Segmenting Human From Photo Images Based on a Coarse-to-Fine Scheme

Huchuan Lu, Member, IEEE, Guoliang Fang, Xinqing Shao, and Xuelong Li, Fellow, IEEE

Abstract—Human segmentation in photo images is a challenging and important problem that finds numerous applications ranging from album making and photo classification to image retrieval. Previous works on human segmentation usually demand a time-consuming training phase for complex shape-matching processes. In this paper, we propose a straightforward framework to automatically recover human bodies from color photos. Employing a coarse-to-fine strategy, we first detect a coarse torso (CT) using the multicue CT detection algorithm and then extract the accurate region of the upper body. Then, an iterative multiple oblique histogram algorithm is presented to accurately recover the lower body based on human kinematics. The performance of our algorithm is evaluated on our own data set (contains 197 images with human body region ground truth data), VOC 2006, and the 2010 data set. Experimental results demonstrate the merits of the proposed method in segmenting a person with various poses.

Index Terms—Graph cuts, human segmentation, multicue coarse torso detection algorithm (MCTD), multiple oblique histogram (MOH).

I. INTRODUCTION

W ith the development of digital cameras, more and more intelligent processing of photos is increasingly demanded, such as photo classification and image retrieval. Automatic segmenting of a person becomes more crucial, so that pose analysis (e.g., running) is available and high-level applications on human (e.g., gait recognition [1], [2]) can be performed.

Image segmentation is still an open problem, and many new segmentation approaches are proposed (e.g., [3]–[6]). However, human segmentation in static images from cluttered background is paid less attention. Precious works find the homogeneous regions and then identify those corresponding to a single object according to their feature properties, such as smoothness and continuity of bounding contours. For human segmentation, there are multiple regions of body parts, such as head, torso, and legs in the image, as a result of large appearance variation. Without additional constraints, it is not clear whether these regions can be grouped into humanlike segmentation. Some top–down cues such as shape are applied to learn the properties to guide segmentation [7]–[9]. The main challenge of these algorithms is to account for large variability of shape and appearance of a given object class. Consequently, the results may not accurately delineate contours in the context of human segmentation. Furthermore, it is rather challenging to learn person shapes that encompass large variation of arbitrary pose. Much success has only been demonstrated in the context of object segmentation with limited pose and shape variation (e.g., pedestrian).

Cour and Shi [10] grouped the oversegmentation results from bottom–up methods into homogeneous regions to achieve best matching against templates. However, it is difficult to extend this approach to human segmentation as it needs to handle high-dimensional pose state and train a large number of shape templates. Although the pictorial structure method [11] can be applied with bottom–up visual cues to infer human pose and, in turn, used for segmentation, it is still a challenge to estimate accurate pose. Consequently, any successful extension of conventional algorithms to segment human usually employs the training-test scheme or model hypothesis on person pose. Unfortunately, it is impossible to build a set of model hypotheses covering all pose variations for matching.

In this paper, we propose a robust framework to recover human body from photo images via integrating top–down body information and low-level visual cues into Graph Cuts framework. As shown in Fig. 1, we divide whole-body extraction into two subtasks, i.e., upper-body and lower-body segmentations. We constrain our researches on those human poses with frontal/side faces, which are common in photos.

The main problem of Graph Cuts is to assign a binary label to each pixel. A common approach [12] that is utilized in our scheme is to construct the foreground and background...
graphs containing the likelihood term of each node being foreground/background and the piecewise smoothness term indicating the pixels in the same region having the same labels. Given an image \( I \) with pixels \( v \), let \( A \) be a binary vector whose elements \( A_i \) specify assignments to pixels \( v \).Thus, an energy function is formulated as follows:

\[
E(A) = \varepsilon \sum_{i \in v} R_i(A_i) + \sum_{(i,j) \in N_i} Q_{i,j}(A_i, A_j). \tag{1}
\]

The first term defines the penalties for assigning pixel \( i \) to label \( A_i \) (foreground or background), calculated by the negative log likelihood of the foreground/background models. The second term describes the boundary properties of segmentation \( A \) and it is large when pixels \( i \) and \( j \) are similar and close to zero when the two are quite different. As a typical choice, we use

\[
Q_{i,j} \propto \exp\left(-\frac{g^2(i,j)}{2\sigma^2}\right) \cdot \frac{1}{dist(i,j)} \tag{2}
\]

where parameter \( \sigma \) is related to the level of variation between neighboring pixels. Function \( g \) measures the difference of pixels, and function \( dist \) is their spatial distance. Parameter \( \varepsilon \) in (1) is utilized to control relative importance, and \( N \) is a neighborhood such as a four-connected and eight-connected neighborhood. \( \delta(A_i, A_j) \) is 0 if \( A_i = A_j \), and 1 otherwise.

A consequent graph can be constructed with the nodes corresponding to pixels of the image and the edges, which consist of two types of undirected edges, i.e., \( n \)-links (neighborhood links) and \( t \)-links (terminal links). Each pixel has two \( t \)-links connecting to each terminal (i.e., source and sink), and each pair of neighboring pixels \( \{i, j\} \) in \( N \) is connected by an \( n \)-link. Then, the optimal solution of (1) can be obtained by the max-flow/min-cut algorithm [13].

To extract accurate human region and overcome shortcomings of conventional algorithms, a coarse-to-fine strategy is employed to obtain the human shape constraints, which is integrated into the Graph Cuts framework with low-level cues for final segmentation. Fig. 2 shows the flowchart of our method. We first adopt a multiview face detector [14] to locate the face region as \( a \ priori \), and then, the multicore coarse torso detection algorithm (MCTD) is utilized to segment the upper body that adjoins to head, in which the Normalized Cuts [15] and global probability of boundary (gPb) [16] are effectively combined. According to human topology, the accurate lower body is segmented based on iterative multiple oblique histogram (MOH). Different from classic graph cuts [12], in which the point seeds are set by the user, we recover the body region without human intervention. Automatic seed point setting is very important, which is highly dependent on the accuracy of part detection. The proposed algorithm can accurately and robustly detect part regions, thereafter accurately setting the seed points inside and outside the body regions, respectively, and achieving fine results.

The rest of this paper is organized as follows: Section II introduces related work on human segmentation. The upper-body segmentation based on the MCTD algorithm is presented in Section III. Section IV describes lower-body segmentation by initializing the foreground seed region and iterative MOH algorithm. Experimental results and the conclusion of this paper are presented in Sections V and VI, respectively.

II. RELATED WORK

Recently, researchers have developed many effective methods to recover human body from video sequences [17]–[19]. Yin and Collins [18] combined global shape information with local graph links in a Conditional Random Field framework. Video-based methods are proved to be robust and effective, but they achieve human segmentation by using motion cues among video frames, which is not available in the static image. Jiang [20] proposed an efficient method for body part matching and segmentation. The algorithm must be given a single exemplar image, so that the part candidates can be found with Chamfer matching and human segmentation can be performed as a constrained optimization problem. Unfortunately, the part candidates in still image are difficult to detect, and the scales and appearance are usually different. Therefore, the video-based algorithm cannot be applied to recover the human body in a single image.

Most of the algorithms for recovering human body in a static image fall into two categories, i.e., exemplar-based [8], [21], [22] and part-based [23]–[25] approaches. In exemplar-based approaches, an exemplar pool should be constructed first, and then, the test images are matched with the exemplars or models. These models cannot always accurately segment the human body, because human poses are arbitrary and an exemplar pool cannot cover all the situations of poses and appearance variation. It is difficult to extend the method of [22] for human segmentation, in which Kumar and Torr drew a pictorial structure and Markov random fields (MRFs) together for detecting and segmenting instances of a particular object (e.g., cows and horses) with a limited pose variation. Gao et al. [26] proposed an adaptive contour feature (ACF) for human segmentation. After labeling 620 human silhouettes, they trained cascade segmentation classifiers directly with the ACF. Although the ACF is very effective for detection, the results of human segmentation are rough, as the labeled samples are limited. In addition, it is almost impossible to construct a reasonable database for all the human poses. Lin et al. [27] proposed an interactive segmentation method that incorporates local MRF constraints and global shape priors to iteratively estimate segmentations and pose simultaneously. Kohlic et al. [8] utilized pose-specific conditional random and stick figures for segmentation, as well as pose estimation of
humans within a Bayesian framework, which has been successfully used in 3-D human pose tracking.

The part-based approaches recover human body configurations by assembling a set of candidate parts (e.g., [28]). Mori et al. [23] found salient half-limbs and torso by training part detectors with four main cues and Normalized Cuts [29]. In their work, hand-segmented limbs are used for training. However, Normalized Cuts usually do not accurately segment half-limbs and torso. Ren et al. [24] developed a framework that incorporates arbitrary pairwise constraints between body parts. Some part candidates are detected as parallel line segments on probability of boundary (Pb) [30] and then, the human body configurations are recovered by assembling the candidates with integer quadratic programming. One primary problem with part-based methods is that it is very hard to design a robust part detector. In addition, performance of the part detector is limited without global constraint among body parts.

Chen and Fan [31] presented the hybrid body representation method integrating pose recognition, localization, and segmentation into a whole framework with restricted pose class (e.g., walking) to segment human. The method constrains pose variation in a few classes, so the application is very limited. Ferrari et al. [32] utilized the GrabCut algorithm to highlight human foreground based on a head detector to reduce the search space and then detect each body part for estimating pose. Bourdev et al. [33] used a poselet detector to detect the keypoints of the human and train a figure-ground predictor of each poselet to segment the human. The trained poselet masks are merged with boundary cues to locate specific poselets, and finally, the human is segmented. A contour person model containing shape variation, viewpoint, and rotation was defined by Freifeld et al. [34], which is learned from a 3-D scape. Initialized by a pose estimation step [35], the human can be extracted via optimizing cost function with a gradient-free direct search method. However, it is still a challenging problem to obtain correct pose initialization.

III. UPPER-BODY SEGMENTATION

In this section, we describe the details of our scheme for segmenting upper body, which is crucial for whole-body segmentation. Given a photo image, we first use a face detection method [14] to locate the human face, which is available in most current digital cameras. In addition, then, a coarse torso (CT) as shown in Fig. 3(d) is detected using Graph Cuts [12]. To find the torso, Mori et al. [23] first searched for all combinations of Normalized Cuts [15] segments [Fig. 3(b)]. A pixelwise torso is then segmented using Graph Cuts [12]. To find the torso, Mori et al. [23] first searched for all combinations of Normalized Cuts segments that satisfy a scale constraint and then classified these candidates with a set of cues. However, their torso results may be broken or contain background, particularly when some Normalized Cuts segments simultaneously contain background and foreground. In addition, the exhaustive search for all combinations may bring more amphibulous candidates. Hu et al. [36] detected torso on dominant colors generated by using the k-means clustering algorithm. However, the dominant colors are not reliable in appearance variation and cluttered backgrounds.

Fig. 3. Upper-body segmentation. (a) Input image with bounding-box candidates. (b) Normalized Cuts segments. (c) Result of gPb. (d) CT with the bounding box. (e) Upper-body segmentation.

A. CT Detection

We reduce the search for torso with keeping up segmentation quality via fully automatic spectral clustering according to the comprehensive Normalized Cuts criterion [15]. The segmentation structure can reflect the true torso quite well, hence facilitating torso extraction. In our scheme, the Normalized Cuts segments are grouped into a torso candidate region based on the bounding box along different orientations, as shown in Fig. 3(a) and (b), where the bounding boxes are generated according to face priori. In the combining procedure, three cues are employed to select the best candidate as CT: area probability, location probability, and contour probability.

Area Probability: Suppose that there are $N$ bounding boxes along different orientations relative to the face and $L_i$ segments overlapped with the $i$th bounding region, $i = 1, 2, \ldots, N$. We denote the area of the $i$th bounding-box region as $R_i$, the area of the $j$th segment overlapped with $R_i$ as $S_{i,j}$, and the overlapped areas as $O_{i,j}$, $j = 1, 2, \ldots, L_i$. Then, the area probability $AP_{i,j}$, which indicates the $j$th segment under the $i$th bounding box belonging to torso or not, is defined as

$$AP_{i,j} = \left( \frac{O_{i,j}}{S_{i,j}} \right)^{\alpha} \left( \frac{O_{i,j}}{R_i} \right)^{\beta}$$

where parameters $\alpha$ and $\beta$ are the weighting coefficients in (0, 1) to control the importance of these two terms. This probability evaluates how a segment contributes to torso, given bounding box in terms of region information.

Location Probability: As the segments are not irregular, we define a location probability that describes the likelihood of each pixel in a segment unit belonging to the given bounding-box candidate. Therefore, the contribution of the segment unit to the candidate can be estimated by the location cue. Let $W$ and $H$ be the width and height of the bounding-box regions, respectively. Given a pixel $(x, y)$, we define

$$\rho_i(x, y) = \left[ \left( \frac{d_i(x, y) \sin \theta_i}{W/2} \right)^2 + \left( \frac{d_i(x, y) \cos \theta_i}{H/2} \right)^2 \right]^{1/2}$$

where $\theta_i$ is the vertical angle of the $i$th bounding region, and $d_i(x, y)$ is the Euclidean distance from pixel $(x, y)$ to the center of the bounding-box region. Then, the location probability $LP$ of $S_{i,j}$ is defined as

$$LP_{i,j} = \sum_{(x, y) \in S_{i,j}} \exp \left( - \frac{\rho_i(x, y)^2}{W} \right)$$
where parameter $\gamma$ is the weighting factor. By considering location term, some segment candidates that are clearly not torso components can be removed.

**Contour probability:** Maire et al. [16] developed a high-performance contour detector employing a combination of local and global cues, i.e., $gPb$. The detector of [30] is used to predict local boundary, and the first $k$ generalized eigenvectors of Normalized Cuts are utilized to obtain global constraint. Thus, $gPb$ gives more salient edge information than local $Pb$, which is confirmed by Maire et al.

As shown in Fig. 3(b) and (c), the boundary of torso indicates a very strong global information that can constrain the segment grouping into a CT region. The contour cue $CP$ of a region is defined as the average $gPb$ along its boundary. Let the contour probability be $gPb$; hence, the boundary constraint of $j$ segments to be CT under the $i$th bounding box $R_i$, i.e.,

$$CP_{i,j} = \frac{1}{N} \sum_{n=1}^{N} gPb(n)$$

(6)

where $n$ denotes the index of pixel $(x, y)$ on the boundary of $R_i$.

**MCTD:** Based on the aforementioned three cues, the CT can be estimated with the MCTD algorithm.

Given a bounding-box region $R_i$, we first find all segments $S_i$ that are overlapped with $R_i$ without considering the head region. For each segment unit in $R_i$, we compute the area and location probability presented in (3) and (4), which can be treated as local information

$$P_{i,j} = (AP_{i,j})^{-1}(\lambda(LP_{i,j})^{-1})^2$$

(7)

where parameter $\lambda$ is the weighting term. The closer to the center of the bounding-box region, the more likely the segment to be a component of torso. Thus, we group each segment unit into the torso according to $P_{i,j}$ in sequence. Once a segment is added into the torso region, we compute and recompute the contour probability (6) to constrain the unlimited increase in CT, therefore generating a CT pool. With the global contour constraint, the grouping region will converge to the CT.

Then, the best group for the segments corresponding to the $i$th bounding-box region and its contour probability are selected by

$$CT_i = CT_{i,j} \left\{ \hat{j}^{*} = \arg \max_{j} (CP_{i,j}) \right\}$$

$$CP_{i} = CP_{i,j} \left\{ \hat{j}^{*} = \arg \max_{j} (CP_{i,j}) \right\}$$

(8)

where $CT_{i,j}$ describes that the CT candidate consists of $j$ segments under the $i$th bounding box, and $CP_{i,j}$ denotes the corresponding contour probability. This means that, in the CT pool, the grouping region with the strongest contour cue (i.e., the maximum mean $Pb$) is selected as CT.

In addition, considering all bounding box, the final torso is determined by the local and contour cues, i.e.,

$$CT = CT_{i} \left\{ i^{*} = \arg \max_{i} (P_{i} + CP_{i}) \right\}$$

(9)

where $P_{i} = P_{i,j} \left\{ j^{*} = \arg \max_{j} (CP_{i,j}) \right\}$, which denotes the local cues with the strongest contour. The details of the MCTD algorithm are described in Algorithm 1.

### Algorithm 1: MCTD

1: for $i = 1$ to $N$ do
2: for $k = 1$ to $K$ do, where $k$ is fine tuning parameter of orientation
3: Find all segments overlapped with $R_{i,k}$
4: Remove the head region
5: for $j = 1$ to $L_{i,k}$ do
6: Sort segments by descending $P_{i,k,j}$, where $P_{i,k,j} = (AP_{i,k,j})^{-1}(\lambda(LP_{i,k,j})^{-1})^2$
7: $CT_{i,k,j} = CT_{i,k,j} + S_{i,k,j}$
8: if $CT_{i,k,j} > S_{max}$ break
9: end for
10: $CT_{i,k} = CT_{i,k,j} \left\{ \hat{j}^{*} = \arg \max_{j} (CP_{i,k,j}) \right\}$
11: $CP_{i,k} = CP_{i,k,j} \left\{ \hat{j}^{*} = \arg \max_{j} (CP_{i,k,j}) \right\}$
12: Update $\theta_{i,k}$ according to $CT_{i,k}$
13: end for
14: $Eval_{i,k} = PCT_{i,k} + CP_{i,k}$
15: end for
16: $CT = CT_{i,k} \left\{ \langle i, k \rangle^{*} = \arg \max_{i,k} (Eval_{i,k}) \right\}$, $R_{CT} = R_{i,k} \left\{ \langle i, k \rangle^{*} = \arg \max_{i,k} (Eval_{i,k}) \right\}$

### B. Upper-Body Segmentation on CT

Hu et al. [36] used Canny operator and CDT to estimate the background seeds of the upper body. However, their scheme will fail in cluttered background and may segment some background when the Canny operator misses some clothing edge, particularly the lower edge. The main reason is that all information they use is Canny edge, which is not reliable in cluttered sense. In addition, the CT detection algorithm based on dominate colors [37] is not reliable enough.

The CT and the detected face region directly provide strong hard constraints for upper-body segmentation, and the $t$-links connecting to the upper body can be constructed by adopting kernel density estimation (KDE). Given a pixel $x$, the similarity between the pixel and the torso region $\{x_i\}$ and the face region $\{x_j\}$ is defined as

$$f_u(x) = \frac{1}{mnh^2} \sum_{x_i \in S_f} K\left(\frac{x-x_i}{h}\right) \sum_{x_j \in CT} K\left(\frac{x-x_j}{h}\right)$$

(10)

where $m$ is the number of pixels in the face region $S_f$, $n$ is the number of pixels in $CT$, and $K$ is the kernel function. Here, we use the Gaussian kernel with mean zero and variance $h$ as follows:

$$K\left(\frac{x_1-x_2}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{d^2(x_1,x_2)}{2h^2}}$$

(11)

where $d^2(x_1, x_2)$ is the Euclidean distance between $x_1$ and $x_2$.

According to the anthropometry, we initialize a circle region centered at the neck for upper-body segmentation, as shown
in Fig. 4(b). Therefore, we can set background seeds outside the initial circle region. In addition, the pixels in the lines close to the waist and head belonging to the background or not can be determined according to the CT and detected face region, by comparing the similarity of the pixels to the torso and face regions, thereby much noise removed in segmentation. For example, as shown in Fig. 4(a), line pixels (pixels in a line), which are parallel to the right and left sides of the bounding box of CT $R_{CT}$ but without overlapping with the area of CT, are chosen to be background seeds when $f_u(x)$ in (10) is less than a threshold. Hence, the $t$-links connecting to background can be calculated as

$$b_u(x) = \frac{1}{l h^2} \sum_{x_i \in B_u} K \left( \frac{x - x_i}{h} \right)$$ \hspace{1cm} (12)

where $l$ is the number of background seeds, and $B_u$ is the background region. In addition, then the probability of pixel $x$ belonging to foreground $p(x = fg)$ is

$$p(x = fg) = 1 - p(x = bg) = \frac{f_u(x)}{f_u(x) + b_u(x)}. \hspace{1cm} (13)$$

Combined with $n$-links computed by (2), the upper body is segmented using the max-flow/min-cut algorithm [13] shown in Fig. 3(e).

IV. LOWER-BODY SEGMENTATION

Lower-body segmentation is more challenging than upper-body segmentation, because the poses of legs are unpredictable. Mori et al. [23] detected “half-limbs” as single Normalized Cuts segments and then extended half-limbs to full legs using superpixels, as well as contour cue. However, the performance of half-limb detection is not robust, as it is uncertain that all legs can be entirely split into half-limbs. In addition, the extension may fail when the Pb [30] of legs is not salient, as shown in Fig. 3(c). So does the scheme in [24], in which the legs are located by finding parallelism based on Pb.

We develop a coarse-to-fine scheme to segment the lower body. As shown in Fig. 5, at first, the pixels on the dominate color [37] of the foreground seed region (c), which is estimated using the segmented upper body (b), are used to construct $t$-links connecting to figure. Meanwhile, the ground connected to $t$-links are set based on the segmented upper-body region, as well as the region around the lower body and between the legs if necessary (d). Then, the coarse lower body (e) can be obtained by employing the max-flow/min-cut algorithm. The final segmentation can be achieved by iterative MOH.

A. Coarse Lower-Body Segmentation

Foreground Seed Region Initialization: We first fit a torso rectangle exactly, which is more accurate than the bounding box to estimate the foreground region. We estimate the torso rectangle on the segmented upper-body image, in which all the background is set in black, which is different from the work [36] where the torso is estimated by formulating maximum $a posteriori$ with a deformable torso model with four parameters (torso width, torso height, torso inclining orientation, and the neck position in the image). In addition, a local search can speed up the fitting because the orientation of $R_{CT}$ is determined, and thus, we just need to adjust the torso orientation in a small range.

According to body topology, we initialize a rectangle in the foreground region along the inclining orientation of the torso shown as the red rectangle in Fig. 5(b). The width of the foreground bounding box is set to the width of the torso rectangle, and the height is in direct proportion to the width. The pixels with dominant colors [37], which are the maximal rectangle [red rectangles in(b)] near the bounding box [red in (a)] by adjusting the width, height, inclining orientation, and neck position. Meanwhile, a background bounding region [blue rectangle in (b)] is generated according to the CT bounding box. We then calculate the foreground and background distributions based on the regions in each rectangle candidate and the back-ground bounding region, respectively. Finally, Bhattacharyya coefficients on these distribution pairs [the histograms in (c) and (d)] are computed. The rectangle with minimal Bhattacharyya coefficient is selected as the final fitting result [red rectangle in (e)].
distribution, which is described as follows: the line pixels can guarantee more reliable ground orientation as background seeds, as shown in Fig. 7 (green line). The line pixels in the inner lower legs region with minimal average similarity with the lower body above a threshold along the torso line. Therefore, the shape and structure of segmentation can be effectively analyzed. In our method, MOH can obtain the missed parts and judge the integrity of the lower body, so that it is used to update Graph Cuts seeds. Lower-Body Segmentation: The inner blocks employing MOH, as shown in Fig. 8(c) (top), are classified into between-leg and in-leg blocks, as shown in Fig. 8(c) (bottom). The gray block is regarded as the between-leg block, which is selected under two evaluations. Symmetry rule: Notate the number of inner blocks as \( N \), the barycenter of the \( i \)-th \((i = 1, 2, \ldots, N) \) inner block as \( C_i \), and the midpoint on the hemline of the torso rectangle as \( C_T \). The line through \( C_T \) and \( C_i \) divides inner blocks into two pieces \( A_{i1} \) and \( A_{i2} \). The area evaluation is defined as \( AR_i = \frac{\min(A_{i1}, A_{i2})}{\max(A_{i1}, A_{i2})} \) (16). The symmetry term indicates that the between-block for background seeds should be through the vertical projection line. Scale rule: We constrain the scale of the between-block as a scale rule term. Notate the height of the \( i \)-th inner block as \( h_i \) and the distance from \( C_T \) to \( C_i \) as \( d_i \). The scale evaluation is defined as \( SR_i = \frac{h_i}{h_i + d_i} \) (17).

We select the inner between-leg block with the maximal sum of area and scale evaluations. Hence, we can update the foreground and background distributions using inner in-leg and between-leg blocks, respectively.

Denote the first lower-body segmentation result as \( TS \), the barycenter of \( TS \) as \( C_s \), and the inner between-leg block as \( N^* \). Notate the lower-body orientation as \( l \), which is adjusted by

\[ b_t(x) = \frac{1}{nh} \sum_{x_i \in B_t} K \left( \frac{x - x_i}{h} \right) \] (15)

where \( B_t \) is the background seed area, \( n \) is the pixel number, and \( K \) is the kernel function in (11). Thus, the \( t \)-links and \( n \)-links of lower-body segmentation can be set, and the coarse lower body is separated from background, similarly to the upper body.

B. MOH-Based Lower-Body Segmentation

In a coarse lower body, there are still many false negatives. Here, we introduce an iterative MOH algorithm to refine the figure/ground distributions, thereby obtaining fine results. MOH: First, we introduce the MOH projection basis utilized to improve segmentation. As shown in Fig. 8(a), the basis consists of a vector (red arrow line) with the orientation from the torso center pointing to legs’ and its perpendicular. MOH is used to describe the projection information of the coarse lower body, which can be used to find the false negatives, as shown in Fig. 8(b) and (e). Each bin of MOH represents multiple cues of coarse segment results: accumulation, span, number of line segments, and boundary points of figure/ground on each projection line. The accumulation refers to the number of all segmented pixels that divide the projection line into multiple segments in a given bin; and the span is defined as the length of a line segment. Therefore, the shape and structure of segmentation can be effectively analyzed. In our method, MOH can obtain the missed parts and judge the integrity of the lower body, so that it is used to update Graph Cuts seeds.

Symmetry rule: Notate the number of inner blocks as \( N \), the barycenter of the \( i \)-th \((i = 1, 2, \ldots, N) \) inner block as \( C_i \), and the midpoint on the hemline of the torso rectangle as \( C_T \). The line through \( C_T \) and \( C_i \) divides inner blocks into two pieces \( A_{i1} \) and \( A_{i2} \). The area evaluation is defined as

\[ AR_i = \frac{\min(A_{i1}, A_{i2})}{\max(A_{i1}, A_{i2})} \] (16)

The symmetry term indicates that the between-block for background seeds should be through the vertical projection line.

Scale rule: We constrain the scale of the between-block as a scale rule term. Notate the height of the \( i \)-th inner block as \( h_i \) and the distance from \( C_T \) to \( C_i \) as \( d_i \). The scale evaluation is defined as

\[ SR_i = \frac{h_i}{h_i + d_i} \] (17)

We select the inner between-leg block with the maximal sum of area and scale evaluations. Hence, we can update the foreground and background distributions using inner in-leg and between-leg blocks, respectively.
repeating MOH. The lower-body segmentation procedure based on MOH is organized in Algorithm 2.

**Algorithm 2: Lower-body segmentation**

1: Do the first segmentation: $TS$
2: Initialize lower-body orientation: $l \leftarrow \hat{C}_T C_S$

3: for $k = 1$ to $K$ do
4: Generate MOH on $l$
5: Find inner blocks
6: Select the inner between-leg block: $N^* = N_i \{i^* = \arg \max_i (AR_i + SR_i)\}$
7: Update seeds using inner between-leg and in-leg blocks

8: Update lower-body orientation: $l \leftarrow \hat{C}_T C_i$
9: end for
10: Do Graph Cuts and segment lower body

V. EXPERIMENTS AND RESULTS

**Data Set:** We have collected 197 real-world photo images with the size of $208 \times 156$ pixels, covering various individuals and appearance, different poses and illumination, and cluttered backgrounds. Some samples are in Fig. 9.

**Parameters:** In our experiments, the scale of the bounding boxes used for CT detection is set to three and four times the
width and height of the detected face, respectively. The radius of the circle outside which the pixels are set as background seeds is 3 and 3.5 times as the width of the fitted torso rectangle for upper-body and lower-body segmentations, respectively. The weighting terms $\alpha$ and $\beta$ used to estimate the torso area probability are set to 0.8 and 0.5, whereas the parameter $\gamma$ for location probability is set to 4, and factor $\lambda$ is set to 0.4 to balance them. The width of kernel function $h$ to calculate figure/ground distributions is 5.

**Evaluations:** Our current implementation with C++ runs about 1 min per image on a 2.70-GHz machine. The proposed method can automatically recover the human body region with a detected face.

Ramanan [38] proposed an iterative parsing (IP) process to estimate articulated body pose. Although there is no explicit segmentation of the foreground and background, the soft segmentation results can provide the figure/ground distributions. We treat the iterative Graph Cuts algorithm employing the distributions derived from the IP method (we denote IPGC) to segment images as a baseline to compare with ours on the collected pool, as shown in Fig. 9. Overall, our method is able to segment out humans in cluttered background. The IPGC method poorly performs when a human subject is in cluttered background, because the IP method depends on human edge information, which is difficult to extract from the noisy scene. In addition, much background is often segmented out along with human, and some important parts are often missed (e.g., heads or arms).

For quantitative evaluation, there are three rules on the performance. The evaluations are carried out by comparing the recovered human body region with ground truth. Sinop and Grady [39] introduced a normalized overlap defined as evaluation one, which can be regarded as a global similarity between the segmentation result and the ground truth (that are manually collected). However, the rule does not consider the segmentation result how much ground truth is covered and how much background is contained. Therefore, we define another two rules to evaluate the missed ground truth and the background contained rates on the segmentation results, respectively.

**Evaluation One:** The similarity between the segmentation and the ground truth is defined as the proportion of pixels correctly segmented as foreground or background by comparing with ground truth binary results [39]

$$\text{Eval}_1 = \frac{|S \cap G|}{|S \cup G|}$$

(Eq. 18)

where $S$ is the set of pixels within the segmentation, and $G$ is the set of foreground pixels in ground truth.

**Evaluation Two:** To evaluate how much the parts are missed, we define the second evaluation as

$$\text{Eval}_2 = \frac{|G \cap \bar{S}|}{|G|}$$

(Eq. 19)

where $\bar{S}$ is the complementary set of $S$.

**Evaluation Three:** The third evaluation, which is to evaluate the background contained in the segmentation result, is defined as follows:

$$\text{Eval}_3 = \frac{|S \cap \bar{G}|}{|S|}$$

(Eq. 20)

where $\bar{G}$ is the complementary set of $G$.

The evaluations on the segmentation results are shown in Fig. 10. We achieve a stabler and more accurate performance than IPGC with less missed part and background. Table I shows...
the statistics of the evaluations on IPGC and our method. Our human body segmentations can overlap 83% of ground truth and contain only 4% background, and about 14% is regarded as background on average. On the other hand, the IPGC method achieves 61%, 30%, and 18% correspondingly. However, the main reason for regarding background as foreground by 14% is that our scheme ignores the hair, shoes, and straps, and we occasionally lose some bared half-limbs. We believe that the overlapped percentage will be higher, if we employed the part detector in [23] based on our results to detect the lost half-limbs. Overall, the evaluations indicate that the proposed algorithm is very effective and robust.

We compare our body part segmentation with the part detectors used in [23]. As shown in Table II, our performance is better than that of the method proposed in [23] validated on our own data set. First, we find out 15.09% more half-limbs than their method, as we focus on the top-down segmentation scheme whereas Normalized Cuts usually do not segment half-limb accurately. Second, there are 12.15% more torsos we found. Because we employed MCTD to detect torsos by grouping blocks with region and boundary cues, whereas [23] exhaustively search torso candidates and score the candidates using four low-level cues, as well as relationship between head and torso candidates.
In addition, we test the impact of four important parameters, which are used in the proposed algorithm on our whole segmentation system. As shown in Table III, we change one parameter (e.g., \( \alpha \)) from 0.1 to 0.9 with other three fixed parameters (e.g., \( \beta, \lambda, \) and \( \gamma \)) and then obtain five sets of the mean accuracy on the whole data set. From each row, we can see that segmentation performance is stable.

We also perform experiments on other public data sets VOC 2006 and VOC 2010, as shown in Fig. 11. Although these two data sets contain several class objects, we only perform our method on the frontal person images. From the results, we can see that our method is effective to recover the human body region in some more challenging images, e.g., part occlusion and intense illumination.

VI. CONCLUSION

In this paper, we have proposed an efficient coarse-to-fine segmentation-based approach to automatically recover human body in static photo image, which is still one of the most challenging tasks in the computer vision field. The main contributions of this work have been drawn as follows: 1) We have proposed a straightforward segmentation-based framework for recovering human body from a static image. 2) We have presented the MCTD to detect torso. 3) We have introduced a robust iterative MOH algorithm to recover lower-body segmentation. The experiments on our data set, i.e., VOC2006 and VOC2010, show that we can exactly recover human body from still images with face priori.

Our proposed method is still very simple as we used a face detector to locate head position currently. For the future work, we will relax the algorithm to deal with variable face orientations, even in the case that the face is not possible to be detected by general face detectors.

REFERENCES


Huchuan Lu (M’06) received the M.Sc. degree in signal and information processing from Dalian University of Technology (DUT), Dalian, China, in 1998. He is currently a Professor with School of Information and Communication Engineering (DUT), becoming an Associate Professor in 2006. His research interests are computer vision and pattern recognition, including visual tracking and segmentation. Mr. Lu is a member of the Association of Computing Machinery.

Guoliang Fang received the B.E. degree and the M.Sc. degree in signal and information processing from Dalian University of Technology (DUT), Dalian, China, in 2008 and 2011, respectively. He is currently with the School of Information and Communication Engineering, DUT. His research interests include human segmentation and interactive image segmentation.

Xinqing Shao received the B.E. degree from AnQing Teachers College, AnQing, China, in 2008. He is currently working toward the M.Sc. degree in signal and information processing with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. His research interests include human segmentation, pose estimation, human computer interaction, and object recognition.

Xuelong Li (M’02–SM’07–F’12) is currently a Researcher (Full Professor) with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China.