Human Body Segmentation via Data-Driven Graph Cut

Shifeng Li, Huchuan Lu, *Senior Member, IEEE,* and Xingqing Shao

***Abstract*—Human body segmentation is a challenging and important problem in computer vision. Existing methods usually entail a time-consuming training phase for prior knowledge learning with complex shape matching for body segmentation. In this paper, we propose a data-driven method that integrates top-down body pose information and bottom-up low-level visual cues for segmenting humans in static images within the graph cut framework. The key idea of our approach is first to exploit human kinematics to search for body part candidates via dynamic programming for high-level evidence. Then, by using the body parts classifiers, obtaining bottom-up cues of human body distribution for low-level evidence. All the evidence collected from top-down and bottom-up procedures are integrated in a graph cut framework for human body segmentation. Qualitative and quantitative experiment results demonstrate the merits of the proposed method in segmenting human bodies with arbitrary poses from cluttered backgrounds.**

***Index Terms*—Color-based boosting algorithm, human body segmentation, top-down information.**

I. Introduction

IN COMPUTER vision, human detection, pose estimation and segmentation are intertwined problems that can provide cues for each of them. Assume that a person is detected and segmented in an image, pose estimation can be carried out efficiently [1]–[3]. Conversely, a person can also be easily segmented from an image when the pose information is available [4], [5]. Numerous methods have demonstrated that effective figure-ground segmentation can be achieved via the learned shape prior knowledge [6]–[11]. Better results can be obtained via integrating top-down with bottom-up cues for segmenting objects with limited pose variations (e.g., horses and cows) [12]–[17]. However, such shape prior may not be easily inferred for segmenting human bodies in arbitrary poses. In addition, shape matching in high-dimensional pose space remains a challenging problem.

Manuscript received May 13, 2013; revised October 22, 2013; accepted January 7, 2014. Date of publication January 31, 2014; date of current version October 13, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61071209 and Grant 61272372, and in part by the joint Foundation of China Education Ministry and China Mobile Communication Corporation under Grant MCM20122071. This paper was recommended by Associate Editor M. Shin.

The authors are with the School of Information and Communication Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: lhchuan@dlut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2301193

Object segmentation without any prior knowledge is a well-known ill-posed problem as multiple equally plausible solutions often exist. Object-centered clustering approaches group features with learned parametric or nonparametric densities, e.g., k-means clustering and EM algorithms [18]. Pairwise potential-based methods perform figure-ground discrimination by gathering features with minimum energy cost, e.g., normalized cut [19]. Interactive segmentation algorithms combine pairwise potential constraints and object-centered appearance representation in a unified energy minimization paradigm [17], [20], [21]. These algorithms are initialized by some manually labeled pixels (i.e., seeds) of the foreground object and background, thereby imposing hard constraints in the graph cut framework. Recently, numerous methods have been proposed to integrate global shape priors and top-down inference for unsupervised segmentation. In [6] and [8], the top-down shape cues are merged with bottom-up over-segmentation results to locate the specific objects with limited pose variations (e.g., horses and cows). An interactive segmentation approach that incorporates local Markov random field (MRF) constraint and global shape prior to iteratively estimate segmentation and pose simultaneously is developed by [7]. In [9], a formulation using pose-specific conditional random fields (CRF) and stick figures is proposed for segmentation and pose estimation of humans within a Bayesian framework. Notwithstanding the demonstrated success on some datasets collected in well-controlled environments, it is not clear whether this method performs well in cluttered backgrounds.

In this paper, we propose an algorithm for segmenting human body in cluttered backgrounds by integrating bottom-up and top-down cues without user interaction. We integrate the pose derived from top-down inference with the probabilistic estimate of body parts obtained by bottom-up cues using adaptive classifiers. A human body is segmented using an efficient graph cut algorithm with data driven by pose evidence. Meanwhile, we present a 2-D pose estimation scheme via searching for locally optimal configurations by combining top-down inference and bottom-up cues. Fig. 1 shows the flowchart of our method. Experimental results demonstrate the effectiveness of the proposed algorithm for segmenting human body with numerous pose variations in cluttered backgrounds.

II. Related Work

Foreground/background segmentation from static images has been an important problem in computer vision. Here, we discuss the most relevant work in the context of

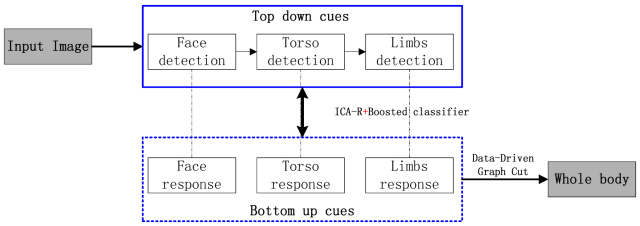


Fig. 1. Flowchart of our method.

object-centered segmentation (e. g., cows, horses, and human bodies) in the following categories.

Bottom-Up Segmentation: The bottom-up methods [18]–[20] segment an image into homogeneous regions and then combine the necessary regions into the object. This kind of approach groups pixels according to their feature property, as well as smoothness and continuity of boundary. The main difficulty is that an object may be segmented into multiple regions, and some of which may be merged into background. For human body segmentation, multiple regions of body parts, such as head, torso and legs, are often found in images for body segmentation. Without additional constraints, it is not clear whether these regions can be grouped into human-like segmentation, especially when the object appears in cluttered backgrounds or has a variety of poses. In our method, we identify lower body by exploiting low level cues using two effective classifiers.

Top-Down Segmentation: The top-down algorithms apply the learned properties about an object, such as shape, color, or texture to guide the segmentation [9], [22]–[24]. The main challenge for this kind of approach is to account for a large variability of shape and appearance of a given object class. Consequently, the segmentation results may not accurately delineate contours of human body. Furthermore, it is rather challenging to learn human body shapes that encompass a large variety of poses. Consequently, much success has been demonstrated in the context of object segmentation with limited poses and shape variations. To enforce the flexibility and reduce the complexity of top-down knowledge, a novel torso model is proposed for torso segmentation.

Top-Down and Bottom-Up Segmentation: In this formulation, deformable templates are often combined with segmentation results obtained from bottom-up methods. That is, over-segmentation results from bottom-up methods are grouped into homogeneous regions to achieve best a matching against templates [8]. However, it is difficult to extend this kind of approach for human body segmentation as it needs to handle high-dimensional pose state space and a large number of shape templates [7], [9], [25]. On the other hand, the combination of top-down and bottom-up approach has been applied to pose estimation [15], [26], [27]. Although, the pictorial structure method [1] can be applied with bottom-up visual cues to infer human poses and in turn used for segmentation, it requires much visibility of the body parts. Consequently, any successful extension with top-down and bottom-up segmentation needs to address the problems of requiring a large number of shape templates, determining shape similarity, and segmenting an input image into homogeneous regions for the images including people. Instead of using a large number of shape templates, based on the human kinematic model, our method introduces a data-driven strategy for efficiency.

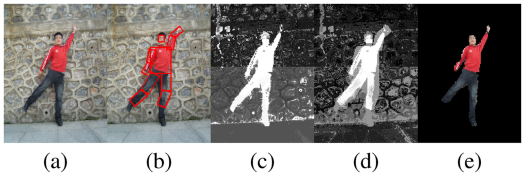


Fig. 2. (a) Input image. (b) Top-down cues. (c) Bottom-up cues. (d) Pose evidence. (e) Our results. Given a still image (a), the human pose (b) is first obtained using top-down inference with dynamic programming. Local classifiers specific to the subject are trained with low-level vision to calculate a body distribution (c). The top-down and bottom-up cues are integrated as evidence of human pose (d), thereby facilitating object segmentation via data-driven graph cut with refined results (e).

Fig. 3. Top-down body pose estimation. (a) Face detection. (b) Torso estimation. (c) and (d) Bottom-up cues searching for limbs. (e) Body pose.

III. Proposed Algorithm

The aforementioned challenges are addressed in our approach. First, body parts are localized via torso estimation and limb searching by a kinematic model. Second, we integrate top-down pose information and bottom-up visual cues to collect prominent pose evidence for human body. Finally, the pose evidence is embedded into the graph cut framework [20] for globally optimal human body segmentation. Fig. 2 shows the overview of our proposed algorithm.

IV. Top-Down Inference for Body Pose

In this section, we present the details of top-down inference process for estimating body pose, which is crucial for the proposed algorithm. Fig. 3 shows the main steps in this process. Given an image I, we first use a multiview face detection method [28] to locate the human face and then estimate the torso using normalized cut [19] with rectangle fitting. This step also provides cues to locate lower body, from which the legs are identified by the low level cues. According to the body topology and human model, limb candidates are derived from Monte Carlo sampling. With the bottom-up cues that indicate the probabilistic estimates of pixels being body parts using boosted classifier and ICA-R local learning, part candidates are locally refined via solving the MAP problem under dynamic programming.

A. Problem Formulation

As shown in Fig. 4(a), the human body is modeled as a kinematic tree similar to that in [1], represented by a graph: $G' = (V', E')$ where the nodes V' denote the body parts, and

the edges E' describe the corresponding kinematic constraints. There are ten parts in the model and their pose is denoted by $\chi = \{X_1, X_2, ..., X_{10}\}$. Each part is represented as a rectangle with five degrees of freedom, i.e., $X_k = \{x_k, y_k, \theta_k, h_k, w_k\}$ [position (x, y), orientation θ, and scale h, w] as shown in Fig. 4(c). Hence, the model has 50 degrees of freedom, which makes pose estimation rather challenging.

Given an image observation I_e, pose estimation can be formulated as a Bayesian inference problem

$$P(\chi|I_e) \propto P(I_e|\chi)P(\chi) \tag{1}$$

where $P(I_e|\chi)$ is the likelihood of observations given a kind of human pose χ, and $P(\chi)$ is the prior distribution that enforces the constraints between body parts. A maximum a posteriori (MAP) estimation is given by

$$\chi^* = \arg\max_{\chi} P(\chi|I_e) \tag{2}$$

that can be solved by numerous methods [1], [2], [26], [27], [29].

B. Torso Estimation

As human head and torso are arguably the most salient body parts for segmentation, we use a head-torso model to guide the search process. This model consists of two rectangles with one representing the face and the other denoting the directed torso (from $-45°$ to $45°$ with the quantized interval of $15°$ in-plane rotation with respect to the head), just as shown in Fig. 5(b).

In our model, the face is first obtained by the method [28] and then the torso can be estimated with region, location, and boundary cues, which is more robust than the method proposed in [30] that also proposed a torso model while using only the color information as feature. We estimate the region probability term based on the superpixels produced by normalized cut [19]. Assume that there are M directed head-torso candidates and each is represented by $R_m, m = 1, 2, ..., M$. For a torso rectangle, it covers L superpixels. The region probability of P_m^R is defined as

$$P_m^R = \sum_l \frac{O_{ml}}{S_{ml}} \left(\frac{O_{ml}}{S_{Rect}^m} \right)^{\xi} \tag{3}$$

where O_{ml} is the overlapped area between the l-th superpxiel and the m-th torso rectangle of head-torso model, S_{ml} is the area of the l-th superpixel under the m-th torso rectangle of the model, S_m^{Rect} is the area covered by m-th torso rectangle. The parameter ξ is often set to weigh more on the first term (*e. g.* $\xi = 0.5$). This probability estimates how each superpixel fits the head-torso model. The location probability describes the likelihood of each pixel belonging to the torso rectangle. The superpixels directly below the face rectangle are likely to be the model that is

$$P_m^L = \sum_{(x,y)\in S_{ml}} \exp\left(-\left|\frac{\rho_{x,y}^{ml}}{C^m}\right|^{\zeta}\right) \tag{4}$$

where $\rho_{x,y}^{ml}$ is the Euclidean distance from one pixel (x, y) in the l-th superpixel covered by the m-th torso rectangle to the midpoint of the bottom edge of the face rectangle, C^m is the

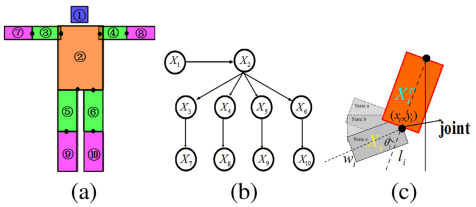


Fig. 4. Human model. (a) Pictorial structure model. (b) Kinematic tree with x_1 representing the face, x_2 standing for the torso and x_3 to x_{10} for the limbs. (c) Limb searching states.

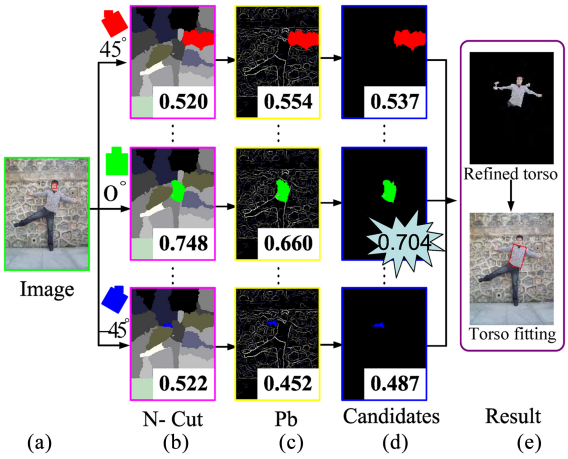


Fig. 5. Torso estimation. (a) Input image with detected face. (b) Scores of head-torso matching with P^{RL}. (c) Scores of boundary matching with P^C. (d) Scores of the coarse torso candidates. (e) Refined torso segmentation.

width of the m-th torso rectangle, and $\zeta = 4$ is a weighting factor. The region probability P_m^R and the location probability P_m^L are combined by

$$P_m^{RL} = \tau P_m^R + (1 - \tau) P_m^L \tag{5}$$

where $\tau = 0.5$ is the weighting term.

The boundary information is also used for more precise torso estimation with the contour detector [31], which predicts the posterior probability $P_b(x, y)$ of a boundary at each pixel(x, y). For all the pixels on the boundary of the torso model R_m, we compute its contour probability

$$P_m^C = \frac{1}{N_t} \sum_{n=1}^{N_t} P_b^n(x, y) \tag{6}$$

where N_t and n denote the total number and index of pixels (x, y) on the boundary of R_m, respectively. All the probabilistic estimates are integrated for torso estimation

$$P_m^{torso} = \nu P_m^{RL} + (1 - \nu) P_m^C \tag{7}$$

where ν is the weighting term (e.g., $\nu = 0.5$), and the most likely region is selected as torso. Once the most likely torso region is determined, the corresponding foreground and background can be used as the constraints in the graph cut algorithm [20] for a coarse torso segmentation. Then, we apply the torso detection once again based on the coarse segmented torso to obtain an accurate torso estimation. The main steps of torso estimation are presented in Fig. 5.

C. Limb Searching

The pose prior, $P(\chi)$, in (1) imposes location (x, y), orientation θ, length h, and width w constraints of body parts. As in [1], let (x_k^p, y_k^p) denote the joint position between the current part k and the parent part p [Fig. 4(b)]. h_k is the height of the current part k, h_k^p is the height of the parent part p. w_k is the width of the current part k, and w_k^p is the width of the parent part p. The prior distribution $P(X_k|X_k^p)$ of part k is

$$\begin{aligned} P(X_k|X_k^p) \;=\; & N(x_k - x_k^p; 0, \sigma_x^2) N(y_k - y_k^p; 0, \sigma_y^2) \\ & N(h_k - h_k^p; 0, \sigma_h^2) N(w_k - w_k^p; 0, \sigma_w^2) M(\theta_k - \theta_\mu; 0, \eta) \end{aligned} \tag{8}$$

where $N(\cdot)$ is a Gaussian distribution with zero mean and standard deviation σ learned from training examples, and the angle θ is modeled by a von Mise distribution [1] with a parameter η. Then $p(\chi)$ can be calculated by the detected face and torso as follows:

$$P(\chi) = P(X_{head}) P(X_{torso}) \prod_{k \notin \{head, torso\}} P(X_k|X_k^p). \tag{9}$$

To compute the likelihood $P(I_e|\chi)$ of (1), we consider three mid-level cues: 1) body distribution $P_1(I_e|\chi)$ obtained by a boosted classifier (Section V-A); 2) body likelihood $P_2(I_e|\chi)$ derived from ICA-R local learning (Section V-B); and 3) color consistency $P_3(I_e|\chi)$ acquired via χ^2-distance

$$\begin{aligned} p(I_e|\chi) = & \; p_1(I_e|\chi)^\alpha p_2(I_e|\chi)^\beta p_3(I_e|\chi)^\gamma = \\ & \prod_{k \notin \{head, torso\}} p_1(I_e|X_k)^\alpha p_2(I_e|X_k)^\beta p_3(I_e|X_k)^\gamma \end{aligned} \tag{10}$$

where α, β and γ represent the importance factors of these cues (e.g., $\alpha = 0.8$, $\beta = 1$, and $\gamma = 0.9$ in our experiments). The body likelihoods p_1 and p_2 are computed by local evidence y' derived from boosted classifier and y obtained from the ICA-R local learning (described in Section V). Hence, the first two terms of (10) are computed by

$$P_1(I_e|X_k) = \frac{1}{N_l} \sum_{n=1}^{N_l} y'(n), \quad P_2(I_e|X_k) = \frac{1}{N_l} \sum_{n=1}^{N_l} y(n) \tag{11}$$

where n and N_l represent the index and the total number of the pixels covered by one body part respectively. For P_3 of (10), we compute the normalized χ^2-distance between all the current part candidates and its parent as well as grandparent. Take the lower arms (la) for example, the color of lower arms may be similar to the face skin color (if short sleeves) or with upper arms (ua) (long sleeves or no sleeves). We first compute the normalized χ^2-distance $P(la, ua)$ and $P(la, face)$, respectively, and then integrate them for likelihood estimate using $P_3(I_e|X_k) = \max(P(la, ua), P(la, face))$.

Hence, the likelihood for each limb can be computed with (10). We can draw a few body part candidates using (1) and the best configuration can be determined with the MAP estimate (2) using dynamic programming [1].

V. Bottom-Up Cues for Body Parts

As mentioned above, the seeds are crucial for solving (19) using graph cut algorithm. On the other hand, the top-down inference of body needs some low-level visual cues to choose the best candidates of human body. For the human body, there usually exist multiple regions that appear to be like face, torso, and legs. Consequently, the locations of these parts may not be modeled well with a simple distribution (e.g., Gaussian). In addition, for the complex poses, we need more information to describe the human body.

In this section, we describe two schemes to compute the body parts distribution: a boosted classifier and a ICA-R local learning, which are considered to be bottom-up cues for top-down pose inference. Meanwhile, we use the response image derived from the boosted classifier as bottom-up cues to segment human body combined with the pose.

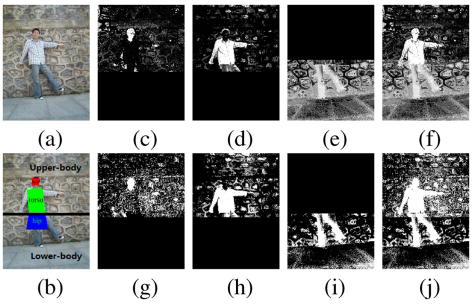


Fig. 6. Locating body parts with bottom-up cues. (a) Input image. (b) Regions for training to obtain bottom-up cues. (c), (d), and (e) Response maps indicating the probabilistic estimates of face, torso, and legs from a boosted classifier. (f) Body-part response map by integrating (c), (d), and (e). (g), (h), and (i) Response maps describing likelihoods of face, torso, and legs from ICA-R local learning. (j) Body-part likelihood map combining (f), (g), and (h).

Algorithm 1 *Boosted classifiers for body parts*

Input: 3 determined regions: B = {face, torso, hip}, and each part RGB feature $\{x_i', y_i'\}_{i=1}^N$, where $x_i' = (x_r', x_g', x_b')$, $y_i' \in \{-1, 1\}$.

1: for B = face, torso, hip
2: Construct all K weak learners
3: Initialize $\omega_1 = 0$, pixel weights $D_1(i) = 1/N$.
4: for t=1 to T do
5: $h_t = \arg\min_{h_{\kappa_1} \in h} \varepsilon_{\kappa_1} = \sum_{i=1}^{M} D_t(i)[y_i' \neq h_{\kappa_1}(x_i')]$
6: Compute u_t and $e_{t,i}$, where $i = \arg\min_{\kappa_1} \varepsilon_{\kappa_1}$
7: Set $\alpha'_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$
8: Update $D_{t+1}(i) = \frac{D_t}{Z_t} e^{-\alpha'_t y_i' h_t(x_i')}$
9: $\omega_1^t \leftarrow \omega_1^{t-1} + \alpha'_t u_t e_{t,i}$
10: end for
11: $y' = \omega_1^\top x$
12: end for

Output: $y' = $ combine $\{y'_{face}, y'_{torso}, y'_{hip}\}$ for a probabilistic map of upper-body parts

A. Boosted Classifier for Body Parts

We introduce a boosted classifier for estimating body parts distribution using low-level cues. Given a set of sample features (e.g., RGB pixel values $x_i' = (r, g, b)$ obtained from the estimated face, torso, and hip), the task is to seek a linear projection $y' = \omega_1^\top x'$ that maximizes the separation between foreground and background (i.e., Fisher criterion). In [32], a

simple model is proposed by defining $\omega_1 \propto (m_1 - m_2)$ where m_1 and m_2 are the sample means of the background and foreground classes, respectively. Though this model reduces the computational complexity, the covariance of both foreground and background distributions cannot be modeled well by the isotropic Gaussian distribution. For this, we propose a boosted classifier to detect body parts.

The boosted classifier is trained via the Adaboost algorithm [33] where weak learners $h(x')$ are added and weighted for a strong classifier $H(x') = sign(\sum_{t=1}^{T} \alpha'_t h_t(x'))$. Since each weak learner is a linear classifier, we have

$$h_t(x') = u_t(q_{t,i}^{\top} x' + b_t) \tag{12}$$

where $q_{t,i}^{\top}$ is the weight of the linear classifier. All elements of $q_{t,i}^{\top}$ are 0 except that the i-th element $q_i = 1$, b_t is a bias term, and u_t is a polarity (-1 or 1). The boosted classifier $H(x')$ is written as

$$H(x') = sign[(\sum_{t=1}^{T} \alpha'_t u_t q_{t,i})^{\top} x' + \sum_{t=1}^{T} \alpha'_t u_t b_t]. \tag{13}$$

Let $\omega_1 = \sum_{t=1}^{T} \alpha'_t u_t q_{t,i}$ and $b = \sum_{t=1}^{T} \alpha'_t u_t b_t$, then

$$H(x') = sign(\omega_1^{\top} x' + b) \tag{14}$$

which means that $H(x')$ is a hyperplane that separates positive and negative samples. For a soft segmentation, we only need α'_t, u_t and $q_{t,i}$ for computing ω_1. After the boosted selection of dimension of feature vectors, we obtain a hyperplane ω_1 with most discriminative power.

The weak learner we use is given by

$$h_{\kappa_1}(x') = sign[p^+(y' = 1|x') - p^-(y' = -1|x')] \tag{15}$$

where p^+ and p^- are modeled by Gaussian distributions for positive and negative classes for each test image. For example, after locating the torso via torso estimation, we collect positive samples in the torso region, and negative examples from all the remaining areas of upper-body for calculating the expectation and deviation of Gaussian distributions of weak classifiers in (15). When obtaining ω_1 from the boosted classifier, we can compute $y' = \omega_1^{\top} x'$ to get the soft values in the homogeneous region. Thus, repeating the boosting training from face, torso to hip [red, green, and blue regions in Fig. 6(b)], we can obtain ω_1^{face}, ω_1^{torso}, ω_1^{hip}, and the probabilistic estimates of body part: $\{y'_{face}, y'_{torso}, y'_{hip}\}$ [see Fig. 6(c)–(e)]. All the probabilistic maps of parts are integrated by choosing the maximum at each pixel, hence rendering a estimate describing the human distribution y' [see Fig. 6(f)], which is crucial in calculating P_1 in (11) and pose-evidence. The steps of boosted local learning model for body distribution are summarized in Algorithm 1, where the bias b is omitted.

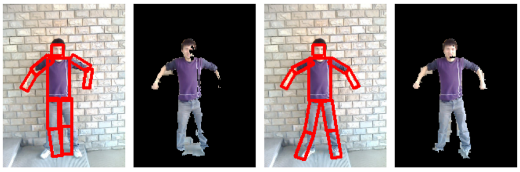
B. ICA-R for Body Parts

As shown in Fig. 7(a) and (b), a worse body segmentation is obtained only using the boosted classifier due to the similar color between lower body and background. To get more information from the low-level vision for precise human body segmentation, we use the one-unit ICA-R algorithm [34], [35], which is capable of extracting an expected information by using the prior knowledge. For an input color image x and the reference image r, ICA-R seeks a projection space $y = \omega_2^T x$ to extract the desired evidence similar to r. The problem can be written as

$$\begin{aligned} &\max\ J(y) \approx \rho[E\{G(y)\} - E\{G(y')\}]^2 \\ &st.\quad g(\omega_2) \leq 0 \end{aligned} \tag{16}$$

where ρ is a positive constant, y' is the Gaussian variable with zero mean and unit deviation, and $G(z) = e^{\frac{z^2}{2}}$. The expectation of z is denoted by $E\{z\}$, and $g(\omega_2) = q'(y, r) - \varepsilon_0 \leq 0$, in which the closeness measure $q'(y, r)$ is defined to achieve its minimum when $\omega_2^T = \omega_2^{*T}$ and ε_0 is a threshold.

To find the optimal ω_2^{*T}, a fast gradient-descent learning method is adopted to iteratively update the vector ω_2^T [34]. For our problem, the input color image x is represented by a $3 * N_I$ vector with R, G, and B values, where N_I is the total number of image pixels. The reference r is as the same size as the original image, with the derived part (face, torso, or hip) labeled as 1 and otherwise as 0. The main steps of ICA-R for local learning model of body parts are described in Algorithm 2, and the body part estimates are finally obtained from repeating the algorithm by changing the reference r, see Fig. 6(g)–(i). All the response maps for upper and lower body parts are integrated by selecting the maximum response at each



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

Fig. 7. (a) and (b) Results derived from only using boosted classifier. (c) and (d) Results obtained by combining boosted classifier and ICA-R local learning.

Algorithm 2 *ICA-R for body parts*

Input: original image x, determined parts ={ face, torso, hip}
1: for B = face, torso, hip
2: Reference image: $r = B$
3: Whiten the centered image $\widetilde{x}$: $E\{\widetilde{x}\widetilde{x}^T\} = I$
4: Initialize a unit-norm vector ω_2
5: Compute $\overline{\rho} = E\{G(y)\} - E\{G(g)\}$
6: Compute $g(\omega_2) = E\{(y - r)^2\} - \epsilon$ where ϵ is a small constant
7: Update μ by $\mu^{\kappa_2+1} = \max\{0, \mu^{\kappa_2} + \gamma' g(\omega_2^{\kappa_2})\}$
8: Update ω_2 by $\omega_2{}^{\kappa_2+1} = \omega_2{}^{\kappa_2} - \eta/\delta(\omega_2^{\kappa_2})$ and $\omega_2^{\kappa_2+1} \leftarrow (\omega_2^{\kappa_2+1})/\left\|(\omega_2^{\kappa_2+1})\right\|$
9: Repeat (5-8) until $\left\|(\omega_2{}^{\kappa_2})^{\top}\omega_2^{\kappa_2+1}\right\| \approx 1$
10: $y = \omega_2^{\top} x$
11: end for
Output: y = combine$\{y_{face}, y_{torso}, y_{hip}\}$ for body parts response

pixel, thereby rendering a map indicating the potential body [see Fig. 6(j)]. This response map is used for computing P_2 of (11). Fig. 7(c) and (d) show that a better result can be obtained by combining the boosted classifier and ICA-R local learning.

C. Data-Driven Graph Cut

Given an image I, denote $p \in I$ as a pixel represented with a RGB vector and N_{p_i} as the set of 4-connected neighbors at the pixel p_i. Assume that the segmentation of the image is given by L with label $l_{p_i} = 0$ for background and 1 for foreground at the pixel p_i. The optimization problem of graph cut is to minimize the energy function [36]

$$E(L) = \sum_{p_i \in V} \lambda \phi(l_{p_i}) + \sum_{p_j \in N_{p_i}} \varphi(l_{p_i}, l_{p_j}) \delta(l_{p_i}, l_{p_j}) \quad (17)$$

where $\phi(l_{p_i})$ defines the penalty for assigning l_{p_i} to object or background, $\varphi(l_{p_i}, l_{p_j})$ describes the boundary properties of the segmentation and it is close to 0 when a pixel p_i and another pixel p_j are utterly different and is large when they are similar, and λ is a constant (e.g., $\lambda = 9$ in our experiments). The term $\delta(l_{p_i}, l_{p_j})$ is defined by

$$\delta(l_{p_i}, l_{p_j}) = \begin{cases} 1 & l_{p_i} \neq l_{p_j} \\ 0 & l_{p_i} = l_{p_j}. \end{cases} \quad (18)$$

The optimal solution of (17) can be found by a min-cut/max-flow algorithm [20]. Considering the label problem and pose estimation in an image simultaneously, we can obtain a new energy function

$$E(L, \chi^*) = \sum_{p_i \in v} \lambda \phi(l_{p_i} | \chi^*) + \sum_{p_j \in N_{p_i}} \varphi(l_{p_i}, l_{p_j}) \delta(l_{p_i}, l_{p_j}) \quad (19)$$

where χ^* represents the body pose, as shown of the red rectangles in Fig. 3. To solve the segmentation L using [20], we need to construct a graph with seeds, which will be introduced in the following.

D. Initialization with Pose Evidence

Fig. 2 illustrates how the top-down inference and bottom-up cues are integrated in this paper. From the top-down inference (described in Section IV), each pixel in an image is labeled by a body part or background. Likewise, each pixel is assigned by a probabilistic measure with visual cues in the bottom-up process (described in Section V). On one hand, the top-down process is able to determine the stable body parts, but usually can not find the precise contours due to the use of bounding box for body part model. On the other hand, the low-level cues from the bottom-up process can provide large details of body but with numerous false positives. The top-down and bottom-up cues are integrated as pose evidence in which the probabilistic estimates of the pixels within body parts (based on top-down inference) are obtained by $\max(P_i, 0.5)$ where P_i is the estimate from bottom-up process. Similarly, the probabilistic estimates for the pixels outside body parts are clamped by $\min(P_i, 0.5)$. Consequently, we have a pose evidence map where the probabilistic estimate at each pixel indicates the likelihood belonging to the human body.

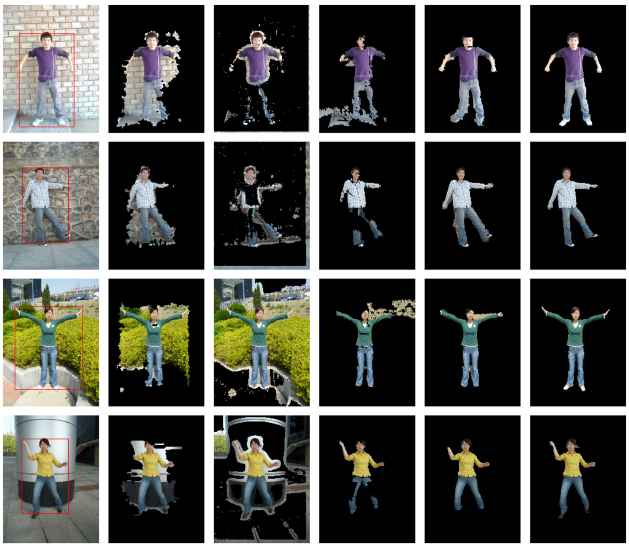


Fig. 8. Sample results on our dataset (the bounding box on images is used for GrabCut). "GT" is the abbreviation of ground truth. (a) Image. (b) GrabCut. (c) GM-EM. (d) BUGC. (e) Ours. (f) GT.

We construct an undirected graph $G = (V, E)$ defined as a set of nodes V, and a set of undirected edges E that connect the nodes. As in [20], the set of edges E consists of two types of undirected edges: n-links (neighborhood links) and t-links (terminal links). Each pixel p_i has two t-links connecting to each terminal (i.e., source and sink), and each pair of neighboring pixels $\{p_i, p_{j \in N_{p_i}}\}$ is connected by an n-link.

For the t-link, the pixels with higher probabilistic estimates in the pose area of the pose evidence map are labeled as foreground seeds, and another pixels with lower values outside the pose area are labeled as background seeds (e.g., 60% and 70% for foreground and background labels). Therefore, we have

$$P(p_i = foreground | \chi^*) = 1 - P(p_i = background | \chi^*) \quad (20)$$

at each node. Similar to the existing MAP-MRF formulation in the literature, their negative log-likelihoods are used as penalty term ϕ of (17). For the n-link, the energy term is computed by

$$\varphi(l_{p_i}, l_{p_j}) \propto \exp\left(-\frac{g^2(p_i, p_j)}{2\sigma^2}\right) \cdot \frac{1}{dist(p_i, p_j)} \quad (21)$$

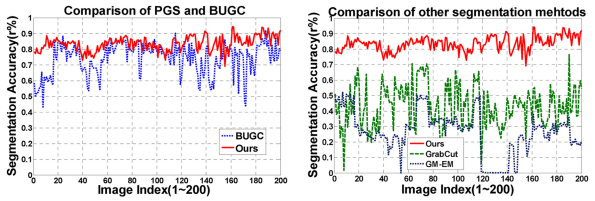
where $g(p_i, p_j)$ measures the difference in the RGB values of pixels p_i and p_j, and $dist(p_i, p_j)$ is their spatial distance. As we only consider the four-neighborhood pixels, it is simple to set $dist(p_i, p_j) = 1$. With this formulation, the optimal solution can be obtained by the max-flow/min-cut algorithm [20] for human body segmentation.

VI. Experiment Results

In this section, we present experiments on the dataset used in [37] and [38], which has 200 real world images with size of 208 × 156 pixels. Our current implementation with a combination of C++ and MATLAB runs about 5 min per frame

TABLE I
IMPACT OF PARAMETERS TESTING

par.	0.1	0.3	0.5	0.7	0.9	par.	0.1	0.3	0.5	0.7	0.9	1
ξ	0.8231	0.8312	0.8135	0.8245	0.8276	α	0.8223	0.8252	0.8308	0.8319	0.8324	0.8246
ζ	0.8031	0.8112	0.8315	0.8323	0.8236	β	0.8094	0.8108	0.8312	0.8313	0.8236	0.8320
$\frac{\tau}{10}$	0.8240	0.8318	0.8335	0.8322	0.8316	γ	0.8069	0.8102	0.8101	0.8332	0.8334	0.8276
ν	0.8332	0.8334	0.8130	0.8102	0.8069							



(a) (b)

Fig. 9. Comparison with other segmentation method. (a) Accuracy of ours (red line) and BUGC (blue dotted line). (b) Accuracy of ours (red line), GrabCut (green dot line), and GM-EM (blue dotted line).

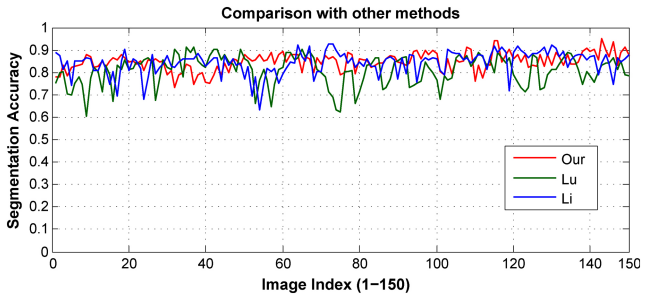


Fig. 10. Comparison with other segmentation methods. Red line: our method. Blue line: Li's approach. Green line: work of Lu.

without optimization on a 2.66 GHz Pentium 4 machine. The images consist of a large variation of pose, clutter, clothing, and lighting. In the following, we first introduce the parameter selection of our method, then discuss the effectiveness of our method and the comparison between proposed ICA-R and boosted classifiers, and at last, provide the comparison with other methods.

Parameters Selection: For selection precision, 50 images are first selected at random, and then the 50 selected images are perturbed by flip and slight rotation to generate 150 images. The parameters in the formulations are obtained by cross validation on the generated images. In our experiments, we use three-folder cross validation, i.e., the whole generated images are divided into three groups, with one group as the training images and the other two as the testing images. Therefore, we can obtain three groups parameters. Finally, we select the parameters as the ones with the highest accuracy in the testing images.

Additionally, we test the robustness of the parameters in torso detection and limb searching. There are four important parameters in torso estimation including ξ in the (3), ζ in (4), τ in (5), and ν in (7). To evaluate the accuracy of torso estimation, we label a rectangle as the ground truth to each torso in the image [see the green rectangle in Fig. 6(b)], and use (22) as the criterion

$$r = \frac{\sum_{p_i}(B(p_i) \bigcap H(p_i))}{\sum_{p_i}(B(p_i) \bigcup H(p_i))} \tag{22}$$

where B is the torso estimation (a rectangle plane), H denotes the corresponding torso ground truth, and the operators of $\bigcap$ and $\bigcup$ represent pixel-wise AND and OR, respectively. We first change one parameter from 0.1 to 0.9 with the other three parameters fixed and then obtain five sets of the mean accuracy on the whole testing images. As shown in the left of Table I, the performance of torso detection will not decrease significantly even though more than one parameter changes within a certain range.

The similar testing is performed on the parameters of limb searching. First, the ground truth of the left upper limb is labeled with a rectangle for each image by hand, and then we test the limb searching based on the accurate detected torso and compute the evaluation using the (22). The robustness of three parameters including α, β and γ in (10) is shown on the right of Table I. As shown in Table I, the parameters in the limb searching are also robust within a certain range.

Effectiveness of Our Method: Some sample results are shown in Fig. 8, and more results (with large images) can be found in the supplementary material. Overall, our method is able to segment out humans in cluttered background. The GrabCut [21] method does not perform well when a person appears in cluttered backgrounds. The GM-EM [18] performs poorly in which background is often segmented out with body, and the bottom-up graph cut (BUGC), which uses bottom-up visual cues and without the top-down inference for segmentation) often misses some small human parts (e.g., head or lower arms). In addition, we tried our method on the challenging images of dataset [27]. This dataset has 206 images with numerous poses, variations and part occlusion. We only perform our method on the frontal person images. As shown in Fig. 13, our method is effective to recover the human body region in some challenging images, for example, part occlusion and intense illumination.

Comparison Between ICA-R and Boosted Classifiers: It is important to use ICA-R and boosted classifiers to exploit the bottom-up features for lower body segmentation. ICA-R has the ability to extract the similar information with reference information from the source information. In our method, the reference is selected as a white-black image with 1 corresponding to the object and 0 corresponding to the background, which can enlarge the contrast between the object

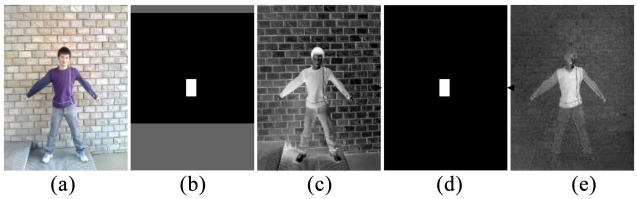


Fig. 11. Example of ICA-R and boosted classifiers. (a) Original image. (b) Samples of boosted classifier. (c) Probability map of boosted classifier. (d) Reference of ICA-R classifier. (e) Probability map of ICA-R classifier.

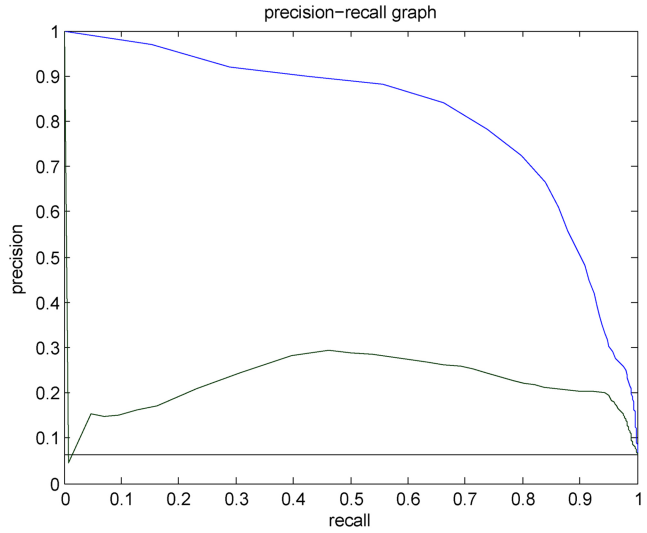


Fig. 12. PR curve of ICA-R and boosted classifiers where the blue curve presents ICA-R classifier and the green is boosted classifier.

and background. Based on Adaboost algorithm, the boosted classifier is powerful to classify the object efficiently.

To compare ICA-R and boosted classifiers, an experiment for torso segmentation is designed. For the ICA-R classifier, the reference images are hand labeled with 1 corresponding to the object and 0 corresponding to the background, as shown in Fig. 11(d). For the boosted classifier, the positive samples are selected from the region which the object part in the ICA-R reference corresponds to and the negative samples are selected from the top and bottom sides of the images, as shown in the gray region of Fig. 11(b). 50 images are selected at random for the experiment. Once finishing the classification, we can obtain the probability response maps, which can be also called saliency maps. According to the evaluation of saliency map [39] and segmentation [40], we can use the PR curve to describe the ability of classifiers. Fig. 12 shows the comparison, it is obvious that ICA-R classifier is better than the boosted classifier. From the experiment, we find that if the positive and negative samples are provided correctly and enough, the probability map will be satisfied. Therefore, the performance of boosted classifier depends on the selected samples. In contrast, ICA-R classifier can classify object better with a little requirement of reference.

Comparison with Other Methods: For a quantitative evaluation, the segmentation accuracy is defined as the proportion of pixels correctly classified as foreground or background

TABLE II
AVERAGE ACCURACY AND DEVIATION OF SEGMENTATION ON 200 TESTING IMAGES

Method	Average	Deviation
GrabCut [21]	0.4322	0.0181
GM-EM [18]	0.2629	0.0189
BUGC	0.7393	0.0101
Ours	0.8359	0.0021

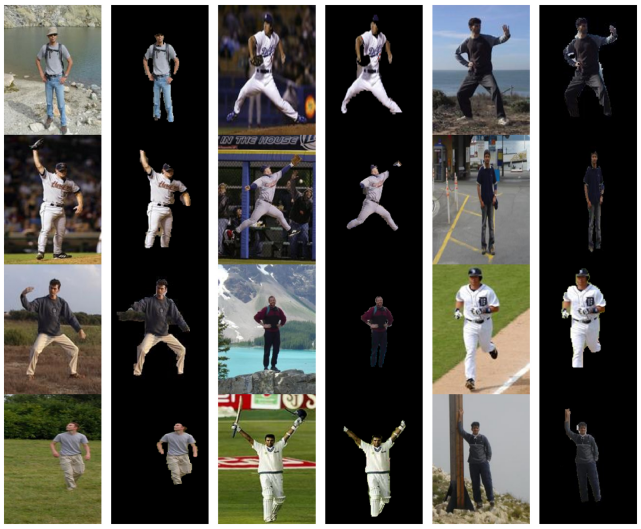


Fig. 13. Sample results of front face images collected from [27].

by comparing with the ground truth segmentation, which is defined as follows:

$$r = (1 - \frac{\sum_{p_i \in V} |B(p_i) - H(p_i)|}{N_I}) \tag{23}$$

where B is the binary segmentation image, H is the hand labeled ground truth, and N_I is the total number pixels in the image. The quantitative results shown in Fig. 9 suggest that our method performs well with a stable segmentation and a high accuracy as well as a low deviation. The BUGC method [Fig. 9(a)] performs much worse than our method with a large deviation as only simple features are considered by BUGC. The GrabCut and GM-EM methods perform poorly with the low accuracy and high deviation [Fig. 9(b)] due to the limited ability of Gaussian model for complexity backgrounds. The average accuracy and deviation are presented in Table II.

To further evaluate our method using (23), we compared our method with the work of Li *et al.* [37] and Lu *et al.* [38], which have the state-of-the-art performance on the same dataset. In these methods, Li designed two deformable models to segment human body on two scale superpixels and Lu used the coarse-to-fine strategy to segment body on the same dataset. Fig. 10 shows the comparison among 150 images as the method of Li only made the evaluation on 150 images. The red line stands for our method, the blue is the method of Li *et al.* [37] and the green is the approach of Lu *et al.* [38]. As implied in this figure, Li's approach is a little better than the method of Lu, and our method is better than Li's. The means and derivations of these methods are provided in the Table III. Our method

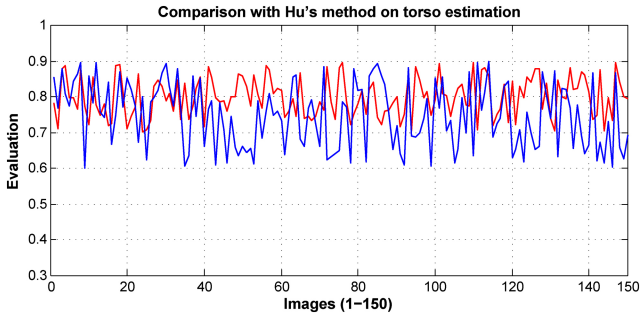


Fig. 14. Torso comparison with [30].

TABLE III
AVERAGE ACCURACY AND DEVIATION OF SEGMENTATION ON 150 TESTING IMAGES

Method	Average	Deviation
Li's method [37]	0.8432	0.0026
Lu's method [38]	0.8084	0.0043
Our method	0.8481	0.0016

has the largest mean and the smallest derivation on the 150 testing images. As shown in Table III, we achieve a more stable and accurate performance than the compared methods. Our method can cover more than 84% of the ground truth and has 0.16% variation on average, which are both better than the state-of-the-art methods.

The torso connects the face and limbs, which is the most important part to body detection, and therefore, we test the accuracy of torso estimation to show the robustness of our method. The method of [30] is used for the comparison as this method is also proposed for torso detection. To evaluate the torso detection, similarly to the parameters robustness testing, we first label the torso rectangle as the ground truth, then estimate the torso based on the accurate detected face, and finally compute the accuracy using (22). Fig. 14 shows the comparison with the method of [30] where the red line stands for our method and the blue one is the method of Hu. In most cases, our method is better than Hu's approach.

VII. CONCLUSION

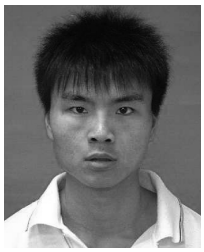
In this paper, we propose an effective data-drive graph based algorithm for human body segmentation without user intervention, which is still one of the most challenging tasks in the computer vision. Our data-driven method incorporates the top-down inference obtained from human kinematics with the bottom-up cues derived from the combination of the boosted classifier and ICA-R local learning for body segmentation in static images within the graph cut framework, which performs well against the existing object-centered graph cut methods for people with a large amount of poses. The experiment results show that our method can achieve a satisfying body segmentation on our dataset, as well other dataset.

Our method currently works with people in near-frontal pose and has limitations on handling the self-occlusion, especially when the occlusion is serious, such as the limb overlapping with body parts. In addition, the current implementation does not run in real time, and our method is difficult to expand to other object classes for our method needs the face detection as the precondition. Our future work will extend the proposed algorithm to deal with these problems.

REFERENCES

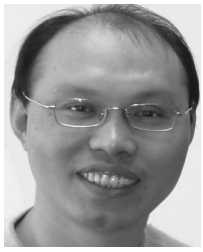
[1] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[2] X. Zhang, C. Li, X. Tong, W. Hu, S. J. Maybank, and Y. Zhang, "Efficient human pose estimation via parsing a tree structure based human model," in *Proc. ICCV*, 2009, pp. 1349–1356.
[3] H. Jiang, "Human pose estimation using consistent max-covering," in *Proc. ICCV*, 2009, pp. 1357–1364.
[4] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *Proc. CVPR*, 2000, pp. 1677–1684.
[5] A. Bissacco, M.-H. Yang, and S. Soatto, "Detecting humans via their pose," in *Proc. NIPS*, 2006, pp. 169–176.
[6] E. Borenstein and J. Malik, "Shape guided object segmentation," in *Proc. CVPR (1)*, 2006, pp. 969–976.
[7] Z. Lin, L. S. Davis, D. S. Doermann, and D. DeMenthon, "An interactive approach to pose-assisted and appearance-based segmentation of humans," in *Proc. ICCV*, 2007, pp. 1–8.
[8] T. Cour and J. Shi, "Recognizing objects by piecing together the segmentation puzzle," in *Proc. CVPR*, 2007, pp. 1–8.
[9] P. Kohli, J. Rihan, M. Bray, and P. H. S. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 285–298, 2008.
[10] J. Wu, W. Cai, and A. C. S. Chung, "Posit: Part-based object segmentation without intensive training," *Pattern Recognit.*, vol. 43, no. 3, pp. 676–684, 2010.
[11] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Objcut: Efficient segmentation using top-down and bottom-up cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 530–545, Mar. 2010.
[12] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," in *Proc. ICCV*, 2003, pp. 18–25.
[13] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Obj cut," in *Proc. CVPR (1)*, 2005, pp. 18–25.
[14] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," in *Proc. ECCV (4)*, 2006, pp. 581–594.
[15] S. Li, H. Lu, and L. Zhang, "Arbitrary body segmentation in static images," *Pattern Recognit.*, vol. 45, no. 9, pp. 3402–3413, 2012.
[16] H. Lu, G. Fang, and X. S. X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 889–899, Jun. 2012.
[17] H. Lu, X. Shao, and Y. Xiao, "Pose estimation with segmentation consistency," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4040–4048, Oct. 2013.
[18] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002.
[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[20] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. ICCV*, 2001, pp. 105–112.
[21] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[22] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *Proc. ECCV (2)*, 2002, pp. 109–124.
[23] J. M. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Proc. ICCV*, 2005, pp. 756–763.
[24] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. ICCV*, 2009, pp. 1–8.
[25] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *Proc. CVPR (2)*, 2004, pp. 326–333.
[26] G. Hua, M.-H. Yang, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," in *Proc. CVPR (2)*, 2005, pp. 747–754.

[27] D. Ramanan, "Learning to parse images of articulated bodies," in *Proc. NIPS*, 2006, pp. 1129–1136.
[28] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proc. ICCV*, 2005, pp. 446–453.
[29] L. Sigal, M. Isard, B. H. Sigelman, and M. J. Black, "Attractive people: Assembling loose-limbed models using non-parametric belief propagation," in *Proc. NIPS*, 2003, pp. 1539–1546.
[30] Z. Hu, G. Wang, X. Lin, and H. Yan, "Recovery of upper body poses in static images based on joints detection," *Pattern Recognit. Lett.*, vol. 30, no. 5, p. 503–512, 2009.
[31] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proc. CVPR*, 2008, pp. 1–8.
[32] P. Ott and M. Everingham, "Implicit color segmentation features for pedestrian and object detection," in *Proc. ICCV*, 2009, pp. 723–730.
[33] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[34] Q.-H. Lin, Y.-R. Zheng, F. Yin, H. Liang, and V. D. Calhoun, "A fast algorithm for one-unit ica-r," *Inf. Sci.*, vol. 177, no. 5, pp. 1265–1275, 2007.
[35] S. Li and H. Lu, "body segmentation based on ica-r at two-scale superpixel," *IET Image Process.*, vol. 6, no. 6, p. 770–777, 2012.
[36] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vision*, vol. 70, no. 2, pp. 109–131, 2006.
[37] S. Li, H. Lu, R. Xiang, and Y. W. Chen, "Human body segmentation based on deformable models and two-scale superpixel," *Pattern Anal. Appl.*, vol. 15, no. 3, pp. 399–413, 2012.
[38] H. Lu, G. Fang, X. Shao, and X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 889–899, Jun. 2012.
[39] J. Sun, H. Lu, and S. Li, "Saliency detection based on integration of boundary and soft-segmentation," in *Proc. ICIP*, 2012, pp. 1085–1088.
[40] H. Lu, R. Zhang, S. Li, and X. Li, "Spectral segmentation via midlevel cues integrating geodesic and intensity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 43, no. 6, pp. 2170–2178, Dec. 2013.



Shifeng Li received the B.E. degree in computer engineering from Liaoning University of Technology, Liaoning, China, in 2006, and the M.E. degree from Northeastern University, Liaoning, in 2008. He is currently pursuing the Ph.D. degree from the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China.

He is currently with the People's Bank of China, Benxi, China. His current research interests include pattern recognition, image processing, and computer vision.



Huchuan Lu (SM'12) received the M.Sc. degree in signal and information processing, and the Ph.D. degree in system engineering from Dalian University of Technology (DUT), Dalian, China, in 1998 and 2008, respectively.

Since 1998, he has been into teaching, and since 2011, he has been a Professor with the School of Information and Communication Engineering, DUT. In recent years, he focused on visual tracking and segmentation. His current research interests include the areas of computer vision and pattern recognition.

Dr. Lu is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS, and a member of the ACM.

Xingqing Shao received the B.E. degree from An-Qing Teachers College, Anhui, China, in 2008, and the M.S. degree from the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China, in 2012.

His current research interests include human segmentation, pose estimation, and object recognition.