Vision. The authors want to thank Professor Mubarak Shah and Arslan Basharat for fruitful discussion on various ideas in this paper.

References

[1] Z. Zhou, A. Prugel-Bennett, and R.I. Damper, "A Bayesian framework for extracting human gait using strong prior knowledge," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.11, pp.1738–1752, Nov. 2006.
[2] R. Urtasun, D.J. Fleet, and P. Fua, "Monocular 3D tracking of the golf swing," Proc. CVPR, pp.932–938, June 2005.
[3] M.W. Lee, I. Cohen, and S.K. Jung, "Partile filter with analytical inference for human body tracking," Proc. Workshop on Motion and Video Computing, pp.159–165, Dec. 2002.
[4] H. Sidenbladh, "Detecting human motion with support vector machines," Proc. ICPR, pp.188–191, Aug. 2004.
[5] J.D.A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering," Proc. CVPR, pp.126–133, 2000.
[6] K. Choo and D.J. Fleet, "People tracking using hybrid Monte Carlo filtering," Proc. CVPR, pp.321–328, 2001.
[7] C. Chang and R. Ansari, "Kernel particle filter: Iterative sampling for efficient visual tracking," Proc. ICIP, pp.977–980, Setp. 2003.
[8] Y. Wu, G. Hua, and T. Yu, "Tracking articulated body by dynamic Markov network," Proc. ICCV, pp.1094–1101, Oct. 2003.
[9] C. Shen, A.V. Hengel, A. Dich, and M.J. Brooks, "2D articulated tracking with dynamic Bayesian networks," Proc. Computer and Information Technology., pp.130–136, Sept. 2004.
[10] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.1, pp.65–81, Jan. 2007.
[11] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," Proc. CVPR, pp.271–278, June 2005.
[12] J. Gao and J. Shi, "Multiple frame motion inference using belief propagation," Proc. Automatic Face and Gesture Recognition, pp.875–880, May 2004.
[13] G. Hua and Y. Wu, "Multi-scale visual tracking by sequence belief propagation," Proc. CVPR, pp.826–833, July 2004.
[14] G. Hua, M. Hsuan, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," Proc. CVPR, pp.747–754, June 2005.
[15] T.X. Han, H. Ning, and T.S. Huang, "Efficient nonparametric belief propagation with application to articulated body tracking," Proc. CVPR, pp.214–221, June 2006.
[16] M. Park, Y. Liu, and R.T. Collins, "Efficient mean shift belief propagation for vision tracking," Proc. CVPR, pp.1–8, June 2008.
[17] M. Isard, "Pampas: Real-valued graphical models for computer vision," Proc. CVPR, pp.613–620, June 2003.
[18] E.B. Sudderth, A.T. Ihler, W.T. Freeman, and A.S. Willky, "Nonparametric belief propagation," Proc. CVPR, pp.605–612, 2003.
[19] E.B. Sudderth, M.I. Mandel, W.T. Freeman, and A.S. Willsky, "Visual hand tracking using nonparametric belief propagation," Proc. CVPRW, pp.189–196, June 2004.
[20] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," Proc. ECCV, pp.705–718, Aug. 2008.
[21] A. Gritai, A. Basharat, and M. Shah, "Geometric constraints on 2D action models for tracking human body," Proc. ICPR, pp.1–4, 2008.

[22] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial structures for object recognition," Int. J. Comput. Vis., vol.61, no.1, pp.55–79, 2005.
[23] G.R. Bradski and J.W. Davis, "Motion segmentation and pose recognition with motion history gradients," J. Machine Vision and Application, pp.174–184, 2002.
[24] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell., vol.3, no.23, pp.257–267, March 2001.
[25] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detection," Proc. ICCV, pp.90–97, 2005.
[26] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," Proc. CVPR, pp.421–428, 2004.
[27] L. Sigal, M. Isard, Sigelman, and M. Black, "Attractive people: Assembling losse-limbed models using non-parametric belief propagation," Proc. NIPS, 2003.
[28] L. Sigal and M. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," Proc. CVPR, pp.2041–2048, 2006.
[29] E. Sudderth, M.I. Mandel, W.T. Freeman, and A.S. Willsky, "Distribution occlusion reasoning for tracking with nonparametric belief propagation," Proc. NIPS, 2004.
[30] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," Proc. ECCV, pp.710–724, 2007.
[31] C. Yizong, "Mean shift, mode seeking, and clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.8, pp.790–799, Aug. 1995.
[32] Y.A. Sheikh, E.A. Khan, and T. Kanade, "Mode-seeking by medoidshifts," Proc. ICCV, pp.1–8, 2007.

Kittiya Khongkraphan received the B.Sc. degree in Mathematics and M.Sc. degree in Computer Science from Prince of Songkla University, Thailand, in 1991 and 2000, respectively. Currently, she is a Ph.D. candidate of King Mongkut's University of Technology Thonburi, Thailand. Her research interests are computer vision and image processing.

Pakorn Kaewtrakulpong received the B.Eng. degree in Electrical Engineering from King Mongkut's University of Technology Thonburi (KMUTT) in 1992. He received M.Sc. and Ph.D. degrees in Systems Engineering from Brunel University in 1998 and 2002, respectively. He is currently an associate professor at the faculty of engineering, KMUTT. His research interests include machine vision, industrial automation and instrumentations.