2012 IEEE International Conference on Robotics and Automation
RiverCentre, Saint Paul, Minnesota, USA
May 14-18, 2012

Iterative Pedestrian Segmentation and Pose Tracking under a Probabilistic Framework

Yanli Li, Zhong Zhou and Wei Wu

Abstract— **This paper presents a unified probabilistic framework to tackle two closely related visual tasks: pedestrian segmentation and pose tracking along monocular videos. Although the two tasks are complementary in nature, most previous approaches focus on them individually. Here, we resolve the two problems simultaneously by building and inferring a single body model. More specifically, pedestrian segmentation is performed by optimizing body region with constraint of body pose in a Markov Random Field (MRF), and pose parameters are reasoned about through a Bayesian filtering, which takes body silhouette as an observation cue. Since the two processes are inter-related, we resort to an Expectation-Maximization (EM) algorithm to refine them alternatively. Additionally, a template matching scheme is utilized for initialization. Experimental results on challenging videos verify the framework's robustness to non-rigid human segmentation, cluttered backgrounds and moving cameras.**

I. INTRODUCTION

Pedestrians, as the principal actors in daily life, have been widely studied in computer vision. Pedestrian segmentation and pose tracking are among the most active research topics for the last decade. This is partly due to their wide applications in video editing, human-computer interaction, surveillance etc., and partly due to their inherent difficulties caused by articulated bodies, occlusion, cluttered backgrounds and camera motions. Although the two problems have been extensively studied in the fields of object segmentation and object tracking, few work has be done on combining them together to improve each other, and the solutions for simultaneous object segmentation and tracking are hard to be applied directly to non-rigid humans.

Basically, the solutions for simultaneous object segmentation and tracking can be classified into three categories: MRF-based methods [1][2][3][4], level-set based methods [5][6] and template matching based methods [8][9]. Methods of the first category extract objects by optimizing a global cost function in a MRF. They either involve cumbersome interactions for indicating figure and ground regions, such as [1][2], or are sensitive to large deformation, such as [3][4]. Methods of the second category extract objects by tracking object boundaries. For highly-articulated pedestrians, the human boundaries tend to disappear in the case of self-occlusion, and hence this category has the drift problem. The third category extracts objects by matching the edge maps of the frames with shape templates. Therefore, a set of templates are required to be stored. Furthermore, considering the high variability in the shape and appearance of pedestrians, it is impossible for the limited templates to capture all detailed information for a particular pedestrian.

The difficulty of pedestrian segmentation lies in the non-rigid characteristic of human limbs. In the past decades, pose estimation for still images and pose tracking for videos have received extensive attention, both of which aim at estimating kinematic parameters of human body. Realizing the complementary merits of pose estimation and pedestrian segmentation, several authors [10][11][12] have combined them together and solved them simultaneously. However, these methods are limited to still images. In this paper, we extend them to handle video pedestrian segmentation and pose tracking using an EM algorithm within a unified framework. Basically, the framework belongs to MRF-based video object extraction. The main difference compared with other MRF-based methods [3][4] is the utilization of pose information to constrain human regions along videos, thus the framework can automatically extract human bodies, avoiding the commonly existing drift problem.

In general, the EM-based framework performs pose tracking in a physical-based Bayesian filtering, and pedestrian segmentation in a pair-wise Markov random field. The silhouette produced by segmentation is utilized as an observation cue in the pose tracking stage, and the skeleton produced by pose estimation is taken for establishing a distance penalty of the energy function in the segmentation stage. For initialization, we employ an Chamfer matching scheme [7] to infer the pose parameters. The major contribution of this paper is the proposed EM-based framework. The main advantage of the framework is being tolerant to large deformation because the adopted pose tracking can directly simulate realistic pedestrian walking. Meanwhile, the available silhouette provided by pedestrian segmentation serves as a useful cue for our pose tracker, which enables our framework more robust to moving cameras compared with the background subtraction methods, e.g. [13][14], which assume the cameras are stationary.

The remainder of this paper is organized as follows. After presenting the framework overview in section II, we describe the stages of segmentation and pose tracking in section III and section IV respectively. Experimental results are demonstrated in section V, and we conclude the paper in section VI.

This work is supported by the National 973 Program of China under Grant No. 2009CB320805, the National 863 Program of China under Grant No.2012AA011803, the Natural Science Foundation of China under Grant No.61170188 and Fundamental Research Funds for the Central Universities of China.

Y. Li, Z. Zhou and W. Wu are with the State Key Laboratory of Virtual Reality Technology & Systems, Beihang University, Beijing, China, zz@vrlab.buaa.edu.cn

978-1-4673-1405-3/12/$31.00 ©2012 IEEE

II. FRAMEWORK OVERVIEW

We handle the case of human walking along the arbitrary trajectory. The inputs to the framework are pedestrian windows corresponding to an individual human, which may be the outputs extracted by a human tracking-by-detection method [15] or manually. The task of our work is to derive spatial-temporal body poses and regions. Firstly, we divide the pedestrian windows into sequences, each of which is a walking cycle. Then we deal with each sequence individually as follows.

A. Problem Formulation

Defining the sequence of pedestrian frames by $I_t, t = 1, \ldots, T$, T is the frame number, we formulate the task as computing the maximum a posterior (MAP) in a first-order MRF, such that :

$$\begin{aligned}\Phi_t^* &= \arg\max_{\Phi_t} p(\Phi_t | I_1, \ldots, I_t, \Phi_1^*, \ldots, \Phi_{t-1}^*) \\ &= \arg\max_{\Phi_t} p(\Phi_t | I_1, \ldots, I_t, \Phi_{t-1}^*) \qquad (1)\end{aligned}$$

where the observation data I_t is a multi-cue composition that combines color cue I_t^c and motion cue I_t^m. The parameter set Φ_t is composed of three parts, $\Phi_t = \{\Omega_t, \Delta_t, \Theta_t\}$, in which Ω_t specifies the segment matte, Δ_t denotes the pose parameters, and Θ_t involves two sets of latent parameters, $\Theta_t = \{\Theta_t^c, \Theta_t^m\}$, which are used in the segmentation stage to model the color and motion distributions.

Maximizing the above posterior with respect to all parameters is intractable as the state space is expensively huge. Instead, we sequentially optimize them along the sequence using an EM algorithm. The E-Step is used to estimate the pose and latent parameters $\{\Delta_t, \Theta_t\}$, and the M-Step is used for pedestrian segmentation, i.e., obtaining matte Ω_t. As shown in Fig. 1, the algorithm consists of four main stages and is performed as: *E-Step I* $\rightarrow$ *M-Step* $\rightarrow$ *E-Step II* $\rightarrow$ *E-Step III* $\rightarrow$ *M-Step* $\rightarrow$ *E-Step II* $\rightarrow \cdots \rightarrow$ *E-Step III* $\rightarrow$ *M-Step.*

1) Initialization (E-Step I): the initial pose and latent parameters $\{\Delta_1, \Theta_1\}$ at the first frame are estimated with an improved Chamfer matching scheme [7].

2) Updating the segmentation (M-Step): the segment matte Ω_t is derived by optimizing a global MRF energy function under the constraints of $\{\Delta_t, \Theta_t\}$.

3) Updating the latent parameters (E-Step II): the latent parameters Θ_t are re-estimated with the refined segment Ω_t.

4) Pose tracking (E-Step III): we predict the pose and latent parameters $\{\Delta_{t+1}, \Theta_{t+1}\}$ at frame $t+1$ through a Bayesian filtering process using the previous segment matte Ω_t and parameters $\{\Delta_t, \Theta_t\}$.

B. Initialization (E-Step I)

At the beginning of each walking cycle, an improved chamfer matching [7] is employed to extract human silhouette and skeleton. As shown in Fig. 2(a), we require that the human legs are furthest apart in the first frame. The reasons are two folds: first, the furthest apart legs preserve least occlusion; second, the walking step length obtained under this condition is served as a main parameter in the pose tracking stage (see Section IV).

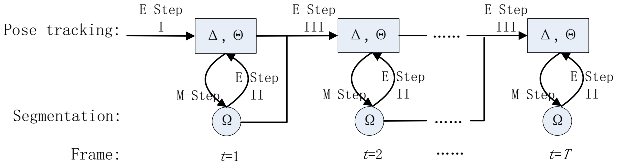


Fig. 1. The framework overview which consists of four kinds of steps: at E-Step I, pose and latent parameters $\{\Delta_1, \Theta_1\}$ are initialized by Chamfer matching; at M-Step, foreground body region Ω_t are extracted based on $\{\Delta_t, \Theta_t\}$ at the current frame; at E-Step II, parameters are re-estimated under the refined segmentation; E-Step III is used to infer subsequent human pose using previous segmentation result and human pose.

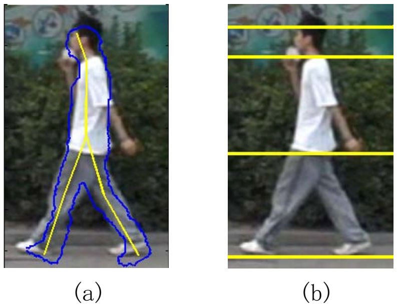


Fig. 2. (a) The derived silhouette (in blue) and skeleton (in yellow) are overlaid in the human body. (b) The four separators (in yellow) separate human body into three blobs, i.e., head, torso and legs.

Under the assumption that human body often consists of three main appearance regions, i.e., head, torso and leg, we decompose human body into three blobs to build appearance models. We divide the three blobs with four horizontal separators, in which the top and bottom separator are directly determined by the highest and lowest pixel in the silhouette, while the head-torso and torso-leg separators are estimated by minimizing the error in color classification. Supposing A, B and C are the head, torso and leg blobs respectively, we estimate the A-B separator as well as the B-C separator as a MAP(see Fig. 2(b)):

$$\begin{aligned}(h_{AB}^*, h_{BC}^*) = \arg \min_{(h_{AB}, h_{BC})} (\|\sigma_{Lab}(R_A(h_{AB})) + \\ \sigma_{Lab}(R_B(h_{AB}, h_{BC})) + \sigma_{Lab}(R_C(h_{BC}))\| + \\ \alpha_0(|h_{AB} - \mu_{AB}| + |h_{BC} - \mu_{BC}|)) \qquad (2)\end{aligned}$$

where $R_A(\cdot)$, $R_B(\cdot,\cdot)$ and $R_C(\cdot)$ denote the head, torso and leg blobs which are encircled by the silhouette and the separators, $\sigma_{Lab}(\cdot)$ is the color variance of the region, μ_{AB} and μ_{BC} are the mean separator locations derived from the skeleton, α_0 is a weighting value (set to 0.15 in our work). Given the separators (h_{AB}^*, h_{BC}^*), the color distributions $\Theta_1^c(\cdot)$ for the three blobs are established independently. The color model Θ_1^c in pixel (p_x, p_y) is taken by:

$$\Theta_1^c(p_x, p_y) = \begin{cases} \Theta_1^c(R_A), & if \quad p_y < h_{AB}^* \\ \Theta_1^c(R_C), & if \quad p_y > h_{BC}^* \\ \Theta_1^c(R_B), & \quad otherwise \end{cases} \qquad (3)$$

Besides, in the learning phase, we have manually clicked joints of all templates, which can be directly transferred to the frame for initializing the pose model Δ_1.

III. MRF FOR SEGMENTATION (M-STEP)

In this section, we utilize a pair-wise MRF to tackle the problem of pedestrian segmentation. Given the collection of pixels $X = \{x_i\}$ in the human window and the binary variant set $\{\Omega(x_i)\}$ associated with X, if x_i belongs to the human region, $\Omega(x_i) = 1$, otherwise $\Omega(x_i) = 0$, the human segmentation can be formulated as inferring a MAP-MRF across all configurations of Ω at the current frame, i.e.,

$$\Omega^* = \arg\max_{\Omega} p(\Omega | I_t, \Delta_t, \Theta_t) \tag{4}$$

The posterior distribution $p(\Omega|I_t, \Delta_t, \Theta_t)$ follows the Gibbs distribution[17]: $p(\Omega|I_t, \Delta_t, \Theta_t) = \exp(-E(X))/Z$. Here, Z is a normalization factor. The term $E(X)$ is known as an energy function which can be written as the sum of unary potentials and pair-wise potentials:

$$E(X) = \sum_i \varphi(x_i) + \sum_{i,j} \psi(x_i, x_j) \tag{5}$$

The pair-wise potential $\psi(x_i, x_j)$, as a smooth term, represents the penalty for assigning two neighboring nodes to any labels. In our work, it is given by:

$$\psi(x_i, x_j) = \begin{cases} \max(\breve{g}(x_i), \breve{g}(x_j)), & \Omega(x_i) = \Omega(x_j) \\ 1 - \max(\breve{g}(x_i), \breve{g}(x_j)), & \Omega(x_i) \neq \Omega(x_j) \end{cases} \tag{6}$$

Here, $\breve{g}(x)$ is the normalized magnitude value in the Pb edge map [16]. This definition suggests that the neighboring nodes should be assigned with different labels if one node has a large magnitude value since the node with a large magnitude value tends to lie in the figure-ground boundaries.

The unary potential, $\varphi(x_i) = -\log(p(\Omega(x_i)|I_t, \Delta_t, \Theta_t))$, is a negative log likelihood. It allows us to utilize multiple cues for human segmentation. In this work, we integrate four cues into the unary potential, referring to: 1) color term $\varphi_c(x)$, 2) motion term $\varphi_m(x)$, 3) pose term $\varphi_p(x)$, and 4) segment coherence term $\varphi_s(x)$, thus the unary potential can be rewritten as:

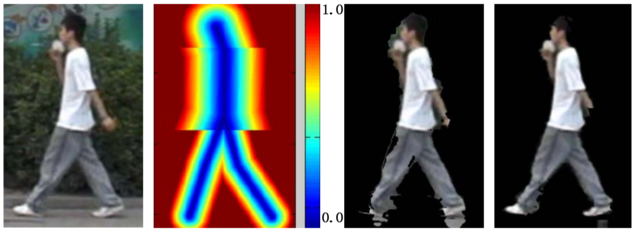
$$\varphi(x) = \lambda_c \varphi_c(x) + \lambda_m \varphi_m(x) + \lambda_p \varphi_p(x) + \lambda_s \varphi_s(x) \tag{7}$$

where $\{\lambda_c, \lambda_m, \lambda_p, \lambda_s\}$ are the weighting values.

Color term. The color distribution across one blob (e.g., torso) is typically compact, thus is considered as a vital cue for segmentation. We define the color model as K-Means clusters : $\Theta_t^c = \{\mu_{k,t}^{c,J} | k = 1, \ldots, K_c, J \in \{B, F\}\}$, in which K_c is the color cluster number (set to 3 in experiments) and $\mu_{k,t}^{c,J}$ is the mean color of the cluster (k, J), B indicates the background while F indicates the foreground. The color term is defined by:

$$\varphi_c(x) = \begin{cases} d_c^F(x)/(d_c^F(x) + d_c^B(x)), & \Omega(x) = 1 \\ d_c^B(x)/(d_c^F(x) + d_c^B(x)), & \Omega(x) = 0 \end{cases} \tag{8}$$

Here, $d_c^J(x) = \min_k \|I_t^c(x) - \mu_{k,t}^{c,J}\|$, $I_t^c(x)$ denotes the color data in Lab color space.



(a) (b) (c) (d)

Fig. 3. The human segmentation results. (a) the input frame; (b) the confidence map produced by the pose term; (c) the initial segmentation result; (b) the final refined segmentation result.

Motion term. Pedestrians typically preserve relative motion with the static background scene. Motion cue, which is invariant to illumination changes, seems to be a more natural and robust cue. For building the motion models, we first obtain the motion field [18] by comparing the current frame with the subsequent frame, and then estimate the mean motion values within the foreground and background regions, obtaining the motion model $\Theta_t^m = \{\mu_t^{m,J} | J \in \{B, F\}\}$, in which $\mu_t^{m,J}$ is the mean motion value. The motion term is calculated as:

$$\varphi_m(x) = \begin{cases} d_m^F(x)/(d_m^F(x) + d_m^B(x)), & \Omega(x) = 1 \\ d_m^B(x)/(d_m^F(x) + d_m^B(x)), & \Omega(x) = 0 \end{cases} \tag{9}$$

Here, $d_m^J(x) = \|I_t^m(x) - \mu_t^{m,J}\|$, $I_t^m(x)$ involves the motion vector in pixel x.

Pose term. The pose cue ensures that pixels falling near to the skeleton would more likely be assigned with object label and vice versa. In our case the skeleton is modeled as a puppet of skeleton lines. As shown in Fig. 3(b), we use the distance field along the skeleton to represent the pose term. The pose term takes the form:

$$\varphi_p(x) = \min(\|x - q^*\|/(r_i |L_{q^*}|), 1.0) \tag{10}$$

Here $q^* = \arg\min_{q \in \{L_i\}} \|x - q\|$. $\{L_i | i = 1, \ldots, 6\}$ are skeleton lines, indicating head, torso, two upper legs and two lower legs. $|L_i|$ is line length, $\{r_i\}$ is width/height ratio for skeleton region, empirically set to $\{1.0, 0.5, 0.34, 0.34, 0.3, 0.3\}$ in our experiments.

Segment coherence term. This term is used to maintain temporal coherence of segmentation along the video sequence, which is defined by:

$$\varphi_s(x) = \begin{cases} c_s, & \Omega(x) = \Omega(x') \\ 1 - c_s, & \Omega(x) \neq \Omega(x') \end{cases} \tag{11}$$

where x' is the matched pixel of x in the previous frame, c_s is a constant value(empirically set to 0.3). Note that the coherence term $\varphi_s(x)$ at the first frame is unavailable.

An energy minimization solver - mincut [19] is run to optimize $\Omega^* = \arg\max_\Omega p(\Omega|I_t, \Delta_t, \Theta_t)$ to obtain the refined pedestrian segmentation. Then at E-Step II, we re-estimate the mean values $\Theta_t = \{\mu_{k,t}^{c,J}, \mu_t^{m,J} | k = 1, \ldots, K_c, J \in \{B, F\}\}$ within the segment matte Ω. The re-estimated parameters Θ_t are further used to refine segmentation. In this way, the two complementary steps (M-Step and E-Step II) are repeated several iterations. For further refining the final extracted silhouette, we invoke the Bayesian matting [19] to soft-segment an eroded narrow region along the silhouette boundaries. Fig. 3(c) and Fig. 3(d) show the initial and refined segmentation results respectively.

IV. BAYESIAN FILTERING FOR POSE TRACKING (E-STEP III)

Pose tracking is to sequentially estimate pose states along the video using available observed data and prior knowledge. With the Markov properties of human dynamic, pose tracking is often formulated as a Bayesian filtering problem:

$$p(\Delta_t|I_{1:t}) \propto p(I_t|\Delta_t)\int p(\Delta_t|\Delta_{t-1})p(\Delta_{t-1}|I_{1:t-1})d\Delta_{t-1} \tag{12}$$

where $p(I_t|\Delta_t)$ is the observation likelihood and $p(\Delta_t|\Delta_{t-1})$ is the prior. Particle filter is mostly implemented to approximate the complicated infinitesimal calculus in the above formula with a set of weighted particles. In particle filter, the posterior is generated in three steps: 1) sample particles $\Delta_t^{(i)} \propto p(\Delta_t|\Delta_{t-1}^{(i)})$; 2) adjust weights $\pi_t^{(i)} = \pi_{t-1}^{(i)}p(I_t|\Delta_t^{(i)})$; 3) normalize $\pi_t^{(i)}$ to make sure $\sum_i \pi_t^{(i)} = 1$.

Various methods have been presented for pose tracking using the above general inference procedure. They differ in the definitions of the pose state, or the observation likelihood, or the dynamic prior. In our work, we build the prior mainly on the bipedal walking motion [21], while presenting a novel observation likelihood in conjunction with a physical-based pose representation.

A. Pose Representation

This physical-based pose is 2.5D, modeled with 6 rigid body regions, including head, torso, upper/lower stance/swing legs, and parameterized with $\Delta = \{\Delta_f, \Delta_v\}$. $\Delta_f = \{sl, rp_1, rp_2, lu, ll\}$, as a fixed model, is set according to the skeleton at the first frame, in which sl indicates the walking step length, rp_1 and rp_2 are the relative positions of the head and neck joints with respect to the body center, lu and ll are the upper and lower leg lengths. Δ_v is a variation model, $\Delta_v = \{v, \theta_{hb}, \theta_{ut}, \theta_{uw}, d\theta_{ut}, d\theta_{uw}, \theta_{lt}, \theta_{lw}\}$. It is used to simulate human walking, in which v denotes the walking speed, θ_{hb} is the turning angle of the body, θ_{ut} and θ_{uw} are the angles for the upper stance and swing legs respectively, $d\theta_{ut}$ and $d\theta_{uw}$ are the corresponding angular velocities for θ_{ut} and θ_{uw} respectively, θ_{lt} and θ_{lw} are the angles for the lower stance and swing legs.

Using the variation model Δ_v, we build a dynamic process to simulate human walking, i.e., sampling particle state $\Delta_t^{(i)}$ based on prior $p(\Delta_t|\Delta_{t-1}^{(i)})$. We build the prior mainly on the 2D physical formulations [21], in which v and θ_{hb} both follow the normal distributions, that is, $v_t \sim N(v_{t-1}, \sigma_v)$, $\theta_{hb,t} \sim N(\theta_{hb,t-1}, \sigma_{hb})$. $(\theta_{ut}, \theta_{uw}, d\theta_{ut}, d\theta_{uw}, \theta_{lt})$ are induced by sl and v according to the physical Motion Laws [21]. The lower swing leg angle θ_{lw} is initialized by the skeleton at the first frame and modeled as:

$$\theta_{lw,t} \sim N(\theta_{lw,t-1} + \varepsilon(\theta_{lw,t-1} - \theta_{lw,t-2}), \sigma_{\theta_{wl}}) \tag{13}$$

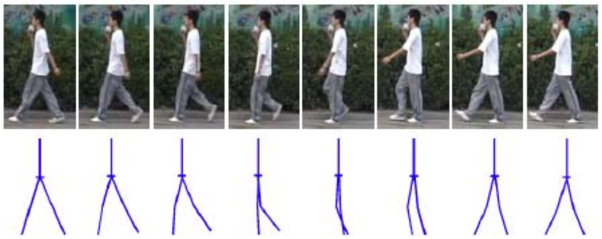


Fig. 4. The pose tracking results. The top row shows the input frames, and the second row demonstrates the corresponding poses.

Here, we use $\sigma_v = 7.0$, $\sigma_{hb} = 3.0$, $\varepsilon = 0.3$, $\sigma_{\theta_{wl}} = 8.0$.

B. Observation likelihood

The observation likelihoods are derived with multiple cues, including color cue I_t^c, motion cue I_t^m and silhouette cue I_t^s. We build an independent likelihood for each cue and combine all likelihoods together to form the final likelihood:

$$p(I_t|\Delta_t) = w_c p(I_t^c|\Delta_t) + w_m p(I_t^m|\Delta_t) + w_s p(I_t^s|\Delta_t) \tag{14}$$

Here, w_c, w_m and w_s are weighting values. We describe the likelihoods according to the cue type in the following.

Color likelihood. The color likelihood is evaluated with a stable component and a wandering component. The likelihood for a new observation conditioned on the previous observation is formulated by:

$$p(I_t^c|\Delta_t) = \lambda_s \exp(-\chi^2(I_t^c, I_1^c)) + \lambda_w \exp(-\chi^2(I_t^c, I_{t-1}^c)) \tag{15}$$

where $\chi^2(\cdot,\cdot)$ is the χ^2 distance, I_1^c and I_{t-1}^c are the color histograms of the leg regions which are determined by Δ_1 and Δ_{t-1} respectively.

Motion likelihood. The motion likelihood is built on the motion field obtained by [16], which can provide the motion information about the tracked limbs between two successive frames. The likelihood of the motion cue is given by the mean square distance (MSD) of the projected positions $\{ps_i\}$ and hypothesized position $\{hs_i\}$ for a set of sample points:

$$p(I_t^m|\Delta_t) \propto \exp(-\sum_i \|ps_i - hs_i\|/\tilde{N}) \tag{16}$$

Here $\tilde{N}$ is the number of sample points.

Silhouette likelihood. Silhouette is a binary map indicating human foreground region, which is derived from the projection of previous obtained silhouette Ω_{t-1} (see Section III) with optical flow [16]. The negative likelihood for silhouette cue is calculated as the mean square error (MSE) of the predicted values $\{ss_i\}$ and the observed values $\{bs_i\}$ for a set of sample points inside the limb region.

$$p(I^s|\Delta_t) \propto \exp(-\sum_i \|ss_i - bs_i\|/\tilde{N}) \tag{17}$$

Here $\tilde{N}$ is the number of sample points.

Based on the above definitions, we use particle filter to sequentially estimate pose states. Fig. 4 demonstrates the tracking results for a walking cycle.

TABLE I

THE EMPIRICAL VALUES SET FOR THE INVOLVED PARAMETERS.

Formula No.	Parameter Name	Parameter Value
(7)	$\{\lambda_c, \lambda_m, \lambda_p, \lambda_s\}$	{0.3,0.3,0.2,0.2}
(14)	$\{w_c, w_m, w_s\}$	{0.3,0.3,0.4}
(15)	$\{\lambda_s, \lambda_w\}$	{0.8,0.2}

V. EXPERIMENTAL RESULTS

Experimental setting: The proposed framework is implemented in a personal computer with a 2.26 GHz CPU and 3 GB RAM. We made experiments on Ethz dataset [23] (5 sequences), Weizmann dataset [24] (15 sequences), and several sequences we captured using a hand-hold camera. In those sequences, all human windows have been resized to 320 $\times$ 240. We use the Ethz dataset as the template set in the initialization stage. For the weighting parameters involved in this framework, we empirically set them in Table I, in which Formula No. refers to where the parameters exist.

Quantitative evaluation. For quantitative evaluation, we obtain the segmentation accuracy in form of F-$measure$[1]. F-$measure = 2 \times precision \times recall/(precision + recall)$, $precision$ is the ratio of the true positive pixels (i.e., the pixels labeled as foreground actually belong to foreground) to the all labeled foreground pixels, and $recall$ is the ratio of the true positive pixels to the ground truth pixels. The pose accuracy is estimated by the Mean Square Distance (MSD) between the lower body joints and the corresponding hand-marked joints.

To verify the influence of pose cue in the segmentation stage, we measure the F-measures with and without pose information on an Ethz sequence. The corresponding F-measure results are illustrated in Fig. 5(a), which shows that the segmentation accuracies with pose cue are obviously higher than the results without pose cue. Meanwhile, we estimate the influence of silhouette cue in the pose tracking stage (E-Step II). The comparison results on the same sequence are shown in Fig. 5(b), which verifies that the pose can be estimated more precisely with silhouette cue.

We also compare the segmentation accuracy with the template matching method [8] for Ethz dataset [23] and Weizmann dataset [24]. The template matching method [8] achieves an average F-measure of 82.2% for Ethz dataset, and 76.5% for Weizmann dataset. Using our algorithm, the accuracies are improved to 89.6% and 88.1% respectively.

Qualitative evaluation. We demonstrate the segmentation results for four sequences in Fig. 6. The front image of each sequence is the first input frame, indicating the background environment where the pedestrian locates. For saving space, we omit the remaining frames. Following up is the segmentation results for the input sequence. As can be seen, although the human poses in the walking cycle are continuous varying, the coherent optimization of our framework ensures the accurate segmentation across time. We attribute this to the utilization of pose information and pixel correspondence (i.e., motion field) along the sequence.

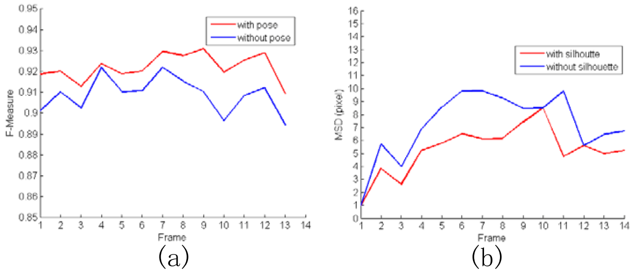


Fig. 5. (a) The segmentation accuracies (F-measure) with and without pose cue. (b) The pose tracking errors (MSD) with and without silhouette cue.

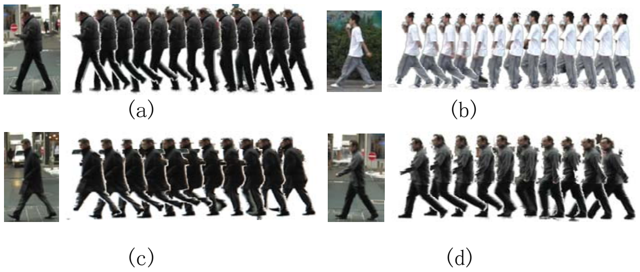


Fig. 6. Segmentation results for four pedestrian sequences.

In order to further demonstrate the performance of our framework, we compare our segmentation results with those obtained by Grabcut [25] and template matching[8]. Grabcut is an interactive MRF-based object cutout method, which commonly requires users to indicate figure and ground regions. Fig. 7(b) demonstrates some figure and ground scribbles we drew on the frames, and Fig. 7(c) shows the corresponding results obtained by Grabcut. Obviously, Grabcut requires cumbersome interactions, and its results are sensitive to the interactions. Comparably, our framework can automatically extract human silhouettes (see Fig. 7(f)) based on the inferred pose (see Fig. 7(e)). Template matching [8] is also an automatic foreground segmentation method, yet the segmentation results are sensitive to local variations since it does not consider the local appearance. As can be seen in Fig. 7(d), the head regions, the hip regions are inaccurately segmented by template matching. The average F-measures for Fig. 7(c)(d)(f) are 0.86, 0.64 and 0.91 respectively.

Time cost. The implementation with Matlab takes about 30 seconds per frame. One third of that time is taken by body segmentation, and two thirds are taken by pose tracking. The reasons for the expensive computation are two folds. First, in the MRF-based segmentation, the adopted mincut solution [19] involves time-consuming α-β swapping and α-expending. Second, in pose tracking stage, the employed particle filter requires to predict many particles' observation likelihoods.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper we address the problem of simultaneous pedestrian segmentation and pose estimation from natural

[1] http://www.dcs.gla.ac.uk/Keith/Preface.html

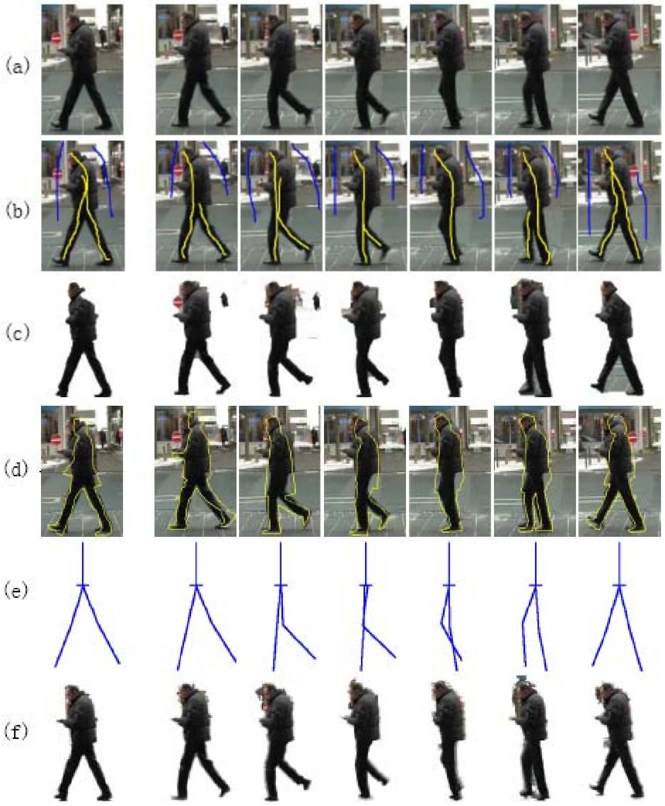


Fig. 7. Comparison results with Grabcut [25] and template matching [8]. (a) the input frames; (b) the foreground (yellow) and background (blue) scribbles drew on the frames; (c) the results obtained by Grabcut based on the scribbles in (b); (d) the silhouettes obtained by template matching are overlaid on the frames; (e) the poses inferred by our framework; (f) the extracted pedestrians by our framework based on the poses in (e).

videos. The problem is formulated in a Bayesian framework in which the two tasks interact closely to provide loop feedbacks for each other to improve estimation quality. As the experiments have shown, our framework can automatically achieve promising results for each walking cycle. We feel this is due to the combination of the high accuracy of MRF with the robustness of physical-based Bayesian pose estimation. The major limitation of the framework is the computation time. To develop a real-time dynamic system, we plan to adopt more efficient optimization solutions or re-design some processes for implementation in parallel graphics hardware. In addition, the framework is currently sensitive to occlusion since the observation cues(color and motion) are unreliable in this case, thus another possible direction is to extend the framework to deal with partial occlusions.

References

[1] Y. Li, J. Sun and H. Y. Shum, “Video Object Cut and Paste”, *ACM Transactions on Graphics*, Vol. 24, 2005, pp. 595-600.

[2] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala and M. F. Cohen, “Interactive Video Cutout”, *ACM Transactions on Graphics*, Vol. 24, 2005, pp. 585-594.

[3] J. Malcolm, Y. Rathi and A. Tannenbaum, “Multi-Object Tracking Through Clutter Using Graph Cuts”, *in Proceedings of IEEE International Conference on Computer Vision*, 2007, pp. 1-5.

[4] C. Wang, M. de La Gorce and N. Paragios, “Segmentation, Ordering and Multi-Object Tracking using Graphical Models”, *in Proceedings of IEEE International Conference on Computer Vision*, 2009, pp.747-754

[5] J. C. Niebles, B. Han and L. Fei-Fei, “Efficient Extraction of Human Motion Volumes by Tracking”, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 655-662.

[6] D. Mitzel, E. Horbert, A. Ess and B. Leibe, “Multi-Person Tracking with Sparse Detection and Continuous Segmentation”, *in Proceedings of European Conference on Computer Vision*, Vol. 1, 2010, pp. 397-410.

[7] Y. Li, Z. Zhong, W. Wu, “Combining Shape and Appearance for Automatic Pedestrian Segmentation”, *in Proceedings of The IEEE International Conference on Tools with Artificial Intelligence*, 2011, pp. 369-376.

[8] D. M. Gavrila, “A Exemplar-Based Approach to Hierarchical Shape Matching”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, 2007, pp. 1408-1421.

[9] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr and R. Cipolla, “Multivariate Relevance Vector Machines for Tracking”, *in Proceedings of European Conference on Computer Vision*, Vol. 3, 2006, pp. 124-136.

[10] Z. Lin, L. S. Davis, D. Doermann and D. DeMenthon, “An Interactive Approach to Pose-Assisted and Appearance-based Segmentation of Humans”, *in Proceedings of IEEE International Conference on Computer Vision*, 2007.

[11] M. Bray, P. Kohli and P. Torr, “PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans using Dynamic Graph-Cuts”, *in Proceedings of European Conference on Computer Vision*, Vol. 2, 2006, pp. 642-655.

[12] M. P. Kumar, P. Torr and A. Zisserman, “OBJ CUT”, *in Proceedings of IEEE International Conference on Computer Vision*, 2005, pp. 18-25.

[13] J. Deutscher and I. Reid, “Articulated body motion capture by stochastic search”, *International Journal of Computer Vision*, Vol. 61, 2010, pp. 185-205.

[14] S. Brutzer, B. Hoferlin and G. Heidemann, “Evaluation of Background Subtraction Techniques for Video Surveillance”,*in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1937-1944.

[15] H. Grabner and H. Bischof, “On-line Boosting and Vision”, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2006, pp. 260-267.

[16] D. Martin, C. Fowlkes and J. Malik, “Learning to detect natural image boundaries using brightness and texture”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.26, 2000, pp. 530-549.

[17] S. Z. Li, “Markov Random Field Modeling in Image Analysis”, *Third Edition, New York: Springer-Verlag*, 2009, Chapter 2.

[18] D. Sun, S. Roth and M. J. Black. “Secrets of optical flow estimation and their principles”, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2432-2439.

[19] Y. Boykov. O. Veksler and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 1222-1239.

[20] Y. Y. Chuang, B. Curless, D. Salesin and R. Szeliski, “A Bayesian approach to Digital Matting”, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2001, pp. 264-271.

[21] M. A. Brubaker, D. J. Fleet and A. Hertzmann, “Physics-based person tracking using the Anthropomorphic Walker”, *International Journal of Computer Vision*, Vol. 87, 2010, pp. 140-155.

[22] A. D. Jepson, D. J. Fleet and T. F. El-Maraghi, “Robust Online Appearance Models for Visual Tracking”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, 2003, pp. 1296-1311.

[23] M. Andriluka, S. Roth, B. Schiele. “People-Tracking-by-Detection and People-Detection-by-Tracking”, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[24] L.Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, “Actions as Space-Time Shapes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, 2007, pp. 2247-2253.

[25] C. Rother, V. Kolmogorov and A. Blake, “Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts”, *ACM Transactions on Graphics*, Vol. 23, 2004, pp. 309-314.