Pattern Recognition Letters 30 (2009) 503–512

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Recovery of upper body poses in static images based on joints detection

Zhilan Hu [a,b,*], Guijin Wang [a], Xinggang Lin [a], Hong Yan [b,c]

[a] Department of Electronic Engineering, Tsinghua University, 9-306, East Main Building, Beijing 100084, China
[b] Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, China
[c] School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 30 September 2007
Received in revised form 16 September 2008
Available online 24 December 2008

Communicated by H.H.S. Ip

Keywords:
Pose estimation
Markov chain Monte Carlo
Torso detection
Gaussian mixture model

ABSTRACT

Recovering human body poses from static images is challenging without prior knowledge of pose, appearance, background and clothing. In this paper, we propose a novel model-based upper poses recovery method via effective joints detection. In our research, three observables are firstly detected: face, skin, and torso. Then the joints are properly initialized according to the observables and some heuristic configuration constraints. Finally the sample-based Markov chain Monte Carlo (MCMC) method is employed to determine the final pose. The main contributions of this paper include a robust torso detector through maximizing a posterior estimation, effective joints initialization, and two continuous likelihood functions developed for effective pose inference. Experiments on 250 real world images show that our method can accurately recover upper body poses from images with a variety of individuals, poses, backgrounds and clothing.

1. Introduction

Recovering human body poses is needed in many computer vision tasks, such as human actions understanding, content-based image retrieval and figure pose classification. It remains a challenging problem however, due to cluttered backgrounds, body posture variations, different individuals, various lighting conditions and occlusions.

Recently, many methods have been developed to estimate human body poses from video sequences in a controlled environment (Agarwal and Triggs, 2006a; Wang and Leow, 2006; Boulay et al., 2006; Sigal et al., 2004; Lee and Nevatia, 2007; Singh et al., 2005; Menier et al., 2006; Fan and Wang, 2004; Sminchisescu and Telea, 2002; Gupta et al., 2008), and especially, body skeletonization (Singh et al., 2005; Menier et al., 2006; Fan and Wang, 2004; Sminchisescu and Telea, 2002) has received considerable attention from pose recovery. However, these methods are performed on human regions (image silhouettes) which are segmented by using background model and motion information which make them infeasible for static images.

For static images, the existing poses recovery methods can be roughly classified into three categories: matching-based (Mori and Malik, 2002, 2006; Agarwal and Triggs, 2004, 2006; Shakhnarovich et al., 2003; Poppe and Poel, 2005), part-based (Mori et al., 2004; Ren et al., 2005; Hua et al., 2005; Roberts, 2005) and model-based methods (Lee and Cohen, 2006; Deva and Cristian, 2006). In matching-based methods, human pose is estimated by comparing the human body features (such as Fourier descriptors (Poppe and Poel, 2005) and shape contexts (Mori and Malik, 2006)) in a tested image with those in a large set of labeled images. These techniques cannot achieve accurate results for images with complex backgrounds. To handle complex backgrounds, Agarwal and Triggs (2006b) modeled the image with a dense grid of local gradient orientation histograms, which requires the human body in the center of the image. Furthermore, such matching-based methods require a large set of training images covering various poses, backgrounds and individuals.

To overcome the limitation of matching-based methods, part-based approaches have been developed. These approaches typically solve the pose recovery problem by detecting the candidates of each body part (such as face, torso and limbs) and inferring the best assembly according to some predefined configuration constraints. Compared to matching-based methods, part-based approaches need smaller training sets and are less sensitive to pose varieties, while their performance depends on the part detectors and they can only deal with special clothing styles. For example, Mori et al. (2004) trained part detectors (limb and torso) by using some special features from baseball players images, but such detectors may not work well for other clothing styles. Ren et al. (2005) detected candidate body parts using parallelism, which is effective for images with simple clothing and backgrounds, such

* Corresponding author. Address: Department of Electronic Engineering, Tsinghua University, 9-306, East Main Building, Beijing 100084, China. Tel.: +86 10 62781291; fax: +86 10 62770317.
E-mail address: huzhilan00@mails.tsinghua.edu.cn (Z. Hu).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.12.005

as gymnastic player and skater, but may fail when lots of edges appear in the background or clothing. For humans with special clothing styles similar to that of a soccer player's, Hua et al. (2005) employed detected skin regions to estimate candidates for the lower arms and thighs, and used line segments to generate torso candidates, thus their performance may be poor when dealing with complex backgrounds and other clothing with different styles.

Unlike the above two categories, model-based methods firstly generate a large number of pose hypotheses by changing the human model parameters, then recover the human pose by minimizing the projection errors between the pose hypotheses and the image. Such methods are less sensitive to part detectors and can achieve accurate poses with proper initialization and fine sampling. To generate pose states, Lee and Cohen (2006) employed the framework of the Data-Driven Markov chain Monte Carlo (DDMCMC) method to estimate human poses. This technique can deal with some images in real world, but initializing all people in a standard upright pose increases its computation cost, also the likelihoods based on discrete distributions and the lack of more reliable observables limit its performance.

We believe that reliable observables benefit accurate pose initialization, and heuristic joints initialization can reduce computational cost and improve the accuracy of pose recovery. For photos of daily life taken in real world, we propose a new method to recover upper body poses via effective joint initialization. Firstly, three observables are detected: face, skin and torso. According to the observables, the joints are properly initialized with some heuristic configuration constraints. Finally the sample-based Markov chain Monte Carlo (MCMC) method Gilks et al., 1996 is employed to generate the final pose.

Experiments show that our method can recover upper poses accurately from images with different poses, backgrounds, lighting conditions and clothing. The main contributions of this paper include:

(1) A model-based torso detection method is developed to accommodate different poses and image conditions. It locates torsos accurately from different images and benefits the pose initialization by providing the coarse locations of shoulders, hips and neck.
(2) An effective pose initialization method is proposed to reduce the computational cost and improve the pose recovery accuracy. Instead of initializing the pose uniformly (such as initializing each pose as standard upright), we find the coarse positions of the joints from detected parts to provide more accurate initialized pose.
(3) Two likelihoods based on continuous distribution, the modified skin likelihood and the foreground and background likelihood, are explored to improve the robustness to different images during the pose inference process.

The rest of the paper is organized as follows: Section 2 presents the framework of our proposed system. The details of torso estimation, joints initialization and pose inference are discussed in Sections 3–5, respectively. Section 6 shows the experiment results. Finall1y, we conclude the paper with Section 7.

2. System overview

Shown as Fig. 1 , our method includes three stages: observables detection, joints initialization and pose inference. Three observables are detected for joints initialization: face, skin and torso. The face is located by an AdaBoost-based face detector (Viola and Jones, 2004) (Fig. 1a). Skin, which is important for the estimation of limbs, is segmented using the Markov Random Field (MRF) based method (Chenaoua and Bouridane, 2006) (Fig. 1b). The torso is the key part connecting to most other body parts and is relatively stable with the detected face, so a robust generative torso detector is developed (Fig. 1c). With the observables, the joints initialization stage is completed as follows: The shoulders, hips and neck are directly located by the detected torso. Assuming the hand skin is visible, the hand candidates are determined by the detected skin regions. Unlike previous works which employ arm detectors (Mori et al., 2004; Ren et al., 2005; Hua et al., 2005; Lee and Cohen, 2006) to estimate elbows (which are difficult to handle various backgrounds, clothing and poses), we initialize the elbows with some heuristic constraints and image color distributions of the foreground (Fig. 1e) and background (Fig. 1f). The coarse hand position is also determined from the hand candidates along with the elbow initialization. In most cases, the initialized joints are not accurate but close to the actual locations. Thus the final pose is inferred by the sampling-based MCMC method, where the proposal distribution is based on the *Random-walk sampler* (Gilks et al., 1996) which is easily sampled and evaluated. Each state is evaluated by four likelihood functions and updated using the Metropolis–Hastings algorithm (Gilks et al., 1996). In our scheme, the scale parameters in the proposal distributions are determined experimentally based on two criteria: the parameters are small so that the points in the neighborhood are close to each other, and they can provide effective initialization to estimate the pose quickly.

3. Torso estimation

A deformable model is designed to represent the torso. As illustrated in Fig. 2a, it is a rectangle cut by a semi-circle and two small triangles. The main advantage of the model is that it guarantees that the torso is mostly covered by one piece of clothing, thus it alleviates the influence of non-clothing pixels during the torso estimation due to different collar styles and diverse poses. By modifying the parameters: torso width w, torso height h, torso inclining orientation θ, and the neck position in the image (x_0, y_0), this model is able to depict different torsos.

Using the torso model, we formulate the torso estimation as maximum a posteriori (MAP) estimation:

$$X_{\mathrm{MAP}} = \arg\max_x p(x|Y) \propto \arg\max_x p(Y|x)p(x), \tag{1}$$

where x is the parameter vector of a torso hypothesis, $x = \{x_0, y_0, w, h, \theta\}$; $p(x|Y)$ is a posteriori probability distribution, evaluating each torso hypothesis x given the image observation Y; $p(x)$ is the prior distribution standing for the prior structure of the torso; and $p(Y|x)$ is the likelihood function measuring how well a torso hypothesis x fits the image observation Y.

3.1. Prior distribution

The prior distribution is composed of two parts. One is the distance from the torso center to face center $d(x)$. Normalized by the corresponding face width, $d(x)$ is relatively stable, thus in our model its distribution is represented by a single Gaussian $p_d(x)$ centered on d_0 with the covariance Σ_d, shown in Eq. (2). The other is the torso appearance $s(x)$, consisting of the torso width w and torso area R. Its distribution is also depicted by a single Gaussian $p_s(x)$ centered on s_0 with the covariance Σ_s, shown in Eq. (3). For robustness to different individuals, w is normalized by the face width and R is normalized by the face area. Finally, we can obtain the prior distribution in Eq. (4).

$$p_d(x) = N(d(x), d_0, \Sigma_d), \tag{2}$$
$$p_s(x) = N(s(x), s_0, \Sigma_s), \tag{3}$$
$$p(x) = p_d(x)p_s(x). \tag{4}$$

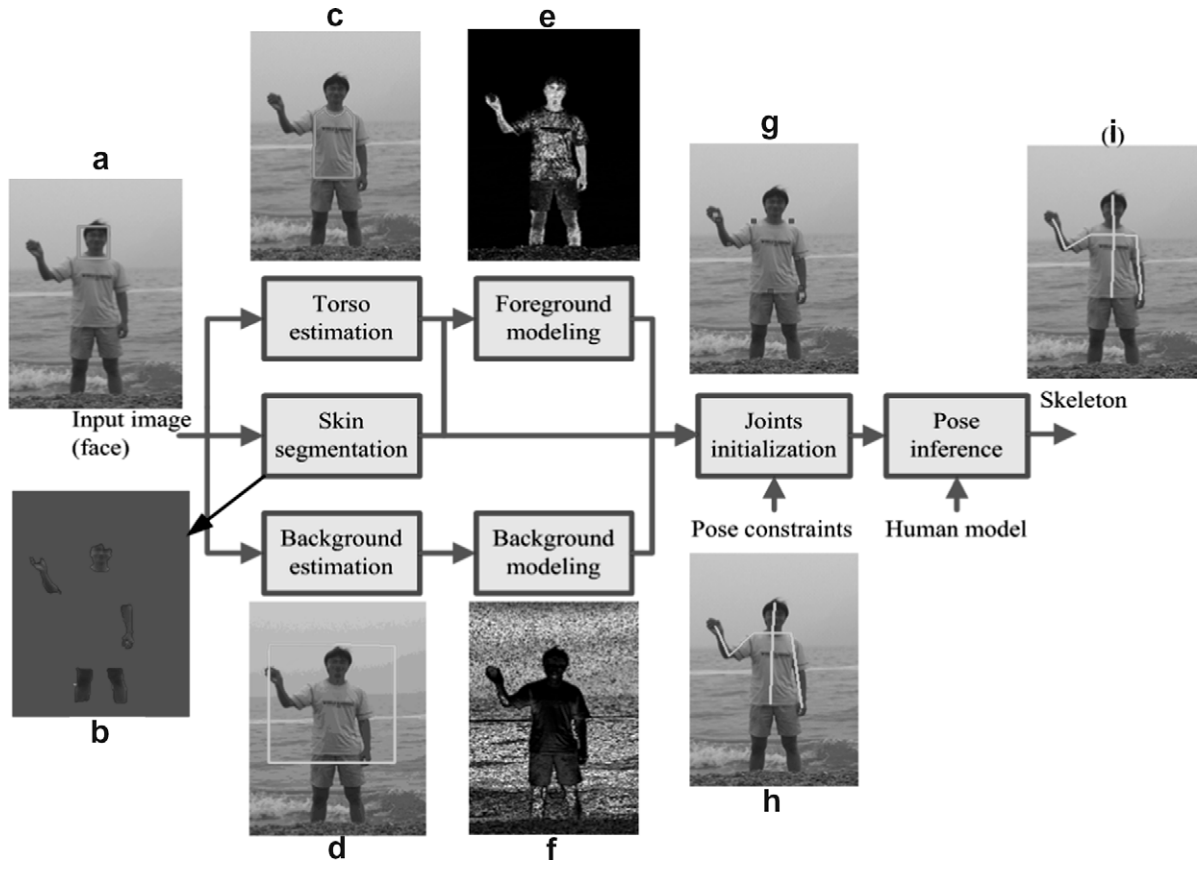

Fig. 1. Flowchart of the proposed upper body poses recovery system. (a) Input image, (b) segmented skin regions, (c) estimated torso (the red curve), (d) outer torso boundary (the green rectangle), (e) foreground color distribution, (f) background color distribution, (g) initialized joints, (h) initialized skeleton, (i) final skeleton. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

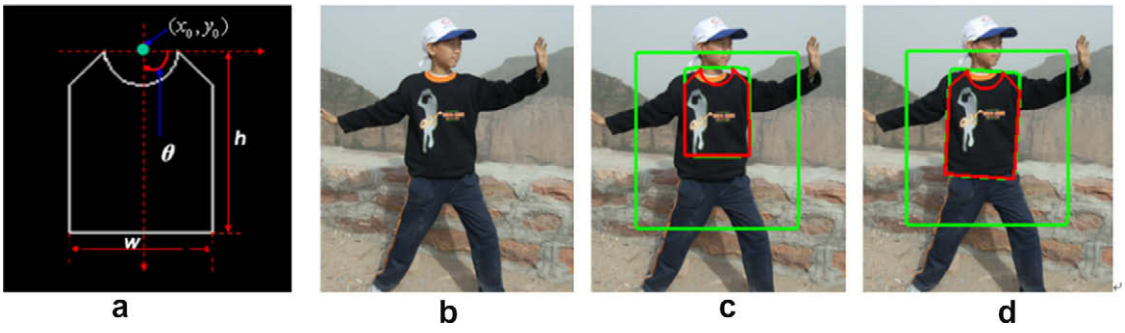

Fig. 2. Deformable torso model and the process of torso estimation. (a) Deformable torso model, (b) input image, (c) initialized torso and (d) estimated torso.

where d_0 indicates the average normalized distance from the face center to torso center. $s_0 = (w_0, R_0)$, w_0 represents the average normalized torso width, and R_0 is the average normalized torso area. All parameters in Eqs. (2) and (3) are derived from the training images.

3.2. Image likelihood function

The likelihood function requires reliable features to obtain optimal solutions. Compared to local features such as edge and texture, holistic features are more stable and effective. The color likelihood, which measures the difference between the foreground (torso region T_x) color distribution and background ($\bar{T}_x$) color distribution, is demonstrated in Zhao and Nevatia (2002) to be a good global criterion and thus adopted here. When the torso hypothesis is correct, the color distribution distance between the foreground and background is expected to be large. In our method, each color distribution is constructed by a normalized histogram with $N_{histogram}$ bins, and the color likelihood function is defined as:

$$p(Y|x) = \exp\left(-\alpha_{color} \sum_{i=1}^{N_{histogram}} \sqrt{t_i b_i}\right), \tag{5}$$

where t_i and b_i are corresponding values of the ith normalized histogram bin for the foreground and background, respectively, and α_{color} is a weighting constant set to 2 experimentally.

In this likelihood function, the foreground region is the predicted torso, and the background is determined as follows. As the torso is often small in the image, if the background is considered as the whole image excluding the predicted torso, the color likelihood may be insensitive to torso variations. So, we limit the

background by defining an outer torso boundary based on the geometric relationship between the torso and face. Additionally, in order to eliminate the ambiguity of whether they belong to the background or foreground, the pixels around the neck and shoulders are ignored. Taking Fig. 2(c) for example, the large green rectangle is the outer torso boundary, the area between two green rectangles is the background used for the color likelihood computation, and the region enclosed by the red curve is the foreground.

3.3. Torso estimation and results

Finally, the optimal solution is obtained by solving a sampling-based MCMC problem, where the proposal distribution is based on *Random-walk sampler*, and each state is updated using the Metropolis–Hastings algorithm. Using the MCMC, the algorithm converges within a small number of iterations. Taking Fig. 2 for example, (d) shows the predicted torso after 10 iterations. The result is much closer to the ground truth than the initialized torso in (c).

This method is able to locate torsos with various appearances and different backgrounds, shown as the examples in Fig. 3.

4. Joints initialization

We use a 2D skeleton to describe the upper poses. The skeleton includes 11 joints: top head, neck, left/right shoulder, left/right elbow, left/right wrist, left/right hand and middle hip. The coarse positions of neck, shoulders and hips are located by the estimated torso. The hand candidates are provided by the segmented skin regions and evaluated by an approximate local posterior probability distribution, through which the elbow is initialized as well.

4.1. Hand candidates determination

The hand should be in a limited distance to the face (the distance is normalized by the corresponding face width), so we firstly remove some skin regions far away from the face, and then detect the hand candidates by clustering the remaining skin regions into two categories. One is the skin region properly fitted by ellipses, and another is those that cannot be fitted by ellipses.

For the first category, the hand candidates are estimated using the major points as well as the major axis of the fitted ellipse by classifying them into one of the following cases:

- If the skin region is small and the length ratio of the major axis to the minor axis is below a threshold, it is considered to be a hand, e.g. the hands in Fig. 4a.
- The skin regions not satisfying case 1 are considered as the lower arms or full arms. We take the ellipse's major points as the hand candidates, e.g. Fig. 4b and c.

The second category always represents a crooked arm (Fig. 5b) or connected arms (Fig. 5a and c), where the corners often indicate the elbow or wrist. Here, we use the Delaunay triangulation (DT) to detect the corners as wrist candidates, from which the hand candidates can be easily obtained as the hand is near the wrist.

The Delaunay triangulation is the unique triangulation of a set of vertices in the plane such that no vertex is inside the circumcircle of any triangle (http://www.geom.uiuc.edu/samuelp/del_project.html). It is widely used in geometric modeling and finite element analysis. Recently, it has also been applied to object extraction (Xiao and Yan, 2004), character skeletonization (Zou

Fig. 3. Examples of torso estimation result.

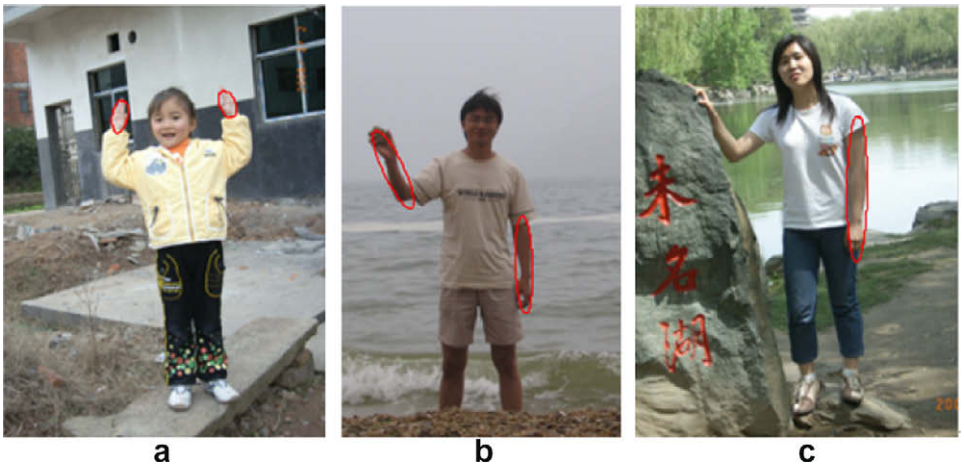

Fig. 4. Examples of skin regions (enclosed by red ellipses) belonging to category 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

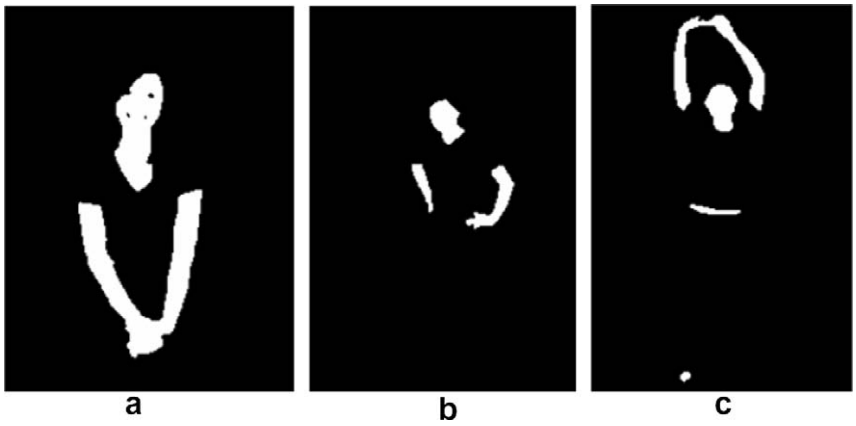

Fig. 5. Examples of skin regions belonging to category 2.

et al., 2001), clothing segmentation (Hu et al., 2008), and so on. By employing the concept of triangle chain (Xiao and Yan, 2004), the corners are found as follows:

Firstly, the skin region boundary (Fig. 6b) is extracted by the *Canny* operator. Taking all boundary pixels as the vertices set, the corresponding DT is generated by TRIANGLE (Shewchuk, 1996) (Fig. 6c). Then, the constructed DT is decomposed into a number of triangle chains. A triangle chain is a set of non-shortest edges shared non-skin triangles, of the form $T, L, \ldots L, T$, where T and L stand for a terminal triangle and linking triangle respectively, e.g. Fig. 6d. Obviously, the longest triangle chain is most probably the link to a corner, so we find the longest triangle chain and extract its axis that connects all centers of each non-shortest edge of the chain (blue line in Fig. 6e). Lastly, the skin region is split by the axis and the center of the axis segment across the skin region is considered as one corner, e.g. Fig. 6f. For the other corners, they are detected from the split regions recursively by the above process, e.g. (Fig. 6)(g).

4.2. Coarse hand position determination and elbow initialization

Each corner detected above is probably a wrist. However, it is difficult to confirm which corner is the correct one. By virtue of the configuration constraints between hand, wrist and shoulder,

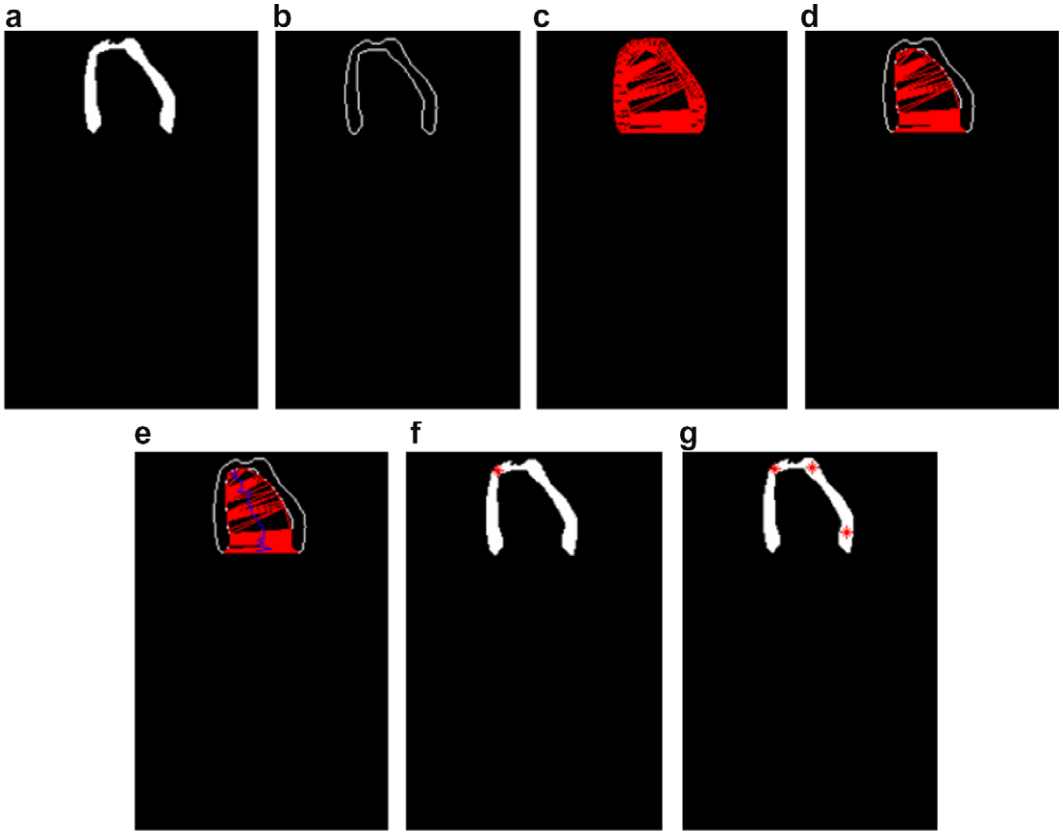

Fig. 6. The processing flow for corner detection. (a) A skin region, (b) skin region boundary, (c) the Delaunay triangulation of the boundary, (d) the longest triangle chain, (e) the axis of the triangle chain (the blue line), (f) the corner detected based on (e), shown as the red point, (g) all the corners detected by our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

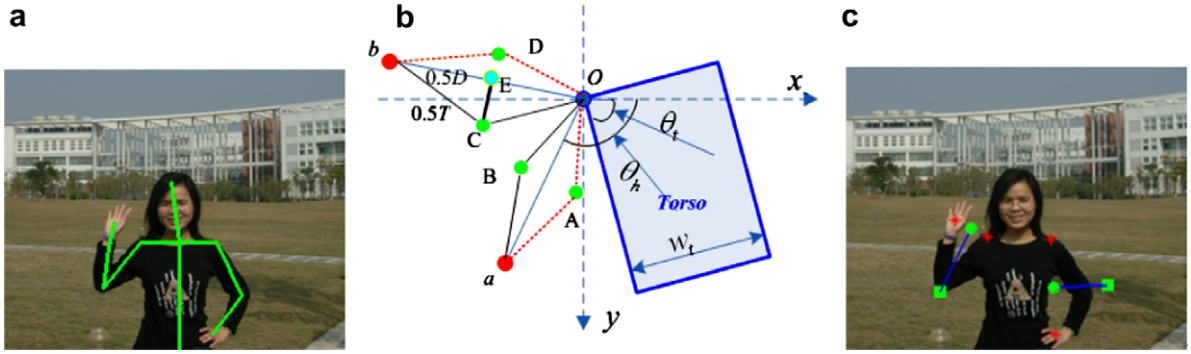

Fig. 7. Configuration constraints between shoulder, hand and elbow. (a) An example, (b) illustration for the pose constraints and (c) elbow solution spaces.

the most probable hand candidate can be determined using the color distributions of the foreground and background, also the elbow can be initialized at the same time.

4.2.1. Configuration constraints between hand, elbow and shoulder

Given the hand and shoulder, the elbow should satisfy some natural constraints. For example, in Fig. 7a, the elbow can move only within a limited range relative to the shoulder and hands. Here, we explore three constraints to provide the solution space for the elbow. Illustrated as Fig. 7b, where the blue rectangle stands for the torso, the blue point denotes the right shoulder, the red points indicate the positions of the right hand, and the other six parameters are explained in Table 1, three constraints are explored as follows (taking the line connecting the shoulder and the hand as marginal line, the area includes the torso or the upper part of the torso is called *inside*, and the another area is called *outside*):

- When $\theta_h - \theta_t \leq \theta_{\min}$, such as the hand at position a, the elbow should be *outside* the line Oa, e.g. position B but not A.
- When $\theta_h - \theta_t \geq \theta_{\max}$, such as the hand at position b, the elbow should be *inside* the line Ob, e.g. position C but not D.
- When $\theta_{\min} < \theta_h - \theta_t < \theta_{\max}$, we take both sides as the solution space.
- $\theta_{\min}$ and $\theta_{\max}$ are experimentally set to 45° and 60°, respectively.

Thus, the solution space for the elbow can be estimated. Given the shoulder (O) and hand (b), we can draw an isosceles triangle ObC with the base as line Ob, and the lengths of the two other sides equal to $0.5T$, the altitude of the base (line CE) is considered as the solution space. Fig. 7c shows a real example, where the red asterisks are the positions of hands and shoulders, the green points are the center between hands and shoulders, the green squares are the vertices calculated by the configuration constraints, and the blue lines are the solution spaces for the elbows initialization.

4.2.2. Coarse hand position determination and elbow initialization

To evaluate each possible elbow position in the solution space and each hand candidate, the color distributions of the foreground and background around the arm skeleton are employed: when the predicted elbow is accurate, the corresponding arm region should have high foreground probability and low background probability, and contain as many skin pixels as possible.

Composed of clothing and skin, the foreground color distribution is constructed by skin color distribution P_{skin} and clothing color distribution P_c. P_{skin} is described by a single Gaussian probability distribution in HCrL space (Sawangsri et al., 2005) using the detected face. While is represented by a Gaussian mixture model (GMM) in RGB space based on the estimated torso as the clothing always consists of more than one color. Finally, the foreground color distribution P_f is determined by the following equation:

$$P_f(z_n) = \max(P_c(z_n), P_{skin}(z_n)). \tag{6}$$

On the other hand, the background is roughly determined by the pixels outside the outer torso boundary, and its color distribution P_b is also modeled by a GMM in RGB space.

With the color distribution of the foreground and background, the most likely arm skeleton can be found by maximizing approximate local posterior probability L, which is defined as:

$$L = \sum_{z_i \in S} (P_f(z_i) + P_{skin}(z_i) - P_b(z_i)). \tag{7}$$

where $\boldsymbol{S}$ is the set of pixels around the arm skeleton, and z_i is a pixel in $\boldsymbol{S}$. L emphasizes the skin probability because of its relative stability as well as its significance in the arm.

Table 1
The meaning of six parameters used in Fig. 7b.

Parameter	Meaning
θ_t	Torso inclination
θ_h	The angle of the hand with the corresponding shoulder
w_t	Torso width
w_f	Face width
D	The distance between the right hand and shoulder, normalized by w_f
T	The maximum value of D, it is learned from the data in (NIST, 1977)

5. Pose inference

Similar to torso estimation, the pose inference problem is formulated as an MAP estimation from Eq. (1), with x as the pose parameter vector including all upper body joints positions in the image. It is also solved by the sample-based MCMC technique. The prior distribution $p(x)$ and likelihood function $p(Y|x)$ in the pose inference are introduced as follows.

5.1. Prior distributions

The parameters used for the prior distribution consist of two subsets: local joints angles and human parts lengths. These two subsets are almost independent, so the prior distribution is approximated by Eq. (8), where $p(j)$ is the joints angles prior distribution, and $p(l)$ is the lengths prior distribution, and we define

$$P(x) \approx p(j)p(l). \tag{8}$$

The joint angles for the prior distribution are composed of seven triples of neighboring joints: {left shoulder, left elbow, and left wrist}, {right shoulder, right elbow, and right wrist}, {left hip, left shoulder, and left elbow}, {right hip, right shoulder, and right elbow}, {left elbow, left wrist, and left hand}, {right elbow, right

wrist, and right hand} and {top head, neck and middle hip}. For each triple-wise, the first and last joint are defined as the father ($j_{father(i)}$) and son ($j_{son(i)}$) of the middle joint j_i) respectively, each corresponding angle is formed by the line from its son to itself and that from itself to its father. As the angle often occurs uniformly within a limited range, we use the uniform distribution $u(j_{father(i)}, j_i, j_{son(i)})$ to describe its prior distribution. $p(j)$ is approximated as:

$$P(j) \approx \prod_i u(j_{father(i)}, j_i, j_{son(i)}), \tag{9}$$

where $j = \{j_i, i = $ *left elbow, right elbow, left shoulder, right shoulder, left wrist, right wrist, neck*$\}$. The range of the first four uniform distributions is $[0, \pi]$, and the last three are $[0.5\pi, \pi]$.

The lengths subset l consists of seven parameters, l_{LWE}, l_{RWE}, l_{LES}, l_{RES}, l_{HN}, h_t and w_t, standing for the length from left wrist to left elbow, from right wrist to right elbow, from left elbow to left shoulder, from right elbow to right shoulder, from top of head to neck, torso height, and torso width respectively. All the parameters are normalized by the face width in each image for robustness to individuals. The lengths prior distribution is approximated by a Gaussian distribution:

$$P(l) \approx N(l, \mu_l, \Sigma_l), \tag{10}$$

where μ_l and Σ_l are the mean and covariance matrices respectively of the Gaussian distribution.

5.2. Likelihood function

The objective of the likelihood function is to discriminate the poses which fit the image well and those that do not. Recently, several likelihood measurements based on discrete color distributions have been proposed and demonstrated to be effective for some images with distinct color differences between the human bodies and backgrounds, such as region likelihood, color likelihood and skin likelihood (Lee and Cohen, 2006). However, discrete color distributions may be disturbed by cluttered backgrounds. So, besides the region likelihood L_{region} and color likelihood L_{color} following the definitions in (Lee and Cohen, 2006), two more likelihoods based on continuous distribution are explored in our work to improve the robustness. One is the modified skin likelihood, and the other is the foreground and background likelihood.

Different from the skin likelihood used in (Lee and Cohen, 2006), where a clothing model is employed to predict the skin region and the likelihood is evaluated by skin color histogram model, our skin likelihood is defined based on the skin color Gaussian distribution, without any clothing model. Considering that the skin pixels always lie in arms and head, we calculate the skin likelihood L_{skin} using the estimated arms and head as Eq. (11)

$$L_{skin} = \alpha_s \sum_{z_n \in S} P_{skin}(z_n)/N_{image}, \tag{11}$$

where $\boldsymbol{S}$ is the set of all pixels in the estimated arms and head, N_{image} is the image size (namely, the total number of the image pixels), z_n is a pixel in $\boldsymbol{S}$, P_{skin} is the skin probability obtained in Section 4.2.2, and α_s is set to 2, a constant determined experimentally.

Cluttered backgrounds may result in incorrect image segmentation or over-segmentation, and then reduce the effectiveness of the region likelihood based on segmentation results. To handle this problem, we employ the foreground and background likelihood L_{fb} built on the continuous color probability distributions of both foreground and background. If a pose candidate is near the ground

Fig. 8. Successful examples of upper poses recovery.

truth, the corresponding upper body region should have high foreground probability and low background probability. Thus L_{fb} is calculated according to the following equation:

$$L_{fb} = \left(\alpha_f \sum_{z_n \in F} P_f(z_n) + \alpha_b \left(\sum_{z_m \in B} P_b(z_m) - \sum_{z_n \in F} P_b(z_n) \right) \right) \Big/ N_{image}, \quad (12)$$

where P_f is the foreground color distribution, P_b is background color distribution, $\boldsymbol{F}$ is the set consisting of all pixels in the estimated upper body region, $\boldsymbol{B}$ is the set including all pixels in the background, z_m is a pixel in $\boldsymbol{B}$, and z_n is a pixel in $\boldsymbol{F}$. This likelihood guarantees that most foreground pixels are included in the estimated upper body region and most background pixels are excluded. Parameters α_b and α_f are constants, determined experimentally as 0.05 and 1, respectively.

The concept of the foreground and background likelihood is used in Zhao and Nevatia (2002), in which the foreground and background likelihood is defined based on the segmented human regions (namely foreground images), and its function is similar to the region likelihood used in our work. However, in our definition, it is unnecessary to segment the human body. Moreover, the likelihood is computed based on continuous distributions instead of binary ones for segmented (binary) images. Thus, our method can improve the pose estimation performance significantly as demonstrated in Section 6.

Finally, the image likelihood function $P(Y|x)$ is determined as:

$$P(Y|x) = L_{region} \times L_{color} \times L_{skin} \times L_{fb}. \quad (13)$$

Table 2
Average 2D error of key joints. Unit: pixel.

Joint	Average 2D error	Joint	Average 2D error
Left wrist	5.28	Left shoulder	6.24
Right wrist	5.76	Right shoulder	6.82
Left elbow	7.40	Middle hip	6.64
Right elbow	8.14		

6. Experiment results

We have collected 290 real world images with sizes around 250×190 for the experiments, covering various poses, backgrounds, clothing, individuals and lighting conditions. We use 40 images for learning the parameters used in torso estimation and pose inference, and the rest 250 images for testing. Fig. 8 shows some successful results of pose recovery, where the poses are different from each other, including standing (*a–n*), twisting (*o–p*), walking (*q*), exercising (*r*), running (*s*) and sitting (*t–u*). The backgrounds are also diverse, and even in some images the background colors are highly similar to the foreground in colors (e.g. (*a, h, i*)); the clothing ranges vary from coat (*a, d*), shirt (*b, c*), T-shirt (*g*), sleeveless (*e*) to sweater (*i, m*). The illumination in some images varies drastically (e.g. (*a, e, i*)). Besides, self-occlusion happens in the left arm in (*l*), and the body in (*k*) is occluded by a bag. Nevertheless, all poses are successfully recovered by our method.

We manually annotate each joint position in the collected images and compute the image distance error of the estimated joint positions. Table 2 shows the average error of the key joints (also the most difficult joints to be estimated). Compared with the overall average RMSE of 16.21 pixels examined on 30 test images in (Lee and Cohen, 2006), our method shows much smaller error (overall average error is 6.53 pixels). Furthermore, we only

Fig. 9. Comparisons of pose inferences with/without the foreground and background likelihood.

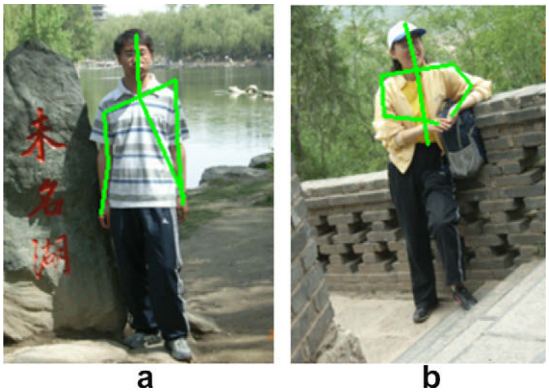

Fig. 10. Examples of the incorrect estimated poses.

take 250 Markov chain iterations during the pose inference in our experiments, while 1000 iterations are needed in (Lee and Cohen, 2006). Although Lee and Cohen, 2006 recovers the whole body (including 15 body joints), our method only focuses on the upper body (including 11 body joints), the experiment shows to a certain degree that our method is more effective in average RMSE and convergence rate, benefiting from effective parts detection, proper joints initialization, and efficient likelihoods.

To validate the foreground and background likelihood function in pose inference, a comparison experiment is conducted by inferring poses in two different ways: one excludes the foreground and background likelihood, denoted as Est1, and the other includes it, denoted as Est2. The image is segmented by a color based segmentation method (Ye et al., 2004).

Fig. 9 illustrates the comparison results, where the first column are the initialized poses, the second column are the segmented regions, the third column are the foreground color distributions, the fourth column are the background color distributions, the fifth column are the results by Est1, and the last column are the results by Est2. In (a), the T-shirt and part of the pants are segmented as one region (see Column 2) due to similar colors, which makes the estimated position of the middle hip far away from the ground truth in Est1 (see Column 5). The belt in (b) divides the clothing into two different regions (see Column 2), which makes the torso incline towards left upper region in Est1 (see Column 5). In (c), the left elbow is wrongly initialized and difficult to converge into the true position by Est1. All above problems can be solved by Est2, shown as the last column in Fig. 9. The experiments show that the proposed foreground and background likelihood is robust for poor image segmentation and joints initialization, and thus lead to a good pose recovery performance.

However, our method may not work well in the following cases.

(1) The background is almost identical to the foreground, illustrated in Fig. 10a, where the color transition from water to clothing is very low.
(2) Self-occlusion occurs and the clothing color is similar to the skin, illustrated in Fig. 10b.

7. Conclusion

In this paper, we present a novel model-based method that recovers upper body poses from static images based on effective joints detection. Guided by the detected face, the torso is described by a deformable torso model and inferred by the MCMC framework. It can deal with various torsos in different images, providing reliable information for the joints initialization. With stable corner detection based on the DT and the color distributions of the foreground and background, the upper joints can be further initialized completely according to some physical configuration constraints. The modified skin likelihood and foreground/background likelihood based on continuous color distributions are proven to be effective for the analysis of background clutter and clothing variations. However, our current technique is limited to the upper body without inter-occlusion between humans. Further research work is needed to develop more accurate part detection methods for pose initialization, and to generalize our algorithm to the analysis of whole-body images.

Acknowledgements

This work is partially supported by a joint project of Tsinghua University, China and Fuji Film Co. Ltd., Japan, the National Natural Science Foundation of China (Project. 60472028), Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20040003015 and a grant from City University of Hong Kong (Project 9610034).

References

Agarwal, A., Triggs, B., 2004. 3D human pose from silhouettes by relevance vector regression. Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2, 882–888.
Agarwal, A., Triggs, B., 2006a. Recovering 3D human pose from monocular images. IEEE Trans. Pattern Anal. Machine Intell. 28 (1), 44–56.
Agarwal, A., Triggs, B., 2006b. A local basis representation for estimating human pose from cluttered images. Proc. Asian Conf. on Computer Vision 1, 50–59.
Boulay, B., Bremond, F., Thonnat, M., 2006. Applying 3D human model in a posture recognition system. Pattern Recognition Lett. 27 (15), 1788–1796.
Chenaoua, K., Bouridane, A., 2006. Skin detection using a markov random field and a new color space. In: Proceedings of IEEE International Conference on Image Processing, 8–11 October, pp. 2673–2676.
Deva, R., Cristian, S., 2006. Training deformable models for localization. Proc. IEEE Conf. on Computer Vision and Pattern Recognition 1, 206–213.
Fan, B., Wang, Z.F., 2004. Pose estimation of human body based on silhouette images. In: Proceedings of International Conference on Information Acquisition, 21–25 June, pp. 296–300.
Gilks, W., Richardson, S., Spiegelhalter, D., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall.
Gupta, A., Mittal, A., Mittal, L.S., 2008. Constraint integration for efficient multi-view pose estimation with self-occlusions. IEEE Trans. Pattern Anal. Machine Intell. 30 (3), 493–506.
Hu, Z.L., Yan, H., Lin, X.G., 2008. Clothing segmentation based on the constrained Delaunay triangulation and foreground and background estimation. Pattern Recognition 41 (5), 1598–1609.
Hua, G., Yang, M.H., Wu, Y., 2005. Learning to estimate human pose with Data Driven Belief Propagation. Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2, 747–754.
Lee, M.W., Cohen, I., 2006. A model-based approach for estimating human 3D poses in static images. IEEE Trans. Pattern Anal. Machine Intell. 28 (6), 905–916.
Lee, M.W., Nevatia, R., 2007. Body part detection for human pose estimation and tracking. In: Proc. IEEE workshop on Motion and Video Computing, pp. 23–31.
Menier, C., Boyer, E., Raffin, B., 2006. 3D skeleton-based body pose recovery. In: Proc. 3rd International Symposium on 3D Data Processing, Visualization and Transmission, pp. 389–396.
Mori, G., Malik, J., 2002. Estimating human body configurations using shape context matching. Proc. Eur. Conf. on Computer Vision 3, 666–680.
Mori, G., Malik, J., 2006. Recovering 3D human body configurations using shape contexts. IEEE Trans. Pattern Anal. Machine Intell. 28 (7), 1052–1062.
Mori, G., Ren, X., Efros, A., Malik, J., 2004. Recovering human body configurations: Combining segmentation and recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2, 326–333.
NIST, 1977. Anthrokids – Anthropometric data of children, http://ovrt.nist.gov/projects/anthrokids/.
Poppe, R., Poel, M., 2005. Example-based pose estimation in monocular images using compact Fourier descriptors. Technical Report TR-CTIT-05-49, University of Twente, Enschede, The Netherlands.
Ren, X.F., Berg, A.C., Malik, J., 2005. Recovering human body configurations using pairwise constraints between parts. Proc. IEEE Conf. on Computer Vision1, 824–831.
Roberts, T.J., 2005. Efficient human pose estimation from real world images. Doctor Thesis, Univ. of Dundee, Scotland.
Sawangsri, T., Patanavijit, V., Jitapunkul, S., 2005. Face segmentation based on Hue-Cr components and morphological technique. Proc. IEEE Symp. Circuits Systems 6, 5401–5404.

Shakhnarovich, G., Viola, P., Darrell, T., 2003. Fast pose estimation with parameter sensitive hashing. Proc. IEEE Conf. on Computer Vision 2, 750–757.
Shewchuk, J.R., 1996. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In: Proc. 1st Workshop on Applied Computational Geometry, pp. 124–133.
Sigal, L., Isard, M., Sigelman, B., Black, M., 2004. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. Proc. Conf. Advances Neural Inform. Process. System 16, 1539–1546.
Singh, M., Mandal, M., Basu, A., 2005. Pose recognition using the Radon transform. In: Proc. 48th Midwest Symposium on Circuits and Systems, pp. 1091–1094.
Sminchisescu, C., Telea, A., 2002. Human pose estimation from silhouettes: A consistent approach using distance level sets. J. WSCG 10 (1), 232–240.
Viola, P., Jones, M., 2004. Robust real-time face detection. Internat. J. Computer Vision 57 (2), 137–154.
Wang, R.X., Leow, W.K., 2006. Human posture analysis under partial self-occlusion. Proc. IEEE Conf. Image Anal. Recognition 1, 874–885.
Xiao, Y., Yan, H., 2004. Extraction of glasses in human face image. In: Proc. 1st International Conf. Biometric Authentication, pp. 214–220.
Ye, Q.X., Gao, W., Wang, W.Q., Huang, T.J., 2004. A color image segmentation algorithm by using color and spatial information. J. Software 15 (4), 522–530.
Zhao, T., Nevatia, R., 2002. Stochastic human segmentation from a static camera. In: Proc. IEEE Workshop on Motion and Video Computing, pp. 9–14.
Zou, J.J., Chang, H.H., Yan, H., 2001. Shape skeletonization by identifying discrete local symmetries. Pattern Recognition 34 (10), 1895–1905.