AUTOMATIC CLOTHES SEGMENTATION FOR SOFT BIOMETRICS

D. Freire-Obregón, M. Castrillón-Santana, E. Ramón-Balmaseda and J. Lorenzo-Navarro

SIANI - Universidad de Las Palmas de Gran Canaria
Spain

ABSTRACT

During the last decade, researchers have verified that clothing can provide information for gender recognition. However, before extracting features, it is necessary to segment the clothing region. We introduce a new clothes segmentation method based on the application of the GrabCut technique over a trixel mesh, obtaining very promising results for a close to real time system. Finally, the clothing features are combined with facial and head context information to outperform previous results in gender recognition with a public database.

Index Terms— Clothing, GrabCut, Trixels, Gender recognition, TriToM

1. INTRODUCTION

Among the different demographic variables, gender is commonly studied by marketers. Certainly, the problem of automatic gender classification has received notorious attention in the Computer Vision literature. Most solutions have been designed focused on the facial pattern, being of particular interest those recently devoted to large datasets [1, 2, 3]. However, different authors have remarked the benefits of the combination of clothing in addition to body [4, 5] and face [6] information to perform a better automatic gender classification [7].

For less restricted scenarios, compared to Li [8], Chen et al. [6] suggested the combination of clothing and face for this problem, extracting automatically the torso of an individual using a pose estimation approach, computing later features based on SIFT descriptors. Their results in their public dataset, that we refer as ClothesDB, reported a notorious improvement when facial and clothing information was fused.

In this paper, we study an approach that will make use of the face, the local head context and the clothing information of the upper torso for this problem. Its performance is later compared with previous results on the ClothesDB, and state of the art gender recognizers for large datasets.

The contributions of the paper are: 1) the introduction of *trixels* to simplify the image data into perceptually meaningful atomic regions, 2) the description of the novel clothes segmentation technique based on face detection information, trixels and an adaptation of the GrabCut algorithm [9] reducing the processing cost up to 80% of the overall segmentation procedure, and 3) the combination of facial and local head context features with automatically extracted clothing features to improve gender classification accuracy in the ClothesDB.

2. TRITOM

We propose a new method for generating trixels (”Triangular Superpixels”) which is fast, memory efficient and exhibits accurate boundary adherence. The trixels are closely related to the popular and well-known superpixels [10]. However, the most visible difference is the geometrical shape; while superpixels are squares only at the beginning of the process, the geometric shape of the trixel is the triangle. As superpixels, trixels are computed all over an input image. We call the mesh of all obtained trixels the Trixels Topological Map (TriToM).

2.1. TriToM creation

Similarly to superpixels [10], there is an initial geometric configuration where each pixel belongs to a higher order structure. In our case, this higher order structure is the previously mentioned trixel. Nonetheless, some interesting differences between these two approaches must be addressed:

- Unlike to superpixels, the trixels mesh (TriToM) topology is not homogeneous at any time during the process. In spite of this non homogeneity, each trixel keeps three neighbors at most.

- The vertices of each trixel are not randomly placed in the initial step. Neither do they follow a precise geometrical path as the SLIC (”Simple Linear Iterative Clustering”) superpixels version does [11]. The location of the vertices depends on the content of the processed image.

- Trixels are not processed individually with the aim to find borders between foreground and background. This task is done following a search tree process.

Work partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering and the Computer Science Department at ULPGC.

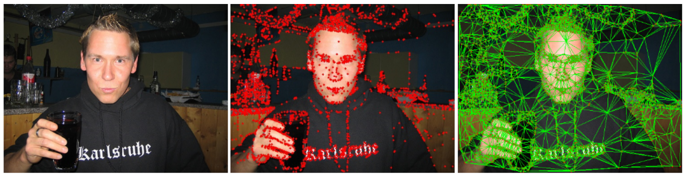

Fig. 1. Process from the source image to TriToM through the DMUM map.

- Contrary to superpixels, the trixels do not change their boundaries. The original shape for each trixel remains unchanged during the clustering process.

The first part of the TriToM creation is related to the image analysis through the DMUM (Distance Maps from Unthresholded Magnitudes) as a preprocessing step [12]. The local minimum points are placed by computing the DMUM algorithm all over the image. Moreover, these points are used as an input parameter for the mesh creation. By using DMUM, we ensure that the relevant image data is kept in the mesh geometry; as it affects the number, the size and the density of the trixels within the mesh. DMUM output values represent the size of the smaller area around each pixel enclosing a relevant amount of accumulated magnitude values that depend solely on the image being studied. A single pass on the magnitude values image is needed, and the integral image is used to speed up area sums.

Once the DMUM map has been obtained, the trixels mesh (TriToM) is built up by the Delaunay triangulation using the local minimum values of the map as input. The Delaunay triangulation leads the process to generate the image TriToM. These trixels group pixels into perceptually meaningful atomic regions, which replace the rigid structure of the classical pixel grid. Figure 1 shows the process to create the TriToM; the image on the left corresponds the source image, the image on the center is the computed DMUM map (each red dot belongs to a local minimum value), and the image on the right is the resulting trixels mesh or TriToM.

3. CLOTHES SEGMENTATION

In order to proceed with the segmentation process, the GrabCut algorithm has been adapted to work with trixels instead of using pixels. In other words, TriToM is the input graph of the GrabCut algorithm. GrabCut is a technique that includes probabilistic color models to achieve a robust segmentation. This technique uses a parameter, the trimap, that initializes the probabilistic models. The trimap works as an input in order to initialize the probabilistic models. It is defined as $T = \{T_U, T_F, T_B\}$ and it consists on an initial mask in which regions are marked as unknown (T_U), foreground (T_F) or background (T_B). Thus, those trixels whose centroids are inside the T_U region belongs to the initial unknown cluster. This unknown cluster will disappear as far as their trixels are assigned to the T_F or the T_B clusters. Hence, the initialization process provides all the necessary data for the segmentation process, and it is a critical stage.

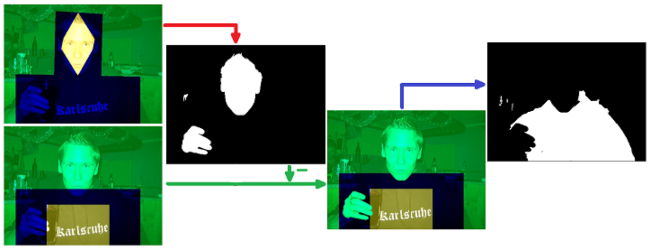

Fig. 2. The geometric mask is automatically computed from the eyes position, and later used to generate the trimap. This occurs for both the skin segmentation process (red line) and the clothes segmentation process (green line). The skin segmentation results is used as input to define the clothes trimap.

For our task, firstly an adaptive geometric model defines the trimap by specifying a mask around the object of interest. Trixels inside the mask are marked as unknown or foreground, and trixels outside the mask are marked as background. The mask is computed based on the information provided by the ENCARA2 face detector [13], that provides face and facial features detection (eyes, nose, mouth), and passes this data through the geometric model in order to generate the input mask. The mask depends on the area enclosed between several geometric points. Thus, given the distance f_{dist} between the eyes, the k geometric points $pt(0), pt(1), ..., pt(k)$ are defined as:

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

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. Figure 2 shows how this trimap has been defined for the clothes extraction. The background region (in green) and the foreground region (in yellow) help to spread the unknown region trixels (in blue) into these two main clusters during the segmentation stage.

After defining the geometric model, K components of the Gaussian Mixture Models (GMM) are created and initialized for both; the foreground region and the background region respectively. This means that $2K$ components are created to divide both regions into K trixels clusters. Then, the Gaussian components are initialized from the trixels data in each cluster. Each trixel contributes by only providing the mean color of every pixel inside them, no matter the trixels size. The idea is to seek low variance Gaussian components which are necessary for a good separation between foreground and

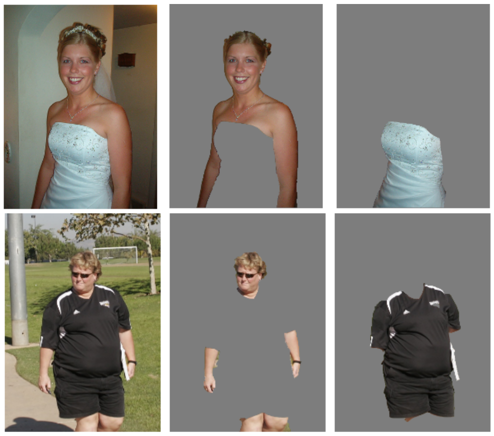

Fig. 3. Automatic skin (second column) and clothes (third column) extraction from two samples of the PASCAL dataset [14].

background. Moreover, this technique uses the eigenvector of the mean color covariance matrix to determine well separated clusters.

3.1. TriToM segmentation

The segmentation process groups the trixels into clusters in order to extract the foreground. The following steps are repeated until the end of the segmentation procedure. The reason of this iterative process is that some trixels may move from the foreground class to the background and vice versa. Therefore the Gaussian distribution may change and it is necessary to update the GMMs to reflect the new color distribution for both, foreground and background.

During the first step, each trixel in the foreground class is assigned to the most likely Gaussian component in the foreground GMM. Similarly, each trixel in the background class is assigned to the most likely Gaussian component in the background GMM. Once the trixels have been clustered, the GMMs are dismissed and new GMMs are generated based on the learning process from the new trixels distribution.

In the second step, the graph is filled with the information computed through the new GMMs. In the original GrabCut algorithm, the nodes (compounded by pixels) are considered the vertices of the graph and they are connected by edges. Also, the necessary GrabCut parameters (T-Links and N-Links) are created from this vertex information. However, this conception changes in our new approach. Pixels are no longer the processing atomic element, but trixels. Hence, vertices are still part of the graph but only on a physical way, they are not functional. In other words, vertices keep together the structure of the graph, but the necessary data to generate the T-Links and the N-Links are not provided by these vertices, they are provided by the trixels instead. T-Links and N-Links are the input parameters for the maximum flow algorithm. N-Links connect trixels in the 3-neighborhood. These links describe the penalty for placing a segmentation boundary between the neighboring trixels. This penalty must be very high in regions of low gradient and low in regions of high gradient, where the figure edges are located. The equation for placing an N-Link between trixels T_i and T_j is:

$$N(T_i, T_j) = \frac{50}{dist(T_i, T_j)} \mathrm{e}^{-\beta \| h_{T_i} - h_{T_j} \|^2} \tag{2}$$

where $\beta = \frac{1}{2 \times mean(\sqrt{\| h_{T_i} - h_{T_j} \|})}$ for each pair of neighboring trixels, h_{Ti} is the color provided by the trixel and $dist(T_i, T_j)$ is the euclidean distance between the two of neighboring trixels. On the other hand, T-Links connect each trixel to the background and foreground nodes. These links describe the probability for a trixel to belong to the background or the foreground. In our approach, as in GrabCut, this probability is contained in the previously trained GMMs. Thus, there are two T-Links for each trixel (one per group) and the weight of these links depends on the state of the trimap. If a trixel is tagged as definitely background or foreground, then the weight of the link assures that the trixel belongs to the appropriate group by using the largest weight in the graph considering the set of all edges joining neighboring trixels. For unknown trixels, probabilities obtained from the GMMs are used to set the weights.

In the final step, the original GrabCut uses the max flow algorithm to build search trees from the source (foreground group) to the sink (background group). The max flow algorithm has been adapted in order to work with trixels instead of pixels, but it follows the same procedure as in [15]; there are three stages repeated until all the edges between the search trees are all saturated (an edge is saturated when its N-Link value is 0). Those stages are known as growth stage, augmentation stage and adoption stage. In the growth stage the active trixels explore adjacent non-saturated edges in order to find an augmentation path when both search trees encounter. Then, the augmentation stage push through the largest flow possible through the augmentation path and some edge(s) in the path can become saturated. When this happens, the search trees can be broken into forests. Finally, the adoption stage restores again the original search threes. At the end of the process, the saturation of the edges provides the segmentation path for each group.

4. EXPERIMENTAL RESULTS

4.1. Clothes extraction performance

We performed a quantitative comparison of the classical GrabCut algorithm and the new proposal using a subset of the PASCAL database [14]. Only pictures where people appear were considered. ENCARA2 was used as a filter to detect faces in the dataset. Hence, it found 1487 images in the database. The tests were made from two points of view;

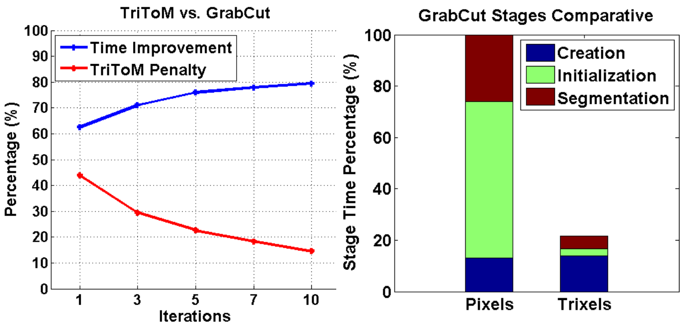

Fig. 4. Temporal improvement of using pixels or trixels as GrabCut input. On the left image, the blue line stands for the time improvement while the red line stands for the penalty influence for computing the trixels mesh at the beginning of the process. A speed comparative between both approaches considering each GrabCut stage is presented on the right image.

first of all, we studied the quality of the segmentation process using the ground truth images provided by the database. Secondly, the performance of the algorithms was studied considering the number of iterations of the segmentation process. Table 1 compares the segmentation accuracy results of both approaches. These measures are taken considering the classification of each image pixel. As can be seen, the proposed approach does not improve the results of the standard GrabCut; in fact it is a 6% worse. An explanation for these rates is that GrabCut works with pixels, while TriToM works with trixels. In other words, 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 TriToM 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. Figure 4 shows the remarkable improvement over the classical GrabCut. The new proposal is between a 60% and a 80% faster than the original GrabCut. We evaluate below the use of the obtained segmentation results for gender recognition.

Table 1. Segmentation performance comparison.

	Pixels GrabCut	TriToM GrabCut
TPR	79%	71%
TNR	84%	80%
FPR	16%	20%
FNR	21%	29%
Accuracy	81.5%	75.5%

4.2. Gender classification

For gender classification, we have adopted the ClothesDB proposed by [6], that has already been tested for this problem combining clothes and facial information. Even though in the original paper there are not many details related to the experiment, with our approach we have repeated 10 times a 5-fold cross-validation experiment.

For comparative purposes, we include the results achieved by Chen et al. [6], but also the results achieved by a recent proposal that combines facial and local head context information [3]. This particular approach is trained with faces taken from The Images of Groups dataset [16], and is taken as source to combine in the final test with clothing information. Table 2 compares the results achieved.

Reference	Data	Accuracy
[6]	Face	71.5%
	Clothing	81%
	Face+Clothing	84.9%
[3]	Face HOG	78.3%
	Face LBP	82.5%
	HS HOG	75.9%
This paper	Clothing (LBP)	67.3%
	Clothing (HOG)	83.9%
	Clothing (HOG) + HS	84.5%
	Clothing + HS + Face (LBP + HOG)	86.8%

Table 2. Gender recognition accuracy in ClothesDB. HS stands for head and shoulders.

As can be appreciated the best results are achieved when several cues are considered; clothing, head and shoulders, and facial information. Local descriptors such as LBP [17] and HOG [18] have been used to characterize each cue. As happens in Chen experiments [6], the clothing features alone are not good enough to define the gender of a person with a great accuracy due to the fashion culture. The use of the face pattern alone reported an interesting accuracy when trained with a large dataset, but could not cope with previous results combining facial and clothing information. However, our proposed final combination of patterns and features outperformed clearly previous literature results.

5. CONCLUSIONS

In this paper, we have presented an automatic clothes segmentation approach close to real-time, that starting from a trixels based image simplification, combines facial information to define a Grabcut inspired procedure for this task. In a quantitative comparison within PASCAL, reported a 6% worse accuracy, but requiring just 20% of the processing time.

Finally, the automatically clothes mask extracted are used to compute local descriptors that fused with face and local head context features are able to outperform previous gender classification accuracy on the ClothesDB. These results suggest the information provided by clothing attributes for automatic gender recognition.

6. REFERENCES

[1] Luis A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1422–1427, 2010.

[2] Juan Bekios-Calfa, José M. Buenaposada, and Luis Baumela, "Revisiting linear discriminant techniques in gender recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 858–864, April 2011.

[3] M. Castrillón, J. Lorenzo, and E. Ramón, "Improving gender classification accuracy in the wild," in *18th Iberoamerican Congress on Pattern Recognition (CIARP)*, 2013.

[4] Liangliang Cao, Mert Dikmen, Yun Fu, and Thomas S. Huang, "Gender recognition from body," in *Proceedings of the 16th ACM international conference on Multimedia*, 2008.

[5] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *International Conference on Computer Vision*, 2011.

[6] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," in *European Conference on Computer Vision*, 2012.

[7] Deng Cao, Cunjian Chen, D. Adjeroh, and A. Ross, "Predicting gender and weight from human metrology using a copula model," in *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, Sept 2012, pp. 162–169.

[8] B. Li, X. Lian, and B. Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 1, no. 76, pp. 18–27, 2012.

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

[10] G. Mori, X. Ren, A. Efros, , and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2004.

[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 22742282, 2012.

[12] Luis Antón-Canalís, Mario Hernández-Tejera, and Elena Sánchez-Nielsen, "Distance maps from unthresholded magnitudes," *Pattern Recognition*, vol. 45, no. 9, pp. 3125 – 3130, 2012.

[13] 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.

[14] C. K. I. Williams J. Winn M. Everingham, L. Gool and A. Zisserman, "The pascal visual object classes (voc) challenge," in *International Journal of Computer Vision*, 2009, vol. 88, pp. 303–338.

[15] Y. Boykov and M. Joly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," *IEEE International Conference on Computer Vision (ICCV)*, pp. 105–112, 2001.

[16] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, p. 886893.