Skeleton-based Human Segmentation in Still Images

Julio C. S. Jacques Junior, Soraia R. Musse*
Computing Science Department
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre/RS, Brazil
Email: juliojj@gmail.com, soraia.musse@pucrs.br

Cláudio R. Jung+
Institute of Informatics
Universidade Federal do Rio Grande do Sul
Porto Alegre/RS, Brazil
Email: crjung@inf.ufrgs.br

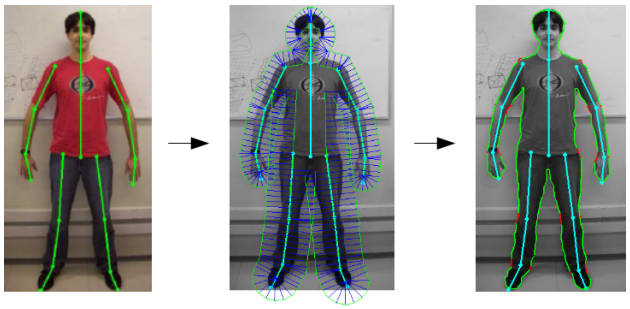


Fig. 1. Overview of our method: input image and their associated skeleton model (left); generated graph (middle); experimental segmentation result (the best path in the graph) with semantic information included (right).

***Abstract*—In this paper we propose a skeleton-based model for human segmentation in static images[1]. Our approach explores edge information, orientation coherence and anthropometric-estimated parameters to generate a graph, and the desired contour is a path with maximal cost. Experimental results show that the proposed technique works well in non trivial images.**

***Keywords*-human body parts segmentation, semantic information**

I. Introduction

The automatic segmentation of human subjects in static images is still a challenge, mainly due to the influence of numerous real-world factors such as shading, image noise, occlusions, background clutter and the inherent loss of depth information when a scene is captured into a 2D image, as well as other factors associated with the dynamics of the human being (great variability of poses, shapes, clothes, etc).

Some studies found in the literature show that people segmentation in static images has become a focus of attention in recent years, and it can be used in several applications, for example, human pose and shape estimation (2D or 3D), image editing, among other. Jacques Junior et al. [8] proposed to solve this problem in an automatic way, starting with a face detection algorithm, and then using color information and anthropometric parameters. It can also be initialized from a pose estimator algorithm as in Freifeld's work [4] combined with a cost function that fits the best pose and shape based on a learned model computed in a previous stage.

[1]Doctoral thesis.
*Advisor.
+Co-advisor.

Another class of techniques tried to detect and segment simultaneously human figures in images, also related to pose estimation algorithms. The approach proposed by Mori et al. [10] is based on segmenting the limb and torso, which are assembled into human figures. Lin et al. [9] proposed a Bayesian approach to achieve human detection and segmentation combining local part-based and global template-based schemes. In a similar way, Gavrila [5] presents a probabilistic approach to hierarchical, exemplar-based shape matching.

Guan and collaborators [6] tried to solve the problem of person segmentation in images using a semi-automatic approach. Basically they compute shape and pose parameters of a 3D human body model directly from monocular image cues, given a user-supplied estimate of the subject's height and a few clicked points on the body, generating an initial 3D articulated body pose and shape estimative. Using this initial guess they generate a tri-map of regions inside, outside and on the boundary of the human, which is used to segment the image using Graph-cuts [2].

This paper proposes a skeleton based approach for silhouette extraction, allowing the precise segmentation of each body part. Our approach is described next.

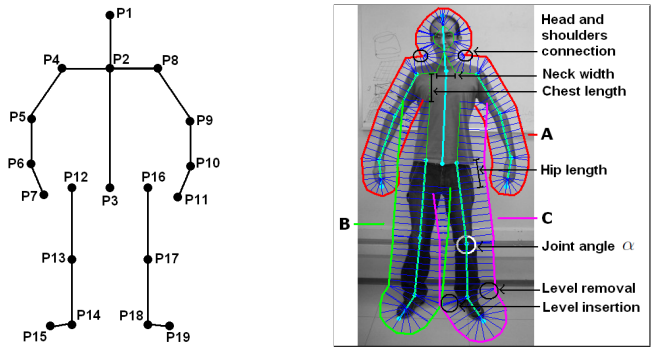
II. Our approach

In this work we propose skeleton-based human segmentation algorithm. The adopted skeleton can be obtained manually or automatically, based on the desired application. Manually, the user must provide a few clicks locating joints of the human structure and then the remainder of the segmentation process is fully automatic. On the other hand, the skeleton could be acquired automatically, through a 2D pose estimation

algorithm, for example [1]. The basic idea is to create a graph around the skeleton model and find out the path that maximizes a certain boundary energy. Next we describe the proposed model using a semi-automatic approach, aiming to prevent problems related to 2D pose estimation algorithms. Some results obtained in a full automatic way using the proposed model are illustrated on Section III.

A. The skeleton model: input data

In our work, the skeleton model is composed by sixteen bones and nineteen joints, as illustrated in Fig. 2(a). All these bones have their widths estimated, parametrized as a function of the height h of an average person based on anthropometric values [12]. More precisely, for a certain body part with label i, the corresponding width w_i is given by $w_i = hf_{wi}$, where the proportionality factors f_{wi} are derived from [12]. Table I presents all body parts used in this work, along with the corresponding values for f_{wi}.



(a) (b)

Fig. 2. (a) The adopted skeleton model. (b) Illustration of the three generated graphs, given the input skeleton, as well some high level information (connections, distances, etc).

There are two different ways to obtain the height of the person through manual intervention. When the person is standing in the photograph and the full body is visible, the user simply clicks on the top of the head and on the bottom of the feet, obtaining the height directly. In any other situation, the height is estimated from the size of the face and anthropometrical relationships. More precisely, the user clicks on the top of the head and on the tip of the chin, to compute the height of the face h_f. The height of the person is then estimated by $h = h_f/0.125$, where 0.125 is a weight derived from anthropometric values [12]. The height of the person could be obtained automatically in a similar way, estimated from a skeleton model acquired trough a 2D pose estimation algorithm [1] and anthropometric relationships, for example.

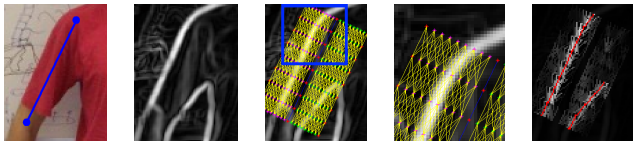
B. Graph generation

The basic idea behind our model is to split the whole contour into groups of body parts, finding the local contour in each of these groups as a path in a graph, and connect them together to obtain the silhouette of the person. For that purpose, we define three main graphs: one for the "upper body" and other two for the "lower body" (left and right sides), as illustrated in Fig. 2(b) in red (A), green (B) and magenta (C).

The assumption made in the proposed model is that each body part, defined from two control points, is limited by a contour, which should have similar orientation to their respective "bone" (except the head, hands and feet) as well to respect some anthropometric distances constraints. The person contour is described by a path in the graph, which should satisfy a predefined condition (for example, the best path is the one that maximizes some kind of energy value). Fig. 3 illustrates this process for one body part.

TABLE I
FIRST COLUMN: THE BODY PARTS INDEX; SECOND COLUMN: THE BODY PART (BONE); THIRD COLUMN: THE TWO JOINTS THAT FORM EACH BONE (LEFT AND RIGHT SIDES); FOURTH COLUMN: THE WEIGHTS USED TO COMPUTE THE WIDTH OF EACH BONE.

i	Bone	Joints	f_{wi}
0	Head	(P1 - P2)	*0.0883*
1	Torso	(P2 - P3)	*0.1751*
2 & 5	Arms	(P5 - P4) & (P9 - P8)	*0.0608*
3 & 6	Forearms	(P6 - P5) & (P10 - P9)	*0.0492*
4 & 7	Hands	(P7 - P6) & (P11 - P10)	*0.0593*
8 & 11	Thigh	(P13 - P12) & (P17 - P16)	*0.0912*
9 & 12	Calf	(P14 - P13) & (P18 - P17)	*0.0608*
10 & 13	Foot	(P15 - P14) & (P19 - P18)	*0.0564*
14 & 15	Shoulders	(P4 - P2) & (P8 - P2)	*0.0608*



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

Fig. 3. (a) input image. (b) gradient magnitude. (c) nodes and edges of a graph. (d) zoom on image (c). (e) the best path.

Let $G_i = (S, E)$ be a graph generated for each body part i, 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 body part is expected to appear (Fig. 3(d) shows the graph related to the external contour of the right arm). The vertices form levels along the grid, and each level is orthogonal to the line segment connecting two control points ($\boldsymbol{P}_{i'}$ and $\boldsymbol{P}_{i''}$), which are associated with the respective body part ("bone"). The extent of each level, as well as the number of vertices along the levels are based on anthropometric values (described in Table I) that provide the expected width of each body part. 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, so that smaller values of n are closer to the corresponding bone. The values of M and N were set experimentally to $M = 0.1\|(\boldsymbol{P}_{i'} - \boldsymbol{P}_{i''})\|$ and $N = 0.33w_i$ (where $i = 2$ for the graph A and $i = 8$ for the graphs

B and C), i.e. the number of levels for each body part is proportional to the length of the corresponding bone, and the number of vertices per level is proportional to the width of the arm for the upper graph A and to the width of the thigh for the lower graphs (B and C).

Connecting individual graphs and special cases: The graph definition described so far is focused on a single body part. The three main graphs used in our work are formed by several body parts, so that the graphs related to each individual body part must be connected. When a body part is connected to another, the regions delimited by the corresponding graphs may overlap or leave gaps, depending mostly on the angle α formed by the connection joint. In a general way, there is overlap when $\alpha < 180°$, so that levels must be removed, and gaps when $\alpha > 180°$, so that levels must be inserted to fill the gaps. An example of creation and removal of levels is shown in Fig. 2(b), in particular the connection of the calf and the foot. In the outer contour of such connection, the graphs overlap and levels must be removed; in the inner part of the connection, however, there is a gap between the individual graphs, so that new levels must be created. Also, it should be mentioned that if $\alpha = 180°$, then the graphs are simply concatenated. As the number of nodes at each level is the same for each main graph (A, B and C), the connectivity is maintained along the "bone" direction.

In addition, as it will be discussed next, a path in a main graph is generated from the first level to the last one, which ensures that the computation of each arm/leg will be made separately. These constraints ensure the connectivity of the limbs in most cases, except when the movements of the limbs are not approximately on the image plane (which affect the anthropometrical estimates in the projected image).

Although most body parts generate a graph placed in a rectangular region, as described previously, some specific body parts present a different shape, such as head, hands and feet. The hands and feet are modeled as circular sectors, and the levels are radial lines discretized by an angle of 22.5°, chosen experimentally.

The head is modeled by an hexagonal shape. Basically, the "bone" of the skeleton associated with the head is decreased in the top by a factor (set experimentally to $w_{neck}/2$) and in the bottom to w_{neck}, where $w_{neck} = 0.0333h$ (based on [12]). So the initial graph of the head is generated with the same width of the arms around the hexagon, to maintain a certain global coherence in the graph A. Finally, the graph of the head is connected to the graph of the shoulders in the intersection point of their boundaries, as illustrated in Fig. 2(b).

Each "bone" of the shoulders is initially decreased by a factor (set experimentally to the half of neck's width) in the side of the neck. The graph generation is similar to a regular body part (arms and legs, for example), but now the graph is created only for one side of the "bone" (the upper side), using the same width used for the arms.

The torso is modeled by two different graphs (one for each side - left and right). Basically, we create two line segments, connecting each femur (points $\boldsymbol{P}_{12}$ and $\boldsymbol{P}_{16}$) to their associated shoulder, in the average point of each respective shoulder "bone". This line segment is decreased in the top by a factor (set experimentally to the length of the chest l_{chest}, where $l_{chest} = 0.0980h$), to deal with the underarms, and the graph of the torso is generated as other regular body parts. Finally, the internal parts of the graphs of the legs are also cut on their upper extremity by an estimated distance (the length of the hip l_{hip}, where $l_{hip} = 0.0492h$).

Weights of the edges: The edges in the proposed graph relate to line segments connecting two nodes. The weight $w(e_k)$ of each edge e_k is given by the average energy of the pixels that lie in the corresponding line segment, i.e.,

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

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 map is composed by several factors: edge, anthropometry and angular constraints, as explained next.

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 trough 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. 4.

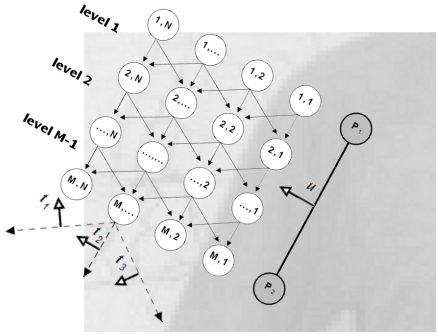


Fig. 4. Illustration of the graph generation for the outer part of the right arm.

Another useful information is provided by anthropometric measures, since the expected width w_i of each body is related to the person height, as shown in Table I. In this work, we also prioritize edges that lie at close to a distance $w_i/2$ from the respective "bone". More precisely, we create two line segments parallel to the "bone" (each one at a distance $w_i/2$) and then compute the Distance Transform (DT), generating an anthropometric distance map R_i for each body part given by

$$R_i(x, y) = e^{-\frac{D_i(x,y)^2}{(w_i/4)^2}}, \tag{2}$$

where D_i is the DT for body part i, and the scale factor of the Gaussian is given by $w_i/4$. For the sake of illustration, the energy term combining gradient magnitude and anthropometric distances ($\|\nabla I\| R_i(x,y)$) for the right arm is illustrated in Fig. 5(c).

(a) (b) (c)

Fig. 5. (a) Input image and their associated "bone". (b) Illustration of the energy term $\|\nabla I\|$ (without the anthropometric distances constraint). (c) Illustration of the energy term $\|\nabla I\| R_i(x,y)$.

The graph is influenced by the adjacent body parts close to the joints. In such portions of the graph, the anthropometric distance map is computed as a weighted average of the distance maps related to the adjacent body parts, and the weights are proportional to the distance of the pixel under consideration to each body part. Hence, the overall distance map $R(x,y)$ presents smooth connections.

The third term in the energy map (Eq. 3) aims to prioritize graph edges that are approximately parallel to the orientation of the corresponding bone. In fact, such term is characterized by $|\boldsymbol{u} \cdot \boldsymbol{t}_k|$, where $\boldsymbol{t}_k$ are unit vectors orthogonal to e_k, as already explained, and $\boldsymbol{u}$ is a unit vector orthogonal to the bone, as illustrated in Fig. 4.

Finally, the energy map for pixels related to a graph edge e_k is given by

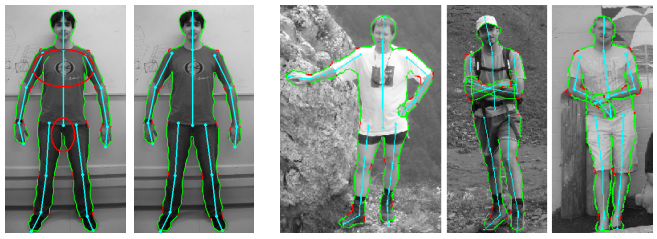
$$E_k(x,y) = |\boldsymbol{t}_k \cdot \nabla I(x,y)| R(x,y) |\boldsymbol{u} \cdot \boldsymbol{t}_k|. \quad (3)$$

C. Finding the maximum cost paths

The procedure defined so far is used to create three main graphs (A, B and C), related to the upper and lower (left and right) body parts. The silhouette of the person in each of these parts is defined as the maximum cost path along the corresponding graphs. Since the graph is acyclic, such path can be computed using dynamic programming, as in Dijkstra's algorithm [3]. In the connection of the main graphs, the contours may intersect or leave gaps, as illustrated in Fig. 6(a) - left. The connection points of the arms are those nearest to the beginning of the contour of the torso. The internal points of the thighs with the smallest distance from one another are connected (if there are more than one, we use the one closest to $\boldsymbol{P}_3$). The final silhouette is shown in Fig. 6(a) - right.

III. Experimental Results

In this section we illustrate some results of the proposed model[2]. It is important to notice that each point of the contour has a label associated to it, so all body parts are identified.



(a) (b)

Fig. 6. (a) Connecting the three main paths (b) Results.

The semantic information is also illustrated in the results by a red contour, which divides two consecutive body parts. One limitation of the proposed approach is to deal with poses when the movements of the limbs are not approximately on the image plane (which affect the anthropometrical estimates in the projected image). Fig. 6 shows some experimental results. Fig. 7 shows a comparison of the proposed method and the approach described in [4]. As it can be observed, the proposed method adapts better to the contours, while the human body shape priors in [4] enforce a smoother contour. It is important to notice that in Freifeld's work [4] the skeleton is acquired automatically (Fig. 7(c)) whilst in our approach the skeleton (shown in Fig. 7(a)) is acquired through user intervention.

Freifeld et.al [4] compared their results with those acquired through Grab-Cut [11] segmentation algorithm, as shown in Figure 8(b-c). In such case, we can lose a lot of information, when comparing a semantic contour against a blob. For example, consider a person with the arms in front of the torso, as illustrated in Figure 9(a). It can be observed that the connectivity of the contour, as well as the semantic information, are maintained in the proposed model (Figure 9(b)), whilst it is lost inside a blob (as illustrated in Figure 9(c) – result acquired through user intervention, using the Graph-Cuts [2] algorithm).

The proposed model can work fully automatically if the input data is obtained in a similar way. For example, Figure 10(b) illustrates the result obtained using the 2D pose estimated from Kinect [3], whilst Figure 10(c) illustrates the result achieved through user intervention. In Figure 10(b-c) and Figure 11, the red lines illustrate the results of the proposed model, whilst the blue ones are those informed by a user (for qualitative comparison). It can be observed that the results obtained automatically do not outperform those acquired semi-automatically, although the latter may considered very convincing. Such problem can be correlated to the challenges related to pose estimation algorithms (for example, the estimated pose, or a part of it, could be not well centered over the associated human figure, as illustrated on Figure 11).

The experimental results shown in Figure 10(b) and Figure 11 are generated in a fully automatic manner using the 2D estimated poses given by the *Kinect* sensor. We believe

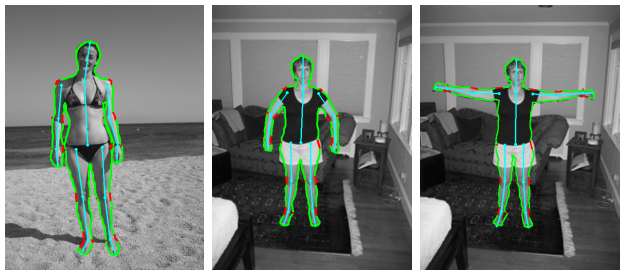
[2]See www.inf.pucrs.br/~smusse/ICIP12 for more results.

[3]*Kinect* for Windows: http://www.microsoft.com/en-us/kinectforwindows/

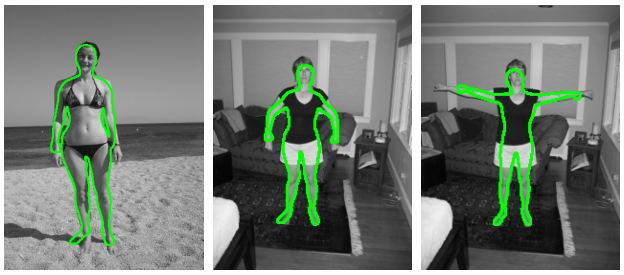
that the depth map (given by the *Kinect* sensor and not used in our experiments) could be incorporated in the Energy equation (Eq. 3) in a future work, aiming to include depth information and deal with low contrast or camouflage, for example. A very simple way to include such feature (depth information) to the model could be made by computing the gradient of the depth map and merge it with the current gradient map (generated from the grayscale image). Of course, other approaches could be used. It is important to notice that such new feature (depth) will make a hard restriction to the model: the need to work with the *Kinect* sensor or similar devices.

The Figure 12 illustrates some experimental results using images with very low contrast (in some specific body parts). As it can be observed, the proposed model works very well.

In addition, it was observed during the experiments that the model is very robust regarding the user provided skeleton (small variations in the input data - informed joints and height of the person - do not drastically affect the results). For example, if the user/algorithm provides a bad joint, it will affect mainly their associated body part and the adjacent body parts, without propagating the error to the whole body.



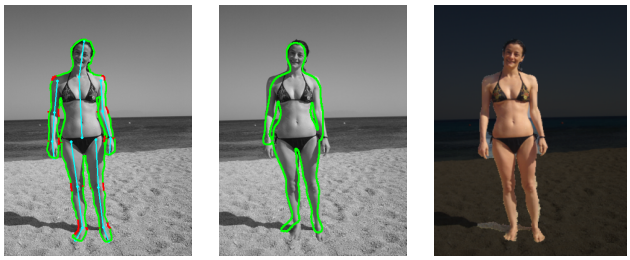
(a)



(b)

(c)

Fig. 7. (a) Our results. (b) Results obtained with [4]. Skeleton acquired through a 2D pose estimation [1], used in the Freifeld's work [4].

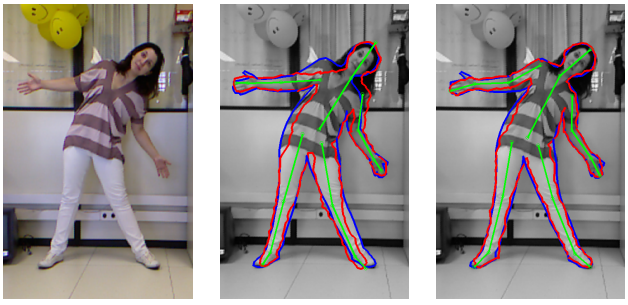


(a) (b) (c)

Fig. 8. (a) Our result. (b) Result obtained with [4]. (c) Result shown in Freifel's work [4], obtained using the Grab-Cut [11] algorithm.

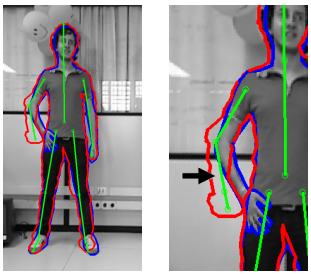
(a) (b) (c)

Fig. 9. (a) Input image. (b) Proposed model. (c) Blob, obtained using the Graph-Cuts [2] algorithm.



(a) (b) (c)

Fig. 10. (a) Input image. (b) Proposed model (full-automatic), using the 2D pose estimated from Kinect. (c) Proposed model (semi-automatic).



(a) (b)

Fig. 11. (a) Proposed model (full-automatic). (b) Zoom on image (a) - problem related to 2D pose estimation algorithms: body part not well centered over their associated shape.

(a) (b) (c)

Fig. 12. Experimental results using images with very low contrast.

IV. Conclusion

In this paper we propose a skeleton-based model to segment human in images (part of the work presented in this paper was accepted for publication on ICIP'12 [7]). The proposed approach does not use complex 3D models of the human form (as in [6]) or databases to learning appearance/shape/pose models (as in [4], [5], [9], [10]). Based on the provided skeleton, a graph is built around the expected contour region, and the silhouette of the person is obtained as a combination of maximum cost paths in the graph, where the weights of the edges are based on edge information, anthropometric distances and orientation constraints.

The experimental results showed that our method performs visually well for a variety of images, being able to handle non trivial images containing self-occlusions. When comparing to a competitive approach that also provides the segmentation of individual body parts [4], our method shows to produce more accurate (but less smooth) contours.

Future work will concentrate on exploring color information, including extensions to multi-view images and/or depth-color data (e.g. Microsoft's *Kinect* sensor), as well as quantitative evaluations of the proposed model.

References

[1] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pp. 1014–1021, 2009.

[2] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," *In Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 105–112 vol.1, 2001.

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

[4] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black, "Contour people: A parameterized model of 2d articulated human shape," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[5] D. M. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1408–1421, 2007.

[6] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *In Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 1381 – 1388.

[7] J. C. S. Jacques Junior, C. R. Jung, and S. R. Musse, "Skeleton-based human segmentation in still images (accepted for publication)," in *IEEE International Conference on Image Processing (ICIP)*, Orlando, Florida - USA, 2012, pp. 1 – 4.

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

[9] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[10] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," *Proc. of IEEE CVPR*, vol. 2, pp. 326–333, 2004.

[11] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, August 2004.

[12] A. R. Tilley, *The measure of man and woman - Human factors in design*. John Wiley & Sons, inc, 2002.