Improved head-shoulder human contour estimation through clusters of learned shape models

Julio Cezar Silveira Jacques Junior and Soraia Raupp Musse
PUCRS - Faculdade de Informática
Porto Alegre, Brazil
Email: julio.jacques@acad.pucrs.br - soraia.musse@pucrs.br

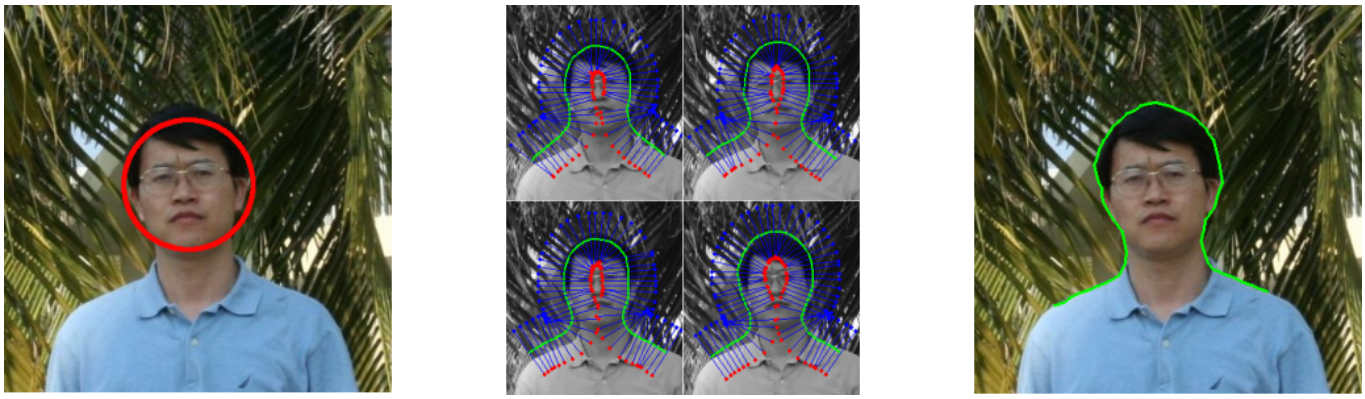


Fig. 1. Overview of our method: from a detected face (left), we build different structured graphs, guided by different learned shape models (middle). Each graph will generate an initial estimative of the head-shoulder contour, defined by the path with maximum cost. The final contour is defined by the path with maximum average energy (right).

***Abstract*—In this paper we propose a clustering-based learning approach to improve an existing model for human head-shoulder contour estimation. The contour estimation is guided by a learned head-shoulder shape model, initialized automatically by a face detector. A dataset with labeled data is used to create the head-shoulder shape model and to quantitatively analyze the results. In the proposed approach, geometric features are firstly extracted from the learning dataset. Then, the number of shape models to be learned is obtained by an unsupervised clustering algorithm. In the segmentation stage, different graphs with an omega-like shape are built around the detected face, related to each learned shape model. A path with maximal cost, related to each graph, defines a initial estimative of the head-shoulder contour. The final estimation is given by the path with maximum average energy. Experimental results indicate that the proposed technique outperformed the original model, which is based on a single shape model, learned in a more simple way. In addition, it achieved comparable accuracy to other state-of-the-art models.**

***Keywords*-human head-shoulder estimation; omega-shaped region; human segmentation.**

I. Introduction

The automatic detection and segmentation of human subjects in static images is still a challenge, due to several real world factors, such as illumination conditions, shadows, occlusions, background clutter, etc. It can also be challenging due to problems associated to image quality, image noise, resolution, or even related to factors associated to the dynamic of the human being, such as the great variety of poses, appearance and shapes. Automatic segmentation models can be widely used in many computer vision based applications, including surveillance systems, people counting, robotics, natural user interfaces, photo analysis and editing and so on.

As related in the work of Li et al. [1], computer vision methods focused on the upper part of the human body are receiving significant attention in the last years. Whang et al. [2] mentioned a special case of pedestrian detection, the head-shoulder detection, which has its significance in scenes where only the upper part of the body can be seen due to occlusion. 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, as mentioned by the authors, head-shoulder contour estimation models can also be used to help the extraction of general contextual information, such as gender [4], clothing appearance (considered the most widely used cue for people re-identification [5]) and hair style [6], which could be very useful for people identification (usually related to soft-biometric based applications [7], [8]), especially when the facial features alone do not provide sufficient information.

Recently, Jacques et al. [9] proposed an approach for human head-shoulder contour estimation in still images, captured in a frontal pose. In their work, the contour is estimated by a path in a graph with maximum cost energy. The graph construction as well as the energy computation is guided by a learned shape model. One drawback of such approach is the use of one single shape model to capture and to estimate the human

head-shoulder contour of a great variety of people, usually with great variations in head shapes/poses, neck width/length and shoulders orientations. Such variations can effectively deteriorate their estimative when the adopted shape model is very different from the head-shoulder contour related to the person in the image under analysis.

In this work we improved the work of Jacques et al. [9] for human head-shoulder contour estimation by using a more sophisticated learning approach, derived from a clustering algorithm, combined with a selecting scheme in the segmentation stage. The advantages of the proposed model (compared to the original one) relies on the usage of different learned shape models to capture and to estimate the human head-shoulder contour, which consider a great variety of head shapes/poses, neck width/length and shoulders orientations. We analyzed quantitatively the results of the proposed approach using the same dataset as in [9] and the same evaluation protocol, achieving an improvement of 22.49% for the average measured error, 48.03% for the minimum computed error and 7.61% for the maximum error (when using grayscale images) and an improvement of 20.63% for the average measured error, 45.98% for the minimum computed error and 6.81% for the maximum error (when color images are employed), detailed in Section IV. In addition, the results obtained by proposed model were compared against two other state-of-the-art models for human head-shoulder segmentation [10] and [3], achieving 8.03% and 7.5% of improvement, respectively.

The remainder of this paper is organized as follows. Section II presents related work concerning head-shoulders detection and segmentation approaches. The proposed technique is described in Section III, and some experimental results are provided in Section IV. Finally, conclusions and suggestions for future work are given in Section V.

II. Related work

Much work has been done on head based human detection and tracking last years (e.g., [1], [11], [12], [13]), with several possible applications. On the other hand, automatic head-shoulder segmentation models for still images seems to be little explored.

Li et al. [1] proposed a method for rapid and robust head-shoulder based human detection and tracking, which is an improvement of their previous work [11]. The detection is achieved by combining a Viola-Jones type classifier and a local Histogram of Oriented Gradients (HOG) feature based AdaBoost classifier, applied in predefined "Entrance" zones. Then, each detected head-shoulder is tracked by a particle filter tracker using local HOG features to model target's appearance.

Zeng and Ma [12] proposed a robust and rapid head-shoulder detector for people counting by combining multilevel HOG with the multilevel Local Binary Pattern (LBP) as the feature set. To further improve the detection performance, Principal Components Analysis (PCA) is used to reduce the dimension of the multilevel HOG-LBP feature set. Tu et al. [13] introduced a robust and rapid head-shoulder detection method for video applications, which is invariant to pose and viewpoint, applied in a surveillance problem. The method combines an attention-based foreground segmentation module and a multiview head-shoulder detection cascade to achieve high performance in both accuracy and speed.

In the work of Jacques et al. [9] an approach to estimate the human head-shoulder contour in still images is proposed. In their work, the contour estimation is guided by a learned head-shoulder shape model, initialized by a face detector [14]. A graph is generated around the detected face with an omega-like shape, and the estimated head-shoulder contour is defined by a maximal cost path in the graph, given by a combination of edge and geometric information. The model is scaled according to the detected face size to be scale invariant. In addition, the model is proposed to be robust when the contour is partially occluded (e.g. by a large amounts of hair, accessories and/or clothes).

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. The model is initialized by a face detector, which is used to get the position and size of the human face. In addition, a watershed algorithm is used to over segment the image under analysis into superpixels, followed by an iterative shape mask guided graph cut algorithm with sketch constraint that is applied to the superpixel level graph to get a contour that segments the head-shoulder from its background. The final estimation is refined by a contour detection algorithm, which is trained by AdaBoost.

Bu et al. [10] proposed a structural patches tiling procedure to generate probabilistic masks which can guide semantic segmentation, applied to a head-shoulder segmentation problem. In this work a local patch structure classifier trained by random forest is firstly applied to the input image in a sliding window manner, followed by the construction of a Markov Random Field (MRF) iteratively optimized to assemble a high quality probabilistic mask from responses collected from the previous stage. In the work of Mukherjee and Das [15], 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. In their work a head-neck-shoulder signature is used to exploits inter person variations in size and shape of people's head, neck and shoulder regions. The model is limited to video applications, due to adaptive background modeling used to extract foreground regions.

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. Li et al. [4] proposed a gender recognition based on the head-shoulder information. In their approach, the Partial Least Squares (PLS) method is employed to learn a very low dimensional discriminative subspace (to extract gradient, texture and orientation information from the head-shoulder area) and a linear Support Vector Machine (SVM) is used for classification.

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], [10] are the most similar to the work proposed in [9]. 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 in [9] they try to segment the most omega-like head-shoulder contour (for the sake of illustration, see Fig. 2), focusing on a well known shape/feature of the human body.

(a) (b)

Fig. 2. Different goals: (a) foreground segmentation [3] and (b) the most omega-like head-shoulder contour estimation [9].

III. Proposed Model

In this work we propose an improvement to the work of Jacques et al. [9] for human head-shoulder contour estimation. The improvement is achieved through two main steps. Firstly, we propose a more sophisticated learning stage, which includes feature extraction from labeled data and the usage of an unsupervised clustering algorithm [16] to automatically estimate the number of shape models to be generated. Given the K-generated shape models, the initial head-shoulder contour estimation, assigned to each shape model and generated graph, is performed in the same way as in [9]. Finally, given the K-initial estimated contours, the final estimative is defined by a selecting scheme, which consider the computed average energy of each initial estimative. The proposed improvement, as well as the main steps related to the original work, used to achieve the final segmentation, is described in details in the next sections.

A. Head-Shoulder Shape Model Generation

As in the original work [9], the shape model of the head-shoulder is generated based on ground truth data associated to the adopted dataset. The dataset is composed by 402 RGB images (256 images collected from public datasets [17], [18], [19], 170 images of the dataset used in [10] and 24 images generated in their work [9], gently sent by the authors), varying in people ethnicity, appearances, shapes, views, orientations, image resolution and scenes.

Following the same protocol as in the original work, the dataset was divided into training and testing dataset, each one with $1/3$ and $2/3$ (randomly chosen) of the 402 images, respectively. In addition, each image of the dataset has an associated ground truth data, defined by the contour points of the expected person's contour, manually formed. The ground truth data is used to quantitatively analyze the results and to create the head-shoulder's shape models. To increase the number of samples and to deal with small angles orientation on the image plane, the ground truth data related to the training set are also flipped in the y axis (vertical).

1) Feature Extraction: The first step of the head-shoulder shape model generation is the feature extraction. In our model, we propose to extract two geometric features from the labeled data: in a general way, the distance from the center of the face to the basis of the neck and the average orientation of the shoulders. The first feature can be used to cope with different neck lengths and the second one to cope with different shoulder orientations, caused by the pose people assume or even by the geometry of their body or clothes.

Firstly, we run a face detector [14] for each RGB image of the training dataset, to get the center point $\mathbf{C}_f$ and radius R_f of the face (Fig. 3(a)). For each detected face, a binary image (illustrated in Fig. 3(b-c)) of its upper body is generated from the points presented in the ground truth data. The binary image is then divided into left and right sides, according to $\mathbf{C}_f$ (illustrated by a dotted line in Fig. 3(b)).

Considering the left side, the binary points are projected onto the vertical axis, generating a curve, which is smoothed with an average filter (with length = 9, set experimentally), illustrated in Fig. 3(d) by a blue line. From this smoothed curve we compute $\mathbf{M}_1$, denoting the position of the first local minimum, illustrated in Fig. 3(d) by a red circle (zoomed in Fig. 3(e)), and compute the first derivative to extract angle information. Points in this smoothed curve with angle orientation higher than a predefined threshold T_α (where $T_\alpha = 0.8391$, related to $40°$, set experimentally) are retrieved, illustrated by blue dots in Fig. 3(d). So, the estimated left neck point $\mathbf{N}_{p1}$ is defined by the nearest retrieved point in relation to $\mathbf{M}_1$, positioned after $\mathbf{M}_1$ (it means, $\mathbf{N}_{p1} > \mathbf{M}_1$), illustrated in Fig. 3(d) by a red plus sign (zoomed in Fig. 3(e) for a better visualization). In some situations there is no local minimum, due to the curvature of the smoothed curve. In such situations $\mathbf{M}_1$ is defined by the most left point of the smoothed curve with orientation higher than T_α, positioned beyond $\mathbf{C}_f$ (in this case, the threshold related to $\mathbf{C}_f$ is illustrated by a dotted line in Fig. 3(d)).

The procedure described above is repeated for the right side of the binary image to estimate its respective neck point $\mathbf{N}_{p2}$. The extracted feature points, related to the neck ($\mathbf{N}_{p1}$ and $\mathbf{N}_{p2}$), illustrated in Fig. 3(b) by green signs, are adjusted accordingly to $\mathbf{C}_f$. Then, the first extracted feature, used in the clustering stage (described next), is defined by the distance (N_d) to the center point $\mathbf{C}_f$ to the line that passes through $\mathbf{N}_{p1}$ and $\mathbf{N}_{p2}$, normalized by R_f, illustrated by a blue line in Fig. 3(b).

The second extracted feature relates to the orientation of the shoulders. As described in the extraction of the previous feature, consider the smoothed curve, generated for the left side by projecting the points of the binary image onto the vertical axis. Similarly, the points of this curve with angle orientation higher than T_α are retrieved. From these retrieved points, the ones positioned before the respective neck point $\mathbf{N}_{p1}$ (in relation to the horizontal axis) are removed, as well as those positioned after $\mathbf{N}_{p1}$ with distance d_c higher than

R_f (where d_c is the arc length measured from $\mathbf{N}_{p1}$). The points satisfying such condition are sent to a curve fitting algorithm, from which the shoulder orientation is derived. Fig. 3(f) illustrates the output of the curve fitting algorithm for the left shoulder.

To cope with clockwise and counterclockwise angle orientations, as well as to reduce the number of features, the signal of the estimated angle is ignored. The reduction of features is desired in this case, as the combination of a very small number of samples (as we are doing) with a high dimensional feature space could forbid the generation of coherent shape models in the clustering stage. Of course, the usage of a larger dataset for learning could encourage the addition of features.

Finally, the procedure described to estimate the orientation of the left shoulder is repeated for the right one and the average angle orientation (S_α), considering both shoulders, is retrieved as the second geometric extracted feature. Fig. 3(c) illustrates the output of the curve fitting algorithm for both shoulders.

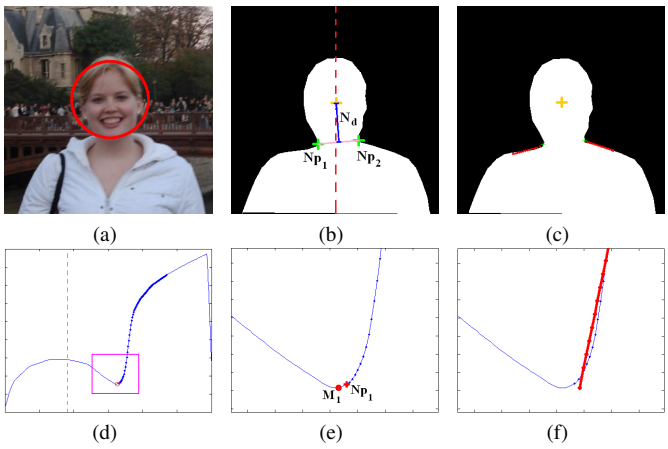


Fig. 3. Feature extraction. (a) Input RGB image and the detected face; (b-c) Binary image generated from the ground truth data, containing illustration of some extracted features, such as the neck points and distance N_d in (b), and shoulder orientations in (c); (d) Computed local minimum $\mathbf{M}_1$ (circle) and the neck point $\mathbf{N}_{p1}$ (plus sign) for the left side of the binary image; (e) Zoom in the rectangular region of image (d); (f) Output of the curve fitting algorithm (red line), used to estimate the angle orientation of the left shoulder.

2) Determining the number of Clusters: Regarding the clustering procedure, each image i of the learning dataset is represented by a 2D feature vector $\mathbf{f}_i$, obtained by combining the extracted geometric features N_d and S_α as below:

$$\mathbf{f}_i = (N_{d,i}, S_{\alpha,i}), \tag{1}$$

with $N_{d,i}$ and $S_{\alpha,i}$ values being normalized according to Eq. 2 and Eq. 3, respectively.

$$N_{d,i} = \frac{N_{d,i} - min(N_d)}{max(N_d) - min(N_d)} \tag{2}$$

$$S_{\alpha,i} = \frac{S_{\alpha,i} - min(S_\alpha)}{max(S_\alpha) - min(S_\alpha)} \tag{3}$$

Coherent shapes tend to produce similar feature vectors $\mathbf{f}$. Hence, a set of coherent shapes is expected to produce a cluster in the 2D space, which is modeled as a Gaussian probability distribution characterized by its mean vector and covariance matrix. Since each cluster relates to a different Gaussian function, the overall distribution considering all feature vectors $\mathbf{f}_i$ is a mixture of Gaussians. The number of Gaussians in the mixture (which corresponds to the number of clusters), as well as the distribution parameters of each individual distribution can be obtained automatically through an unsupervised clustering algorithm [16].

Given the number of clusters and their respective images in the learning dataset, the shape model generation of each cluster j is performed similarly to the original work [9], as described next.

3) Clusters of Shape Models generation: The first step in the shape model generation relates to image resizing. Firstly, all binary images in the learning dataset are resized by a factor $f_l = \frac{R_a}{R_f}$ (where R_a is the average radius of all detected faces in the learning dataset and R_f is the face radius of the image under analysis). The reference point $\mathbf{R}_{p,j}$ of the shape model (assigned to a specific cluster j) is defined by the average of all center points $\mathbf{C}_f$ related to the cluster j. All resized images (assigned to each cluster j) are projected onto a plane, aligned by their respective face center $\mathbf{C}_f$ to $\mathbf{R}_{p,j}$, accumulating the value of each pixel. Such projection generates the initial shape mask $\mathbf{S}_{0,j}$, as illustrated in Fig. 4(a). Aiming to capture the essence of the expected contour of the head-shoulder, the initial shape mask $\mathbf{S}_{0,j}$ is thresholded by an histogram analysis procedure, in the same way as in [9]. The thresholded image is illustrated in Fig. 4(b).

A morphological thinning operation is performed over the thresholded image, from which pixels away from $\mathbf{R}_{p,j}$ more than $3R_a$, as well as undesired branches are ignored. The resulted skeleton curve $\mathbf{S}_{s,j}$ (with 1 pixel-wide) is illustrated in Fig. 4(c) (dilated for visualization purpose). The skeleton curve $\mathbf{S}_{s,j}$ is used to build the final shape mask $\mathbf{S}_{f,j}$, as well as to guide the construction of the graph, as described in the next section.

Finally, the final shape model $\mathbf{S}_{f,j}$ (Fig. 4(d)) related to each cluster j is computed using a Gaussian function (as in [9]), defined in Eq. 4.

$$S_{f,j}(x,y) = e^{-\frac{D_{t,j}(x,y)^2}{(R_a/2)^2}}, \tag{4}$$

where x, y are the spatial coordinates of each pixel, $D_{t,j}$ is the Distance Transform (computed using the skeleton $\mathbf{S}_{s,j}$ illustrated in Fig. 4(c)), 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. 5 illustrates a few generated skeleton curves, using the procedure described above.

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

Fig. 4. Shape model generation. (a) Initial shape model $\mathbf{S}_{0,j}$ and the reference point $\mathbf{R}_{p,j}$; (b) thresholded image; (c) skeleton image $\mathbf{S}_{s,j}$ and (d) final shape mask $\mathbf{S}_{f,j}$ for a given class j.

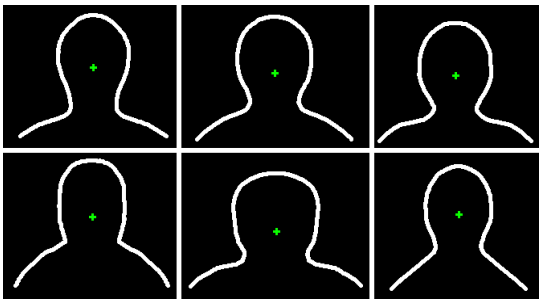
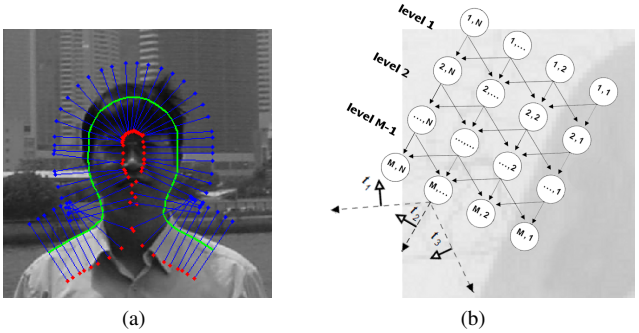


Fig. 5. Illustration of different generated skeleton curves $\mathbf{S}_{s,j}$ and their respective reference points $\mathbf{R}_{p,j}$ (dilated for visualization purpose).

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

The graph generation procedure, the way the edges of the graph are weighted and how the best path is chosen are performed in the same way as in [9]. In this section we briefly describe the original approach (see [9] for a detailed explanation).

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. 6(a)), which is defined by the skeleton curve $\mathbf{S}_{s,j}$ previously computed (resized according to f_s, where $f_s = \frac{R_f}{R_a}$), and aligned to $\mathbf{C}_f$ by $\mathbf{R}_{p,j}$. The goal of using in this stage f_s instead of f_l is to adapt the learned shape model to the input image resolution. The number of the levels, the length of each level of the graph, as well as the number of vertices along the levels are set experimentally.



(a) (b)

Fig. 6. Illustration of the graph. (a) The green line represents the generated skeleton curve $\mathbf{S}_{s,j}$ for a specific cluster j, whilst the blue lines illustrates the levels of the graph; (b) a detailed illustration of the graph, with its nodes and edges.

The edges of the 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. 6(b). The weight $w(e_k)$ of each edge e_k is computed as:

$$w(e_k) = \frac{1}{q_k} \sum_{j=1}^{q_k} E_k(x_j, y_j), \tag{5}$$

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. The energy map for pixels related to graph edges is given by Eq. 6.

$$E_k(x, y) = |\mathbf{t}_k \cdot \nabla I(x, y)| S_{f,j}(x, y), \tag{6}$$

where $\mathbf{S}_{f,j}$ is the shape model, resized according to f_s, aligned to the detected face center $\mathbf{C}_f$ of the person under analysis by $\mathbf{R}_{p,j}$ (also scaled according to f_s); $\mathbf{t}_k$ is a unit vector orthogonal to the measured graph edge (to prioritize contour with similar orientation as the graph edge under analysis), and $\nabla I(x, y)$ is the discrete gradient image computed using the Di Zenzo operator (for color images) or the luminance component I of the original image (for grayscale images).

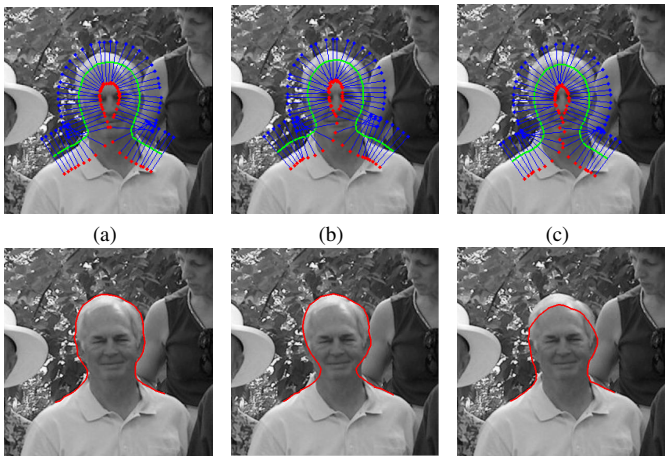
The silhouette of the head-shoulder is defined as the maximum cost path along the graph. Since the graph is acyclic, such path can be computed using dynamic programming, as in Dijkstra's algorithm [20].

The procedure described above is used to obtain the best path for a given graph, related to a specific cluster j. The contribution of the proposed model, beyond the learning stage, is to estimate the head-shoulder contour from different estimated paths (computed from different graphs). In order to achieve this goal, the energy of each path P_j (associated to each cluster/graph j) is defined by the average energy along such path. The final contour estimation is defined by the path with maximum P_j energy. Fig. 7(a-c) illustrates different graphs (generated from different skeleton curves $\mathbf{S}_{s,j}$) and their respective estimated contour, as well as the computed average energy P_j (Fig. 7(d-f)).

IV. Experimental Results

In this section we illustrate some results of the proposed model, also presenting a quantitative comparison with the work of Jacques et al. [9] and against two other state-of-the-art models for human head-shoulder segmentation [10], [3].

In relation to [9], the same evaluation protocol was used to make fair comparison, i.e. we randomly partition the dataset intro training and testing datasets, with $1/3$ of the images of the dataset used for training and $2/3$ for testing. The whole procedure is repeated 5 times, regarding the proposed model, and the results are presented in Table I, in terms of average error (distance in pixels), standard deviation, minimum and maximum error (measured errors assigned to the original work were taken from [9]).



(a) (b) (c)

(d) $P_1 = 0.2836$ (e) $P_2 = 0.2226$ (f) $P_3 = 0.1897$

Fig. 7. Selecting the final head-shoulder contour. The best path, computed for each graph, which has the maximum P_j energy, is assigned to the final estimation. (a-c) Illustrate three generated graphs. In (d-f) their respective estimated contours with respective P_j measured energy.

The estimated contour of each analyzed image (defined by the path with maximum P_j energy) is a set of points, which is confronted to the respective ground truth data using the modified Hausdorff Distance [21]. As mentioned in the original work [9], 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, two line segments, r and g (illustrated in Fig. 8(a)) are created, each one passing through the extremity points of each contour curve (the estimated one and the ground truth, respectively), and points below these line segments are ignored.

Experimental results were obtained by using two different approaches to compute the energy map (∇I):

i) Computing the gradient using the luminance component I of the input image;
ii) Using the Di Zenzo color edge detector, which computes the gradient using RGB information.

Both approaches use the angle constraint combined with the shape mask information ($\mathbf{S}_{f,j}$), as proposed in [9]. The results presented in Table I, regarding the proposed model, were generated from an average number of clusters equals to 13.9 (standard deviation = 2.13). As we can see in this first experiment, the proposed model outperformed the original work, achieving an improvement of 22.49% for the average measured error and 48.03% for the minimum error (approach i), and an improvement of 20.63% and 48.03% for the same measured errors, respectively, regarding the approach ii.

TABLE I
COMPARING THE RESULTS OBTAINED BY THE ORIGINAL WORK [9] WITH THE PROPOSED MODEL.

Approach	Mean	Std	Min	Max
i ([9])	8.1893	6.1711	1.2743	43.9500
i (our)	6.3476	6.0711	0.6622	40.6049
ii ([9])	7.9189	5.8889	1.3015	43.5504
ii (our)	6.2855	5.9632	0.7031	40.5859

The number of clusters used to estimate the head-shoulder contour is hardly related to the computational cost, as well as the size of each face under analysis. The cost is reported using the two most predominant face radii intervals found in the adopted dataset. If we consider the proposed model generates one single cluster/graph, the computational cost[1] varies from about 0.26 ± 0.05 seconds when $R_f < 17$ pixels to 1.46 ± 0.15 seconds when $R_f > 65$ pixels. The computation time can be multiplied by the number of used clusters, which could increase undesirably if considered a large number of clusters.

In order to optimize the computational cost we conducted a second experiment, in which up to K_n components of the mixture with highest prior probability (mixture weigh) were used in the shape model generation (as the number of cluster is related to the number of components of the mixture, set experimentally to $K_n = \{5, 3\}$). Table II summarizes the obtained results for this second experiment. As we can see in Table II, the proposed model have their accuracy decreased after discarding several components of the mixture, but still outperformed the original work (in terms of average and minimal evaluated error, i.e., achieving an improvement of 9.87% for the average error, using $K_n = 3$, and 45.69% for the minimum error, using $K_n = 5$, in the worst scenario).

TABLE II
MEASURED ERROR OBTAINED BY OUR MODEL, CONSIDERING THE K_n COMPONENTS OF THE MIXTURE WITH HIGHER MIXTURE WEIGH.

Approach	K_n	Mean	Std	Min	Max
i	5	6.5598	6.3420	0.6920	42.3141
ii	5	6.5018	6.3454	0.6920	42.3141
i	3	7.1163	6.6406	0.6271	42.3141
ii	3	7.1360	6.7281	0.5391	42.3141

Fig. 8 illustrates some obtained results by the original work [9] (odd rows), compared to the proposed model (even rows), both obtained from grayscale images (approach i).

Trying to compare the proposed model with the state-of-the-art [10], [3] we conducted a third experiment, as described next. Firstly, our dataset with 402 images were again divided into training and testing. The testing dataset contains the 170 images sent by the authors of [10] (aiming to reproduce their experiment, in which a testing dataset with 170 images was used) and the training dataset contains the remaining 232 images. In a second stage, 11 clusters of head-shoulder shape models were generated (as described in Sec. III-A) using the training dataset. Finally, the segmentation is performed over the testing dataset and the results are summarized in Table III in terms of average precision (measured errors assigned to [10], [3] were taken from [10]). In this evaluation, the extremity points of the estimated contour (as well as those from the generated ground truth data) were connected and filled out to generate a blob (to compute the average precision). As in the first experiment, the precision rate for

[1]Measured from a MATLAB implementation, using an HP xw8600 Workstation, with an Intel Xeon processor, Core2 Quad, 2.83GHz and 3Gb of memory (time to detect the faces and I/O procedures was not considered).

Fig. 8. Qualitative comparison: odd rows show the results obtained by the work of Jacques et al. [9] in red lines (ground truth in blue lines), whilst even rows show obtained results by the proposed model (green lines).

each image was measured disregarding those points below the line segments r and g (illustrated in Fig. 8(a)).

TABLE III
COMPARISON TO THE STATE-OF-THE-ART: AVERAGE PRECISION.

Our	**[10]**	**[3]**
92.30	85.44	85.86

It is important to emphasize that such evaluation is quite similar to those made in [10]. The main difference here is that we used our own ground truth data and ignored those points below the line segments r and g, whereas in [10], [3] the whole foreground object is considered. Our goal here is to perform a fair comparison using a very similar protocol (as defined in [10]). Fig. 9 illustrates a qualitative comparison among these works. As we can see in Table III and Fig. 9, the proposed model achieved satisfactory results, with comparable accuracy, according to numbers and visual inspection, to the state-of-the-art.

(a)

Fig. 9. From the left to right: input image, results obtained from [10], [3] and by the proposed model (with ground truth illustrated by blue lines and the estimated contour through the green ones), respectively.

V. FINAL CONSIDERATIONS

In this work we proposed an improvement to the work of Jacques et al. [9] for human head-shoulder contour estimation. The improvement relates to a more sophisticated learning approach, combined with a selection scheme in the segmentation stage.

The proposed model is initialized by a face detector, as in the original approach, and the segmentation is given by a path in a graph with maximal cost. In the learning stage, geometric features are extracted from labeled data. Then, the number of shape models to be learned, used in the segmentation stage, is obtained by an unsupervised clustering algorithm. Different graphs are built around the detected face, according to the number of clusters. The final estimation is obtained by selecting the path with maximum average energy, derived from each graph. The proposed model achieved an improvement of 22.49% when compared to the original work, regarding the evaluated average error. In addition, experimental results indicated that the proposed technique works well in non trivial images, with comparable accuracy to the state-of-the-art. Future work will concentrate on exploring appearance features to increase the accuracy.

ACKNOWLEDGMENT

The authors would like to thank Brazilian agencies FAPERGS and CAPES for the financial support and the authors of [10] for sharing part of their dataset.

REFERENCES

[1] M. Li, Z. Zhang, K. Huang, and T. 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.

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

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

[4] M. Li, S. Bao, W. Dong, Y. Wang, and Z. Su, "Head-shoulder based gender recognition," in *20th IEEE International Conference on Image Processing*, Melbourne, Australia, 2013.

[5] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person reidentification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 5, pp. 869–878, May 2015.

[6] D. Wang, S. Shan, H. Zhang, W. Zeng, and X. Chen, "Isomorphic manifold inference for hair segmentation," in *Automatic Face and Gesture Recognition, 10th IEEE International Conference and Workshops on*, April 2013, pp. 1–6.

[7] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 29:1–29:37, Dec. 2013.

[8] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 739–777, Jan. 2011.

[9] J. C. S. J. Junior, C. R. Jung, and S. R. Musse, "Head-shoulder human contour estimation in still images," in *21th IEEE International Conference on Image Processing*, Paris, France, 2014, pp. 278–282.

[10] P. Bu, N. Wang, and H. 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.

[11] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[12] C. Zeng and H. Ma, "Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting," in *Pattern Recognition, 20th International Conference on*, Aug 2010, pp. 2069–2072.

[13] J. Tu, C. Zhang, and P. Hao, "Robust real-time attention-based head-shoulder detection for video surveillance," in *Image Processing, 20th IEEE International Conference on*, Sept 2013, pp. 3340–3344.

[14] P. A. Viola and M. 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.

[15] S. Mukherjee and K. 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.

[16] M. A. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, March 2002.

[17] 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.

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

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

[20] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Sec. Ed.* McGraw-Hill Science/Engineering/Math, 2001.

[21] M. Hossain, M. Dewan, K. Ahn, and O. 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.