STRIPE BASED CLOTHES SEGMENTATION

Javier Lorenzo-Navarro, Modesto Castrillón-Santana, David Freire-Obregón, Enrique Ramón-Balmaseda

Instituto Universitario SIANI
Universidad de Las Palmas de Gran Canaria - Spain
javier.lorenzo@ulpgc.es

ABSTRACT

In this paper, a clothes segmentation method for fashion parsing is described. This method does not rely in a previous pose estimation but people segmentation. Therefore, novel and classic segmentation techniques have been considered and improved in order to achieve accurate people segmentation. Unlike other methods described in the literature, the output is the bounding box and the predominant color of the different clothes and not a pixel level segmentation. The proposal is based on dividing the person area into an initial fixed number of stripes, that are later fused according to similar color distribution. To assess the quality of the proposed method the experiments are carried out with the Fashionista dataset that is widely used in the fashion parsing community.

Index Terms— Clothes segmentation, people segmentation, fashion parsing

1. INTRODUCTION

Clothing segmentation and recognition has attracted the attention of Computer Vision community since years. One of the most widely used applications is as virtual mirror where augmented reality is achieved by means of person detection and overlapping garments. Moreover, the number of reliable applications in this area has increased due to the use of low cost RGBD cameras as Microsoft Kinect [1].

On the other hand, analysing the clothing that people wear gives valuable information in some applications as image or video demographic analytics. Among other cues, gender, age or social status can be obtained from the clothing [2, 3]. Besides clothing can be used in different contexts as a powerful aid. In the case of re-identification tasks, Satta et al. [4] propose a variation of their Multiple Component Matching framework named Multiple Component Dissimilarity (MCD) based on clothing attributes.

Different problems are tackled in the clothing analysis literature. One problem is related to clothes segmentation. The task aims at segmenting the clothes from the image, clustering those regions that correspond to the same garment. Another issue is clothing attribute classification, where each garment is described by different attributes such as color, pattern, neck type and sleeve among others. A last problem is clothing recognition that focuses on assigning garment categories like t-shirt, dress, trousers, etc. This problem is also referred as clothes or fashion parsing.

In this paper we concentrate on clothes segmentation. Different authors have already proposed several approaches for this task. The approach proposed by Gallagher and Chen [5] is based on a previous superpixel segmentation. Later, color and texture features are computed for each superpixel, obtaining a final clothes mask. Related to image retrieval, the approach proposed by Borras et al. [6] describes the upper body clothing making use of texture and color features where a previous split-and-merge process based on homogeneity measures groups the different clothing parts. Another proposal in this scenario of application is due to Weber et al. [7]. The authors make use of offline trained pose detector to deal with occlusions and different poses, before getting the clothing mask. An approach to segment garments in fashion databases is described by Manfredi et al. [8]. Different features such as color, texture and gradient information are fused by means of a Gaussian Mixture model. Background subtraction is the first step before extracting color and texture features from the clothing regions in the approach described by Yang and Yu [9].

We would like to highlight the works by Yamaguchi et al. [10, 11]. Their proposal makes a previous pose estimation and a superpixel segmentation prior to label the garments making use of a Conditional Random Field (CRF). Simo-Sierra et al. [12] also make use of CRF model for clothes parsing and introduce the concept of *clothelets* to encode the likelihoods for each garment and location in the body. Chen et al. [13] describe clothing making use of 11 attributes integrating a CRF to take into account the relation between some attributes. A similar work [14] makes use of LBP and HOG descriptors. Another remarkable work is proposed by Bossard et al. [15] where the upper torso is described making use of Random Forest.

In most of those works on fashion parsing, classification is done to the level of pixel or superpixel. With the rise in the use of Convolutional Neural Networks (CNN) [16] as object

This work has been partially funded by the Departamento de Informática y Sistemas of the Universidad de Las Palmas de Gran Canaria

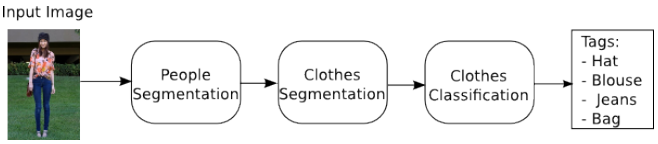


Fig. 1. General schema of a clothing classification system

detector other approaches can be introduced. Moreover, Hara et al. [17] make use of CNN to obtain the bounding box of the clothes that appear in the image.

In this work, a method for obtaining the bounding box and the predominant color of the upper and lower clothes is presented. These bounding boxes can feed a classification system, for example a CNN or SVM, to make up a fashion parsing system forming as the first stage in a more general system (Figure 1).

2. DATASETS

We have considered two popular databases: Color-Fashion and Fashionista. The Color-Fashion dataset [18] is composed of 2682 images of people wearing different outfits. Furthermore, it is a fashion-oriented dataset, which means that poses may have some variety but the number of people shown in such images may not exceed a person. This appropriately labelled database allows us to obtain valuable information to train our classifiers.

For test purposes, we have adopted the clothing Fashionista dataset proposed by Yamaguchi [10, 11], that has already been tested for the addressed problem on this paper. This dataset is composed of 700 images of standing persons wearing different outfits. For each image, pixels are labelled as background or as one of the 35 different garments. Similar to the Color-fashion dataset, it is a fashion-oriented dataset. Moreover, there is some poses variety and a large diversity of clothing elements (hats, purses, glasses, etc.).

3. PEOPLE SEGMENTATION

In our scenario, the background that surrounds the target people may hinder the accuracy of the classification process. For this reason, it is necessary a first stage in which people are completely or mostly segmented from their background. Thus the noise level added by the background can be reduced to achieve a better subsequent clothing segmentation.

Recently, Freire et al. [19] have proposed a novel algorithm based on the GrabCut [20] through a basic computation structure known as *trixel* (superpixels triangle). The trixels simplify the image data into perceptually meaningful atomic regions. This proposal is ideal for real-time systems, because it is much faster (up to 80% reduction of processing cost) than the original GrabCut (which computes pixels instead of trixels). For comparison purposes, we have made use of both techniques, considering pixels and trixels as input to GrabCut.

GrabCut is a segmentation technique that uses an input parameter, known as trimap, to initialize the probabilistic models. The trimap is a probability distribution that labels three regions of interest: those that are likely to be foreground, background and unknown respectively. This initialization process provides all the necessary data for the consequent segmentation process, and it is a critical stage. In order to compute these distributions two techniques are considered:

- A geometric technique that requires a face detector [21] for determining the position of the eyes. Then, this position is used to calculate an estimation of the different trimap regions based on equation 1.

$$\forall n \in [0,k], \quad pt(n) = (f_{dist} + w * \vec{v}) + \vec{p_e} * w_n \quad (1)$$

  Where $\vec{v}$ represents the vector from one eye to the other, w is the symmetric distribution to achieve equidistant points to both sides of the face, and w_n is the individual distribution of the weights to move on the y-axis. The vector $\vec{p_e}$ allows the 90 degrees rotation along the image for each point. The intersection of these points provides a suitable mask for the trimap selection.

- A Bayesian technique that considers a conditional probability map obtained from the Color-Fashion dataset [18]. Once again, the facial detector provides the eyes position in the image. Then, each image is rescaled and translated so that the middle eyes position is placed at a fixed location (x_0, y_0). According to this location, the different trimap regions are computed based on equation 2.

$$P(L|(x_n, y_n)) = \frac{P(L)P((x_n, y_n)|L)}{P(x_n, y_n)} > \varepsilon \quad (2)$$

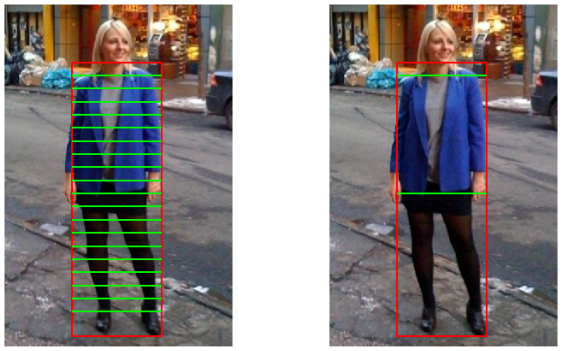
  Where L is the considered label (foreground, background or unknown), (x_n, y_n) is the normalized pixel position according to (x_0, y_0) and ε is the decision threshold.

Then, GrabCut exploits the provided trimap and labels the image into foreground and background. Figure 2 shows a result of the GrabCut based on the Bayesian trimap people segmentation.

4. CLOTHES SEGMENTATION

Clothes segmentation can take advantage of person morphology to obtain the regions that correspond to the different garments: blouse, t-shirt, pants, skirt and so on. Mostly, each piece of clothing has a single predominant color and print so a broad segmentation can be obtained by finding the pattern changes.

Fig. 2. Results of the segmentation process



(a) (b)

Fig. 3. Example of an initial (a) and final (b) division into horizontal stripes.

The proposed segmentation method starts by dividing the segmented area occupied by the segmented person into n horizontal stripes (Figure 3-a). Each stripe, S_i, is characterized by its color distribution, $P_i(c)$, that represents the probability of occurrence of color c in stripe S_i. Making use of the assumption that each piece of clothing has a predominant color and print, the method iteratively merges similar adjacent stripes. On each iteration, two adjacent stripes, S_i and S_j, are fused if their color distributions, $P_i(c)$ and $P_j(c)$, are similar. The similarity of the two distributions, $P_i(c)$ and $P_j(c)$, is measured using the Kullback-Leibler divergence, $D_{KL}(P_i||P_j)$ [22]. The expression of the Kullback-Leibler divergence for discrete probability functions in equation 3.

$$D_{KL}(P_i||P_j) = \sum_c P_i(c) \ln \frac{P_i(c)}{P_j(c)} \tag{3}$$

Given $D_{KL}(P_i||P_j)$ for each pair of adjacent stripes, the pair with minimum divergence, $d_{kl} = \min\{d_{ij} \forall \ S_i \, adjacent \, S_j\}$, is merged ($S'_i = S_k \cup S_l$). After merging two stripes, the color distribution for the new S'_i is computed. The method ends when the minimum divergence is greater than a defined threshold, $d_{kl} > \epsilon_h$.

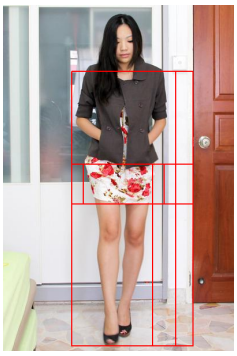


Fig. 4. Result of the clothing segmentation method.

Once the stopping criteria is reached, the region occupied by the segmented person has been divided into a set of horizontal stripes, $S_{final} = \{S'_1, \ldots, S'_h\}$, corresponding tentatively to each piece of clothing (Figure 3-b).

Sometimes the person segmentation process yields a container that includes background. In those situations, some of the horizontal stripes are a mix of piece of clothes and background. To remove the background, a similar process is carried out but for each horizontal stripe. Thus, each horizontal stripe is divided into m vertical stripes S_j^{vert}. As for the horizontal stripes, each vertical stripe, S_j^{vert}, is characterized by its color distribution, P_j^{vert}. Again a merging adjacent vertical stripes is carried out until the minimum Kullback-Leibler divergence is greater than a threshold, ϵ_v. After this second process of merging vertical stripes, the container of each piece of clothing can be obtained (Figure 4).

5. EXPERIMENTS

As we argued before, our proposal is based on two phases (Figure 1). The first one is the people segmentation process carried out to remove non-sensitive information. Then, the second stage allows us to obtain all the necessary information to segment the clothes.

The Jaccard index (JI) is considered to evaluate the segmentation quality. Let us consider Seg and GT as the result of the obtained and the ground truth segmentation respectively. One can directly apply the Jaccard measure to estimate the similarity between these segmentations by performing equation 4.

$$Jaccard\,Index = \frac{Seg \cap GT}{Seg \cup GT} \tag{4}$$

5.1. People Segmentation Results

We performed a quantitative comparison of the classical GrabCut and the trixel version. The tests were made from two points of view: the quality of the people segmentation using the ground truth images, and the processing time. In

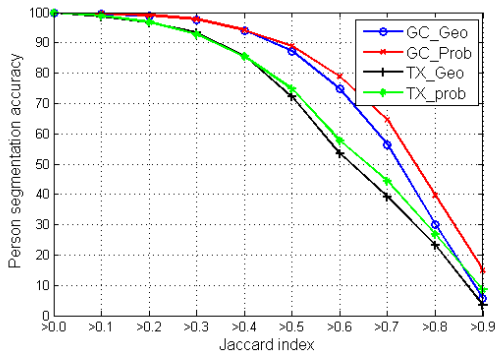


Fig. 5. People segmentation accuracy vs Jaccard index

Fig. 6. Result of the color classification.

total four different approaches are evaluated considering the algorithms and the trimap model:

- GC_Geo: A classical pixel GrabCut approach based on the computation of a geometrical trimap as input.
- GC_Prob: A classical pixel GrabCut approach which computes a probabilistic trimap as input.
- TX_Geo: A trixel GrabCut approach based on the computation of a geometrical trimap as input.
- TX_Prob: A trixel GrabCut approach which computes a probabilistic trimap as input.

Figure 5 shows the accuracy of the four different approaches under evaluation. These measures are taken considering the Jaccard index of the foreground pixels for each image. As can be seen in the graph, the probabilistic trimap outperforms the geometric trimap in every case. The improvement depends on the selected JI threshold. For example, choosing a JI threshold of 0.6, an approximately accuracy improvement of 4% can be appreciated for both, the GC and the TX approaches.

Table 1. A normalized speed comparison among the approaches.

GC_Geo	GC_Prob	TX_Geo	TX_Prob
1.0	0.9	0.2	0.2

As it happens in [19], the trixel approach does not improve the results of the GrabCut approach. The authors claim that there is a simplification of the process due to the fact that trixels only provide a mean value of the color inside them. For an example image of 270×349 pixels, while GrabCut needs 94230 pixels to work properly, the trixel version algorithm only uses 4403 trixels, just 4.6% of the original number of vertices. However, this reduction of information has a significant positive side in terms of speed. Table 1 shows the remarkable improvement over the classical pixel GrabCut. The trixel proposal is roughly a 80% faster than the original GrabCut.

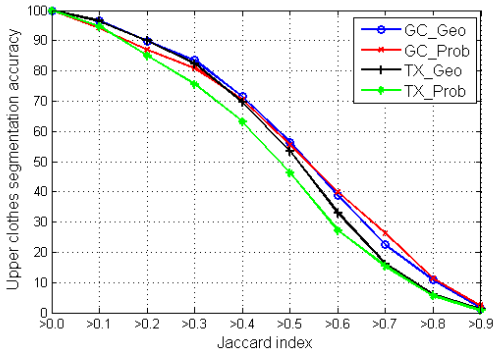
5.2. Clothes Segmentation Results

As stated in Section 4, the proposed clothes segmentation is based on the merge of similar color distribution stripes. For computing the color distribution in each horizontal and vertical stripe, $P_i(c)$ and $P_j^{vert}(c)$, a color classifier was previously trained with the Colorful-Fashion dataset, considering the following colors: *beige*, *black*, *blue*, *brown*, *gray*, *green*, *orange*, *pink*, *purple*, *red*, *white* and *yellow*. The classifier is a decision tree taking as input the YCbCr color space that gives better results than RGB or HSV color spaces for this experimental setup (Figure 6). Some garments such t-shirts, shorts or skirts leave part of the body exposed and the color classifier is not trained for skin. To avoid this, the person segmentation includes skin information and that it is used to assign the color *skin* to those areas.

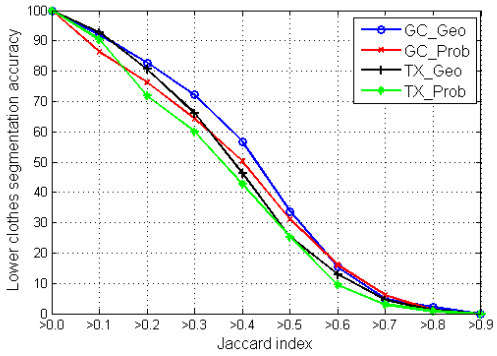
In our experiments we set the initial number of horizontal stripes (n) to 20, and the number of initial vertical stripes (m) for each horizontal stripe to 10. After testing with different thresholds that as stopping criteria for the merging process, they were fixed to $\epsilon_h = 0.6$ and $\epsilon_v = 1.4$.

The ground truth for comparing the results obtained with the proposed method are the upper (GT_{upper}) and lower (GT_{lower}) bounding boxes of the corresponding clothes. All the pixels labelled as any of the upper body garment (t-shirt, cardigan, blouse, etc.) or any of the lower body garments (shorts, skirt, pants, jeans, etc.) in the Fashionista dataset are considered the upper and lower clothes respectively. To compare the accuracy of the segmentation, the Jaccard index between the upper/lower bounding box and the closest horizontal/vertical stripe , $S'_{closest}$, is computed (eq. 4).

The results of the upper clothes segmentation are shown in Figure 7-a when the second stage of method (vertical stripes) is not implemented. The accuracy is plotted against the JI considering a correct segmentation if the index is higher than a given value. This index has already been used to compare clothing segmentation results. In [12] Simo-Serra et al. use the same dataset, Fashionista, to show the influence of pose estimation in the segmentation of the different garments that are labelled in the dataset. In their results, except for three

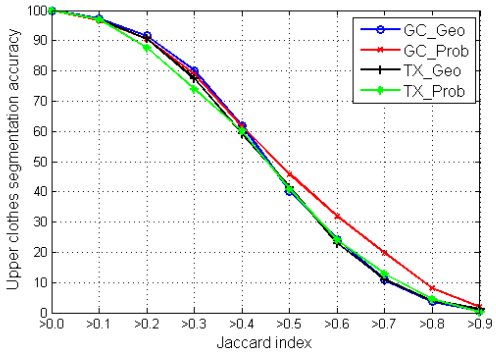


(a)

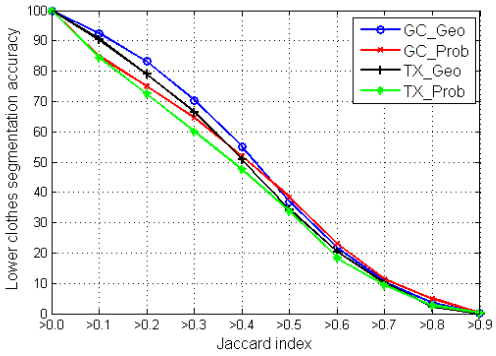


(b)

Fig. 7. Upper (a) and lower (b) clothes accuracy segmentation (horizontal stripes only) vs Jaccard index.



(a)



(b)

Fig. 8. Upper (a) and lower (b) clothes accuracy segmentation vs Jaccard index.

classes, the JI chosen is 0.3. Analysing the results in Figure 7-b it can be observed that for JI values under 0.5 all the person segmentation methods yield similar results except TX_prob that exhibits a worse performance. Taking into account the results of Simo-Serra et al. [12] a threshold of 0.3 can be set and obtaining an accuracy of 83.55%. With a more restrictive threshold of 0.5 the accuracy drops to 56.45%. The results for the lower clothes are shown in Figure 7-b. As can be seen, the GC_Geo approach gives better accuracy although the performance is worse than for upper clothes. This can be explained by the higher variability in appearance of the lower clothes (pants, skirt, shorts, stockings, etc.). As in the previous case, if the JI threshold is fixed to 0.3 the accuracy is 72.11% and it descends to 33.8% taking JI of 0.5.

When the clothes segmentation process includes the vertical stripe segmentation, results can be observed for the upper clothes in Figure 8-a. For lower JI values all the person segmentation methods exhibit similar performance but for values over 0.4 with the GC_Prob, higher accuracies are obtained. If the JI threshold is fixed to 0.3 the accuracy is 80.16% and falls to 45.87% if the JI threshold is raised to 0.5. Figure 8-b shows that the higher and more restrictive JI threshold is, the worse the accuracy is. Thus, for a JI threshold of 0.3 we get 70.37%, that is 10% lower than for upper clothes. The accuracy falls to 38.4% when JI threshold is 0.5. Unlike upper clothes, all the people segmentation methods present similar performance.

6. CONCLUSIONS

In this paper a method has been presented for clothing segmentation based on stripe merging with no previous pose estimation needed. Unlike other approaches, here the bounding boxes of the upper/lower clothes and their corresponding predominant colors are obtained and they can be used as the input for a posterior fashion parsing stage. The performance for upper clothes is higher than for lower clothes. In any case accuracy rates over 70% are achieved with JI equals to 0.3 that improves previous works on the same dataset and JI. An advantage of the proposed method is that it is not necessary a previous pose estimation because it relies on a person segmentation that is based on [19]. With respect to the effect of the people segmentation, the results are very similar. The trixel version reduces the processing cost at 80% with respect to the classical GrabCut.

Regarding to the people segmentation, we have introduced a probabilistic approach to compute the GrabCut trimap that outperforms the previous work in terms of segmentation accuracy and keeping similar processing cost. In

the case of trixels, the execution time varies significantly compared with the classical GrabCut.

7. REFERENCES

[1] Zhenglong Zhou, Bo Shu, Shaojie Zhuo, Xiaoming Deng, Ping Tan, and Stephen Lin, "Image-based clothes animation for virtual fitting," in *SIGGRAPH Asia 2012 Technical Briefs*, 2012, pp. 33:1–33:4.

[2] I. Kwak, A.C. Murillo, P. Belhumeur, S. Belongie, and D. Kriegman, "From bikers to surfers: Visual recognition of urban tribes," in *British Machine Vision Conference (BMVC)*, 2013.

[3] Z. Song, M. Wang, X.S. Hua, and S. Yan, "Predicting occupation via human clothing and contexts," in *IEEE International Conference on Computer Vision*, 2011.

[4] R. Satta, F. Pala, G. Fumera, and F. Roli, *Person Re-Identification*, chapter People search with textual queries about clothing appearance attributes, pp. 371–390, Springer, 2014.

[5] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] A. Borras, F. Tous, J. Llados, and M. Vanrell, "High-level clothes description based on colour-texture and structural features," in *1st Iberian Conference on Pattern Recognition and Image Analysis IbPRIA*, 2003.

[7] M. Weber, BÄuml, and R. M., Stiefelhagen, "Part-based clothing segmentation for person retrieval," in *International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2011.

[8] M. Manfredi, C. Grana, S. Calderara, and R. Cucchiara, "A complete system forgarment segmentation and color classification," *Machine Vision and Applications*, 2013.

[9] Ming Yang and Kai Yu, "Real-time clothing recognition in surveillance videos," in *18th IEEE International Conference on Image Processing*, 2011, pp. 2937–2940.

[10] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[11] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Retrieving similar styles to parse clothing," *Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[12] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "A High Performance CRF Model for Clothes Parsing," in *Proceedings of the Asian Conference on Computer Vision (2014)*, 2014.

[13] H. Chen, A. Gallagher, and B Girod, "Describing clothing by semantic attributes," in *European Conference on Computer Vision (ECCV)*, 2012.

[14] J. Lorenzo-Navarro, M. Castrillón Santana, E. Ramón-Balmaseda, and D. Freire-Obregón, "Evaluation of lbp and hog descriptors for clothing attribute description.," in *First International Workshop Video Analytics for Audience Measurement (VAAM)*, Sweden, August 2014.

[15] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool, "Apparel classification with style," in *11th Asian Conference on Computer Vision*, Korea, November 2012.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, vol. 86, pp. 2278 – 2324.

[17] Kota Hara, Vignesh Jagadeesh, and Robinson Piramuthu, "Fashion apparel detection: The role of deep convolutional neural network and pose-dependent priors," arXiv:1411.5319, November 2014.

[18] Si Liu, Jiashi Feng, C. Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan, "Fashion parsing with weak color-category labels," *Multimedia, IEEE Transactions on*, vol. 16, no. 1, pp. 253–265, 2014.

[19] D. Freire-Obregón, M. Castrillón-Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Automatic clothes segmentation for soft biometrics," in *Proceedings of IEEE International Conference on Image Processing*. IEEE, 2014, pp. 4972–4976.

[20] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut - interactive foreground extraction using iterated graph cuts," *Proceedings of ACM Siggraph*, 2004.

[21] M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, pp. 130–140, April 2007.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 2006.