HEAD-SHOULDER HUMAN CONTOUR ESTIMATION IN STILL IMAGES

Julio C. S. Jacques Junior[*], Cláudio R. Jung*[+], Soraia R. Musse*[*]

[*] Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
[+] Institute of Informatics, Federal University of Rio Grande do Sul, Brazil

ABSTRACT

In this paper we propose a head-shoulder contour estimation model for human figures in still images, captured in a frontal pose. The contour estimation is guided by a learned head-shoulder shape model, initialized automatically by a face detector. A graph is generated around the detected face with an omega-like shape, and the estimated head-shoulder contour is a path in the graph with maximal cost. A dataset with labeled data is used to create the head-shoulder shape model and to quantitatively analyze the results. The proposed model is scaled according to the detected face size to be scale invariant. Experimental results indicate that the proposed technique works well in non trivial images, effectively estimating the contour of the head-shoulder even under partial occlusions.

Index Terms— human head-shoulder estimation, omega-shaped region, human segmentation.

1. INTRODUCTION

Automatic people detection and segmentation can be widely used in many computer vision based applications, including surveillance systems, people counting, photo analysis and editing and so on. As related in the work of Whang et al. [2], a special case of pedestrian detection (head-shoulder detection) has its significance in scenes where only the upper part of the body can be seen due to occlusion, like crowded subway stations, meeting rooms, etc. According to Xin et al. [3], head-shoulder segmentation is an important part of face contextual region analysis for the purpose of human recognition and tracking. In addition, head-shoulder contour estimation models can also be used to help the extraction of general contextual information, such as clothing and hair style, which could be very useful for people identification, especially when the facial features alone do not provide sufficient information.

Much work has been done on head based human detection and tracking. In [4] a method for rapid and robust head-shoulder based human detection and tracking is proposed.The detection is achieved by combining a Viola-Jones type classifier and a local HOG (Histogram of Oriented Gradients) feature based AdaBoost classifier. Xin et al. [3] proposed an automatic head-shoulder segmentation method for human photos based on graph cuts with shape sketch constraint and border detection through learning. In such approach, a face detector is used as a start point and to get the position and size of the human face. Bu et al. [1] proposed a structural patches tiling procedure to generate probabilistic masks which can guide semantic segmentation, applied to a head-shoulder segmentation problem. In the work of Mukherjee and Das [5], a model that employs a set of four distinct descriptors for identifying the features of the head, neck and shoulder regions of a person in video sequences is proposed. Whang et al. [2] proposed an edge feature designed to extract (predict) and enhance the head-shoulder contour and suppress the other contours. The basic idea is that head-shoulder contour can be predicted by filtering edge image with edge patterns, which are generated from edge fragments through a learning process.

Considering methods that focus on extracting the precise head-shoulder contour (and not just an estimate, such a bounding box), the methods proposed in [3, 1] are the most similar to ours. However, it is important to emphasize that they try to segment the whole foreground object, which may include part of the clothes and hair, while we try to segment the most omega-like head-shoulder contour (for the sake of illustration, see Figs. 2(d-e)), focusing on a well known shape/feature of the human body. The main contribution of the proposed method is an automatic, fast and scale invariant approach to segment the head-shoulder contour (of a person), proposed to be robust when the contour is partially occluded (e.g. by a large amounts of hair, accessories and/or clothes).

2. OUR MODEL

In this work we propose to segment the contour of the head-shoulder of a person using a graph-based segmentation model, similar to [6]. The basic idea is to initialize the model with a face detector [7] and create a Directed Weighted Graph (DWG) around the detected face based on a scale invariant shape model (i.e. the graph can be resized according to the size of the detected face). The goal is to find out the path in the graph that maximizes a certain boundary energy that should be large in the boundary between the upper body and the background. The proposed model is described in the next sections.

Authors would like to thank Brazilian agencies CNPq, FAPERGS, CAPES and the authors of [1] for sharing part of their dataset.

2.1. The Head-Shoulder Shape Model

In this work we propose the usage of a shape model to guide the computation of the best path in the graph. The shape model of the head-shoulder was generated based on ground-truth data associated to the adopted dataset. Our dataset is composed by 402 images (acquired in our lab, collected from public datasets [8, 9, 10], including 170 images of the dataset used in [1], sent by the authors), varying in ethnicity, view, angles, resolution, appearances and scenes. The dataset was divided into training and testing dataset, each one with 1/3 and 2/3 (randomly chosen) of the 402 images, respectively. The contour of the head-shoulder of each person, contained in each image of the dataset, was manually formed. This information is used to quantitatively analyze the results and to create an average head-shoulder's shape, as described next.

Firstly we run a face detector [7] for each RGB image of the training dataset. For each detected face, a binary image of its upper body is generated (from the points presented in the ground truth data), which is limited by a bounding-box, as shown in Fig. 1(a). The center point $\mathbf{C}_f$ of each face, radius R_f and the distances of the center of the face to the left and upper sides of the bounding box of the binary image are retrieved (d_1 and d_2, respectively, as illustrated in Fig. 1(a)). The distances d_1 and d_2 will be used as references in a second stage, as explained next. Each binary image is also smoothed with a Gaussian filter to alleviate the high contrast in their boundaries, also taking into account the imprecisions of the manually formed boundaries.

In a second stage, we firstly resize all these binary images by a factor $f_s = \frac{R_f}{R_a}$ (where R_a is the average radius of all detected faces in the previous stage and R_f is the face radius of the image under analysis). Then, we project all the resized images into a plane, accumulating the value of each pixel, to generate the initial shape mask $\mathbf{S}_0$, as illustrated in Fig. 1(b). To increase the number of samples and to deal with small angles orientation on the image plane, the binary images are also flipped in the y axis (vertical). The green plus in Fig. 1(b) represents the reference point $\mathbf{R}_p$, which will be used to align the final shape mask with the center of the detected face in the image under analysis ($\mathbf{R}_p$ is defined by the average of all distances d_1 and d_2, respectively for x and y coordinates – in this case, each distance d_1 and d_2 is normalized by f_s).

The initial shape mask $\mathbf{S}_0$ is thresholded, aiming to capture the essence of the expected contour of the head-shoulder. To do this, two thresholds are defined (T_1 and T_2), based on the analysis of the histogram of $\mathbf{S}_0$ (Fig. 1(c)). Pixels that are related to the background of $\mathbf{S}_0$ will present lower values (dark blue, as shown in Fig. 1(b)), generating a peak close to the origin of the histogram, whereas pixels of other structures (head, shoulders and chest) tend to present larger values (generating a flat region and a second high peak on the right side of the histogram, representing the red area of the initial shape mask). The desired thresholds should lie right after the first peak and right before the second one, and the long valley delimited by the thresholds represents the "uncertainty" region.

More precisely, let $h(S)$ denote the histogram, S_{M1} and S_{m1} denote positions of the first local maximum and first local minimum, respectively (with $S_{M1} < S_{m1}$), and S_{M2} and S_{m2} denote positions of the last local maximum and last local minimum, respectively (with $S_{M2} > S_{m2}$). The desired threshold T_1 is obtained through $T_1 = \min\{S | S_{M1} < S < S_{m1} \wedge h''(S) > 0 \wedge |h'(S)| \leq \alpha\}$, where α is the "flatness" threshold (set experimentally to 0.5774, related to $30°$). The second derivative was included to avoid the selection of points with low derivative close to the local maximum (where $h''(S) < 0$), so that the threshold T_1 is selected after the inflection point. T_2 is computed using the same idea, but for the opposite side of the histogram (e.g. $T_2 = \max\{S | S_{m2} < S < S_{M2} \wedge h''(S) < 0 \wedge |h'(S)| \leq \alpha\}$. This automatic threshold estimation was adapted from [11], which was applied in a color segmentation algorithm. The thresholded image is shown in Fig. 1(d), and pixels away from $\mathbf{R}_p$ more than $3R_a$ (set based on experiments, and shown as a blue circle in Fig. 1(e)) are ignored, so we can get its skeleton that represents the "average" upper body contour, as shown in Fig. 1(e). The final shape model $\mathbf{S}_f$ (Fig. 1(g)) is computed using a Gaussian function, as defined on Equation 1.

$$S_f(x, y) = e^{-\frac{D_t(x,y)^2}{(R_a/2)^2}}, \quad (1)$$

where x, y are the spatial coordinates of each pixel, D_t is the Distance Transform (Fig. 1(f) - computed using the skeleton illustrated in Fig. 1(e)), and the scale factor of the Gaussian is given by $R_a/2$ (set based on experiments). The shape model can be viewed as a prior confidence map on the location of the upper body contour, and it is combined with image data to obtain the final contour, as explained next.

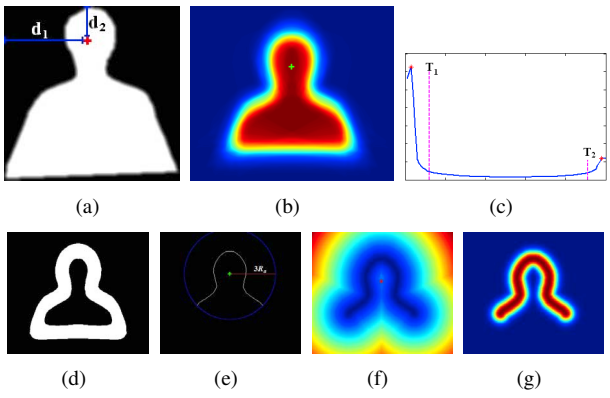
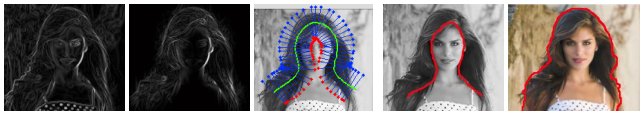


Fig. 1. Shape model generation. (a) The red plus shows the center of the detected face. The distances, d_1 and d_2, to this center point to the left and upper sides of the bounding box of the binary image, respectively. (b) Initial shape mask and the reference point $\mathbf{R}_p$ (plus sign). (c) Histogram analysis. (d) Binarized image. (e) Skeleton of the binarized image. (f) Distance transform of (e). (g) The final shape mask $\mathbf{S}_f$.

2.2. Graph generation, weights of the edges and finding the maximum cost path

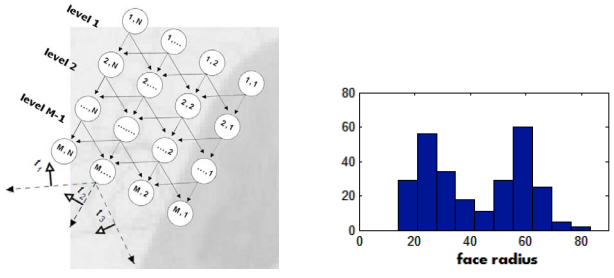
In this work we use an adapted version of the graph-based segmentation model described in the work of Jacques Junior et al. [6]. Let $G = (S, E)$ be a graph generated for a specific face radius, consisting of a finite set S of vertices and a set of edges E. The vertices form a grid-like structure, and they are placed along a region where the contour of the head-shoulder is expected to appear (Fig. 2(c)). The number of the levels of the graph, the length of each level, as well as the number of vertices along the levels are set based on experiments, as described next.

Consider the skeleton curve (Fig. 1(e)), described in the previous section, resized according to a specific face radius R_f and aligned to the center of the detected face $\mathbf{C}_f$ based on reference point $\mathbf{R}_p$. The points of this curve are discretized from one another by $d = 0.15R_f$ pixels. For each two consecutive points of this curve, a level of the graph is generated orthogonal to the line segment that connect these points. Each level, with a length $L = 1.6R_f$ pixels, is centered on its respective line segment. The vertices are labeled $S_{m,n}$, where $m = 1, ..., M$ denotes the level of the vertex, and $n = 1, ..., N$ is the position of the vertex in such level. The number of levels M are based on the number of points of the skeleton curve and the d value, and the number of nodes in each level is set experimentally to $N = L/2$. The levels of the entire graph, for a given image, are illustrated by blue lines in Fig. 2(c) and a detailed illustration of it is shown in Fig. 3(a).



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

Fig. 2. (a-b) Energy terms, without and with the influence of S_f, respectively. (c) Generated graph. (d) The best path (our model – approach ii., as described in Sec. 3). (e) Segmentation presented in [3].



(a) (b)

Fig. 3. (a) Illustration of the nodes/edges of the graph. (b) Histogram of the face radii of the testing dataset.

The edges in the proposed graph relate to line segments connecting two nodes belonging to adjacent levels. More precisely, each node in a level m can be connected to the $k = 3$ (up to) nearest nodes in the level $m + 1$, as illustrated in Fig. 3(a). The weight $w(e_k)$ of each edge e_k is computed as:

$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} E_k(x_j, y_j), \tag{2}$$

where q_k is the number of image pixels in a raster scan along edge e_k, E_k is the energy function, and (x_j, y_j) are the coordinates of the pixels along such scan. The proposed energy function is composed by several factors: edge, shape mask and angular constraints.

As defined in [6], given the luminance component I of the original image, we initially compute the discrete gradient image ∇I using the Sobel operator. If the contour of the person passes through a graph edge e_k, the gradient magnitude $\|\nabla I\|$ is expected to be large in the pixels along e_k, and the orientation of the gradient vector should be orthogonal to the line segment related to e_k. Hence, the first term of the energy map is given by $|\boldsymbol{t}_k \cdot \nabla I|$, where $\boldsymbol{t}_k$ is a unit vector orthogonal to e_k, as illustrated in Fig. 3(a). Another useful information is provided by the learned shape mask $\mathbf{S}_f$. The normalized shape mask tries to decrease energy values far away from the expected head-shoulder contour locations. For the sake of illustration, the energy term combining gradient magnitude and the shape mask ($\|\nabla I\|\mathbf{S}_f$) for one single image is illustrated in Fig. 2(b). Finally, the energy map for pixels related to a graph edge e_k is given by Eq. 3.

$$E_k(x, y) = |\boldsymbol{t}_k \cdot \nabla I(x, y)| S_f(x, y), \tag{3}$$

where S_f is the shape mask aligned to the center $\mathbf{C}_f$ of the detected face by its reference point $\mathbf{R}_p$, and resized according to the factor f_s.

The silhouette of the head-shoulder is defined as the maximum cost path along the graph. Since the graph is acyclic, such path can be computed using dynamic programming, as in Dijkstra's algorithm [12]. It is important to emphasize that the estimated head-shoulder contour is defined by a maximal cost path in the graph, given by a combination of edge and geometric information. In case of partial occlusions, (e.g. long hair in front of the shoulder or strap of a backpack), edge information tends to be weak along the desired contour, but the geometric cue tends to attract the maximal cost path to the desired location. Fig. 2(d) illustrates the estimated head-shoulder contour for a given image.

3. EXPERIMENTAL RESULTS

In this section we illustrate some results of the proposed model[1], also presenting a quantitative evaluation using the

[1] See `www.cpca.pucrs.br/icip2014` for more results.

testing dataset (described in Sec. 2.1). The estimated contour of each analyzed image is a set of points, which is confronted to the respective ground truth data using the modified Hausdorff Distance [13], since the original one is too noise sensitive. In addition, the lengths of the estimated contour curve and the ground truth might be different, increasing the measured error even in very good estimations. To deal with this issue, we create two line segments, r and g (illustrated in Fig. 4(a)), each one passing through the extremity points of each contour curve (the estimated one and the ground truth, respectively), and points bellow these line segments are ignored.

Table 1 summarizes the obtained results in terms of average error (distance in pixels), standard deviation, minimum and maximum errors, using four different approaches[2] to compute the energy map: i) compute the gradient using the luminance component I of the input image, without combining the shape mask information ($\mathbf{S}_f$); ii) compute the gradient using the luminance component I combined with $\mathbf{S}_f$; iii) using the Di Zenzo color edge detector, which computes the gradient using RGB information; iv) using the model proposed[3] in the work of Nezhadarya and Ward [14], which also computes the gradient using color information.

As we can see in Table 1, the use of color in the gradient computation improved the results (considering approach iii.) over the grayscale based gradient method, with little computational overhead (Table 2). Hence, we have selected as default for color images approach number iii., and approach number ii. for grayscale images. Figs. 4(a)-(i) illustrate some obtained results that we consider very promising, using the proposed model (approach ii.). Figs. 4(j)-(m) illustrate some bad segmentation results, due mostly to some of the following reasons: perspective problems (pictures taken from top to bottom or upward); the generated graph does not include part of the expected contour (usually when the face is not centralized to the body); camouflage, hair style and occlusion.

Approach	Feature	Mean	Std	Min	Max
i	Grayscale[1]	8.7331	6.4458	2.0303	45.143
ii	Grayscale[2]	8.1893	6.1711	1.2743	43.95
iii	Di Zenzo	7.9189	5.8889	1.3015	43.5504
iv	[14]	8.5377	6.2511	1.9343	40.927

Table 1. Measured error using the testing dataset (269 images). Grayscale[2] uses the shape mask information (as well as approaches iii and iv) and Grayscale[1] don't.

The computational cost is very important for many commercial applications. As the proposed model is based on the size of each face, we measured the cost using four different face radii intervals, set based on the histogram shown in Fig. 3(b). Table 2 shows the intervals and summarizes the measured times, which we consider very promising (e.g approach ii.), considering that the model was implemented using MATLAB. The hardware used was an HP xw8600 Workstation, with an Intel Xeon processor, Core2 Quad, 2.83GHz and 3Gb of memory. In this experiment the time to detect the faces and I/O procedures was not considered.

[2]All approaches use the angle constraint, defined by the first term in Eq. 3.
[3]Implementation in MATLAB given by the authors.

(a) 2.8084 (b) 4.0319 (c) 3.1981 (d) 2.4468
(e) 2.4453 (f) 1.2743 (g) 2.4790 (h) 2.2531 (i) 6.6618
(j) 15.3635 (k) 14.2974 (l) 8.4676 (m) 43.9500

Fig. 4. Results (red), the ground truth (blue) and the measured error for each image. Image (a) illustrates the two line segments (r and g) used to remove some points of the two curves, which could affect the measured error.

$R_f < 17$	$17 \leq R_f \leq 31$	$51 \leq R_f \leq 65$	$R_f > 65$
0.26 ± 0.04	0.40 ± 0.08	1.11 ± 0.09	1.35 ± 0.14
0.26 ± 0.05	0.41 ± 0.08	1.16 ± 0.10	1.46 ± 0.15
0.30 ± 0.09	0.43 ± 0.08	1.23 ± 0.11	1.58 ± 0.18
36.3 ± 3.72	75.9 ± 18.3	328.4 ± 36.7	455.9 ± 56.6

Table 2. Computational cost (time in seconds) for each approach (in rows, from i. to iv.). Each column shows the mean and standard deviation, considering a specific radius interval.

4. CONCLUSION

This paper presented a new approach for head-shoulder contour estimation for human figures in still images, captured in a frontal pose. The proposed model, which is scale invariant, generates a graph around a face region, guided by a learned head-shoulder shape model. The estimated head-shoulder contour is then defined as the path in the graph with maximal cost. A dataset with manually formed ground truth data was used to validate the experimental results, indicating that the proposed technique works well in non trivial images, effectively estimating the contour of the head and shoulders. Future work will concentrate on exploring other features to increase the accuracy and extend the model to estimate the contour of the upper body.

5. REFERENCES

[1] Pengyang Bu, Nan Wang, and Haizhou Ai, "Using structural patches tiling to guide human head-shoulder segmentation," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 797–800.

[2] Shu Wang, Jian Zhang, and Zhenjiang Miao, "A new edge feature for head-shoulder detection," in *20th IEEE International Conference on Image Processing*, Melbourne, Australia, 2013, pp. 2822–2826.

[3] Hai Xin, Haizhou Ai, Hui Chao, and D. Tretter, "Human head-shoulder segmentation," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 227–232.

[4] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan, "Rapid and robust human detection and tracking based on omega-shape features," in *16th IEEE International Conference on Image Processing*, 2009, pp. 2545–2548.

[5] Subra Mukherjee and Karen Das, "Omega model for human detection and counting for application in smart surveillance system," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 167–172, 2013.

[6] J. C. S. Jacques Junior, C. R. Jung, and S. R. Musse, "Skeleton-based human segmentation in still images," in *19th IEEE International Conference on Image Processing*, Orlando, FL, USA, 2012, pp. 141–144.

[7] P. A. Viola and Michael J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, 2001, pp. 511–518.

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Computer Vision and Pattern Recognition, IEEE Conference on*, june 2008, pp. 1 –8.

[9] Lubomir Bourdev and Jitendra Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Computer Vision, 12th IEEE International Conference on*, 2009, pp. 1365–1372.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2005, vol. 1, pp. 886–893.

[11] J. C. S. Jacques Junior, L. Dihl, C. R. Jung, M. R. Thielo, R. Keshet, and S. R. Musse, "Human upper body identification from images," in *17th IEEE International Conference on Image Processing*, Hong Kong, China, 2010, pp. 1717–1720.

[12] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms, Sec. Ed.*, McGraw-Hill Science/Engineering/Math, 2001.

[13] M. Hossain, M. Dewan, Kiok Ahn, and Oksam Chae, "A linear time algorithm of computing Hausdorff distance for content-based image analysis," *Circuits, Systems, and Signal Processing*, vol. 31, pp. 389–399, 2012.

[14] E. Nezhadarya and R.K. Ward, "A new scheme for robust gradient vector estimation in color images," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2211–2220, 2011.