MULTI-CLASS SEMANTIC SEGMENTATION OF FACES

Khalil Khan, Massimo Mauro, Riccardo Leonardi

Department of Information Engineering, University of Brescia, Italy

ABSTRACT

In this paper the problem of multi-class face segmentation is introduced. Differently from previous works which only consider few classes - typically skin and hair - the label set is extended here to six categories: skin, hair, eyes, nose, mouth and background. A dataset with 70 images taken from MIT-CBCL and FEI face databases is manually annotated and made publicly available[1]. Three kind of local features - accounting for color, shape and location - are extracted from uniformly sampled square patches. A discriminative model is built with random decision forests and used for classification. Many different combinations of features and parameters are explored to find the best possible model configuration. Our analysis shows that very good performance ($\sim 93\%$ in accuracy) can be achieved with a fairly simple model.

1. INTRODUCTION

Pixel-wise semantic segmentation is a critical topic of mid-level vision which aims at jointly categorizing and grouping image regions into coherent parts. Extensive research work has been carried out on the subject, mainly driven by the PASCAL VOC segmentation challenge [1]. Notwithstanding, a limited number of works specifically focus on faces.

Indeed, face labeling is potentially of interest in many situations. Huang et al. [2] showed that simple learning algorithms could be used to predict high-level features, such as pose, starting from the labeling of a face image into hair, skin and background regions. In their vision, intermediate level features such as segmentations, provide important information for face recognition and are extremely useful in estimating other charateristics such as gender, age, color of hair, color of skin, etc. Psychology literature seems to confirm their claim, as important facial features extracted from face regions (forehead, hair) are shown to be informative for the human visual system in order to recognize the face identity [3, 4].

Moving to different application scenarios, hair modelling, synthesis, and animation have already become active research topics in computer graphics [5, 6]. Moreover, face processing and enhancement applications such as skin smoothing [7], skin color beautification [8] and virtual make-up [9] began to appear in literature. In all such applications the precise knowledge - at pixel level - of face segments is of crucial importance.

Fig. 1. Face segmentation as produced by our algorithm

1.1. Related work

Several authors have built systems for segmenting hair, skin, and other face parts [2, 10–13] The work of Yacoob and Davis [10] is the first work specifically addressing hair labeling. The authors build a Gaussian Mixture Model (GMM) for hair color and then adopt a region growing algorithm to improve the hair region. Lee et al. [11] extend the GMM approach by learning six distinct hair styles, and other mixture models to learn color distributions for hair, skin, and background. Huang et al. [2] use a superpixel-based conditional random field (CRF) [14] trained on images from LFW dataset [15] to disambiguate among the same classes. Scheffler et al. [12] learn color models for hair, skin, background and clothing, and also introduce a spatial prior for each label. They combine this information with a CRF that enforces local label consistency. Finally, Kae et al. [13] propose a GLOC (GLObal and LOCal) model that combines the strengths of CRFs and Shape Boltzmann Machines [16] to jointly enforce local consistency and a proper global shape structure. To our knowledge, this is the best-performing algorithm for hair-skin-background segmentation to date.

With respect to state of the art, extending the face categories into more semantic classes may open new research scenarios and enhance the performance and the flexibility of most previously cited applications. E.g., an application for skin beautification could certainly benefit of a method which

[1]The labeled dataset is downloadable at `http://massimomauro.github.io/FASSEG-dataset/`

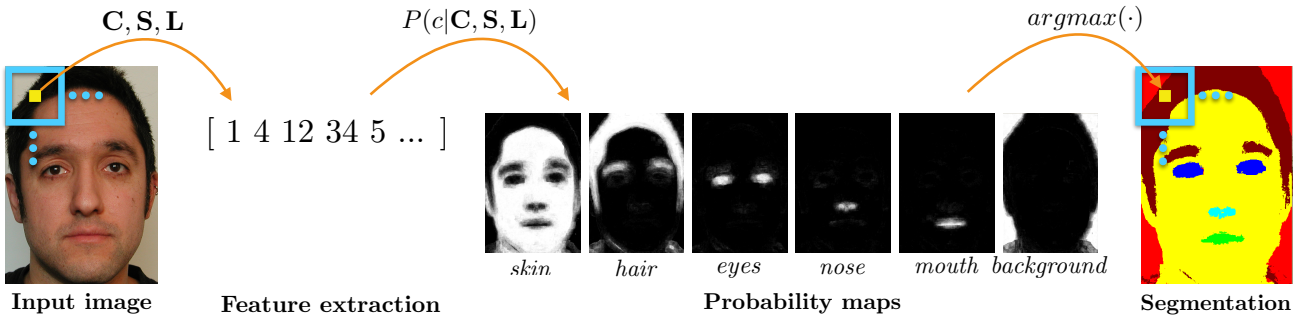

Fig. 2. Schema of the proposed algorithm described in Section 2.

disambiguates "real" skin from mouth and eyes. For these reasons, in our work we extend the label set to six classes: skin, hair, eyes, nose, mouth, and background.

Differently from several previous works which use a generative mixture model approach, we purely rely on labeled data and build a discriminative model by using a random forest [17] classifier. We classify the image content considering square patches as processing primitive. We adopt 3 kind of local features, accounting for color, shape and location. The spatial cue is combined in two different settings, explained in Section 2. We investigate the impact of each feature, of its parameters and of the spatial combination setting in order to find the best possible configuration. We build - and make publicly available - a dataset of 70 manually labeled images taken from the MIT-CBCL [18] and the FEI [19] face databases[1]. We analyse the different settings and show that very good performance is obtained with a fairly simple model.

The rest of the paper is arranged as follows: in Section 2 we explain our algorithm, in Section 3 we present the dataset and analyse experimental results, while Section 4 draws conclusions and discusses future work.

2. PROPOSED ALGORITHM

A schema of the proposed algorithm is in Fig.2. We divide the presentation of the algorithm in two parts: patches and features extractions are described in Subsection 2.1, classification and combination with spatial information are explained in Subsection 2.2.

2.1. Patches and Feature Extraction

Many of the semantic segmentation algorithms work at pixel or super-pixel level. Here we use square patches as processing primitive: we classify the image content of every patch and transfer the labeling to the center pixel of the patch. Our approach enjoy some benefits: information contained in these patches is more comprehensive than a single pixel. At the same time, every pixel is classified individually, differently from a super-pixel approach where a mistake may compromise the classification on the whole super-pixel region. In both training and testing we rescale the original images to have a constant height $H = 512$ pixels while width W is varied accordingly to keep the original image ratio. As a result, the type of content for a given patch dimension is comparable for different face images.

We use color and shape local features for our classification, combined with spatial information. As color features we adopt HSV color histograms: hue, saturation, and variance histograms are concatenated to form a single feature vector. We explore different parametrizations for the patch dimension (D_{HSV}= 16×16, 32×32 and 64×64) and for the number of histogram bins (N_{bins} = 16, 32 and 64). For each patch we get a feature vector $f^{16}_{HSV} \in R^{48}$, $f^{32}_{HSV} \in R^{96}$, or $f^{64}_{HSV} \in R^{192}$ depending on N_{bins}.

To account for shape information we extract the widely used HOG feature [20], changing the patch dimension among D_{HOG} = 16×16, 32×32 and 64×64. Using these values each patch generates feature vectors $f^{16\times16}_{HOG} \in R^{36}$, $f^{32\times32}_{HOG} \in R^{324}$, and $f^{64\times64}_{HOG} \in R^{1764}$ respectively.

As spatial informations we use the relative location of the pixel. Given a pixel at position (x, y) the relative location is defined as $f_{loc} = [x/W, y/H] \in R^2$.

2.2. Classification with spatial information

Since the classification is performed indipendently at every location, it consists at labeling every pixel with its maximum-probability class:

$$\hat{c} = \arg\max_{c \in C} p(c|\mathbf{C}, \mathbf{S}, \mathbf{L})$$

where $C = \{skin, hair, eye, nose, mouth, background\}$ and random variables $\mathbf{C}$, $\mathbf{S}$, and $\mathbf{L}$ are the features f_{HSV} (Color), f_{HOG} (Shape) and f_{loc} (Location) respectively.

We investigate two different settings to integrate the spatial information into the classification: as *feature concatenation* and as *spatial prior*. In the first case, the 2D feature f_{loc} is concatenated to f_{HSV} and f_{HOG} in a unique feature vector which is given as input to the classifier. In the second case,

Fig. 3. Example of segmentation results. Labeled ground truth on the second line, algorithm output on the third.

Features	Settings	Accuracy
color + location	$\mathbf{D_{HSV} = 16 \times 16}$, $N_{bins} = 16$, FC-setting	$\mathbf{92.27}\%$
color + location	$\mathbf{D_{HSV} = 32 \times 32}$, $N_{bins} = 16$, FC-setting	91.70%
color + location	$\mathbf{D_{HSV} = 64 \times 64}$, $N_{bins} = 16$, FC-setting	90.25%

Table 1. Impact of D_{HSV} color parameter

Features	Settings	Accuracy
color + location	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 16}$, FC-setting	92.03%
color + location	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 32}$, FC-setting	$\mathbf{92.27}\%$
color + location	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 64}$, FC-setting	91.71%

Table 2. Impact of N_{bins} color parameter

f_{loc} is used to estimate a spatial prior $p(c|\mathbf{L})$ and then the classification is performed as:

$$\hat{c} = \arg\max_{c \in C} p(c|\mathbf{L}) \cdot p(c|\mathbf{C}, \mathbf{S})$$

We use random forests to train the model, exploiting the C++ ALGLIB [21] implementation.

3. EXPERIMENTS

Our experiments are presented here. The experimental setup is explained in Subsection 3.1, while Subsection 3.2 and 3.3 contain the analysis and the discussion of results.

3.1. Experimental setup

The dataset we use for training and evaluation is made of 70 frontal face images, taken from the MIT-CBCL and FEI databases. The faces present a moderate degree of variability, as we included people of different ethnicity, gender, and age. Moreover, faces are not perfectly aligned in position and scale. This makes the algorithm suitable for performing face segmentation on the bounding-boxes derived from a previous face detection. We select a random subset of 20 images for extracting patches during training, while the remaining 50 images are used for testing. Accuracy is used as performance metric.

3.2. Results

Impact of HSV parameters. The HSV color feature has two important parameters to be considered: the patch dimension D_{HSV} on which the histogram is computed and the number of bins N_{bins} of the histogram itself. To evaluate the impact of both, a first stage of experiments is performed by only using location and color features and ignoring shape. We consider all the 9 combination of values from the sets $D_{HSV} = \{16\times16, 32\times32, 64\times64\}$ and $N_{bins} = \{16, 32, 64\}$. We find that the best accuracy - 92.27% - is achieved with $D_{HSV} = 16 \times 16$ and $N_{bins} = 32$. Results are reported in Table 1 and

Features	Settings	Accuracy
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 16 \times 16}$, FC-setting	92.44%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 32 \times 32}$, FC-setting	92.82%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 64 \times 64}$, FC-setting	$\mathbf{92.95\%}$

Table 3. Impact of D_{HOG} shape parameter

Features	Settings	Accuracy
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 16 \times 16$, **FC-setting**	92.44%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 16 \times 16$, **SP-setting**	91.20%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 32 \times 32$, **FC-setting**	92.82%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 32 \times 32$, **SP-setting**	91.27%
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 64 \times 64$, **FC-setting**	$\mathbf{92.95\%}$
color + shape + location	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $D_{HOG} = 64 \times 64$, **SP-setting**	91.67%

Table 4. Impact of spatial setting

2. Feature concatenation (FC) setting is used for the inclusion of spatial information.

Impact of HOG feature and parameters. We then introduce HOG feature and run a second stage of experiments to evaluate the impact of the patch dimension D_{HOG}. We find that the best accuracy - 92.95% - is achieved with $D_{HOG} = 64 \times 64$ and FC-setting. Results are reported in Table 3.

Impact of spatial setting. We run all previous tests by using both the feature concatenation and the spatial prior (SP) settings for the location feature. In Table 4 we show the results obtained in the two cases with different feature and parameter configurations. Results highlight that the FC-setting constantly outperforms the SP-setting in terms of accuracy.

3.3. Discussion of results.

A few considerations emerge from the results. The first is that the right choice of features and parameters matters. Among all the configurations we experimented there is indeed a big difference between the worst - which has an accuracy of 79.89% (not shown here) - and the best, which achieves 92.95% and is obtained with $D_{HSV} = 16 \times 16$, $N_{bins} = 32$, $D_{HOG} = 64 \times 64$, and FC-setting.

A second observation is that the HOG feature boosts the accuracy from 92.27% to 92.95%. This may seem a small improvement, but it corresponds to a 9% reduction in error rate. Moreover, the classes which benefit the most from the introduction of HOG are eyes, nose and mouth, which are mostly distinguishable from their shape. Since these classes are the least frequent, they have a smaller impact on the accuracy.

A third and last note is that, though the overall accuracy of FC spatial setting is always higher, for certain classes the SP-setting performs better. In Table 5 and 6 we show the confusion matrices corresponding to the best possible configuration in the FC and SP settings respectively. Such results highlight that if we are more interested in the skin region, the SP-setting could be preferable, as it improves the accuracy for the skin category from 93.39% to 96.61%.

		Predicted class					
		skin	*hair*	*eyes*	*nose*	*mouth*	*back*
True class	*skin*	**93.39**	5.19	0.64	0.06	0.26	5.19
	hair	3.17	**95.14**	0.11	0.00	0.00	1.58
	eyes	15.24	2.55	**82.2**	0.00	0.00	0.00
	nose	68.08	0.29	1.24	**29.83**	0.54	0.00
	mouth	29.47	0.02	0.26	0.00	**70.23**	0.00
	back	2.46	5.02	0.00	0.00	0.00	**92.50**

Table 5. Best results in the FC spatial setting

		Predicted class					
		skin	*hair*	*eyes*	*nose*	*mouth*	*back*
True class	*skin*	**96.61**	2.75	0.3	0.00	0.14	0.19
	hair	7.59	**91.84**	0.01	0.00	0.00	0.56
	eyes	35.77	1.64	**62.59**	0.00	0.00	0.00
	nose	97.00	0.00	0.00	**3.00**	0.00	0.00
	mouth	53.3	0.00	0.00	0.00	**46.97**	0.00
	back	3.65	12.64	0.00	0.00	0.00	**83.71**

Table 6. Best results in the SP spatial setting

4. CONCLUSION AND FUTURE WORK

In this work we introduce the problem of multi-class semantic segmentation of faces. For the purpose, we collect and make publicly availble a dataset of 70 face images taken from FEI and MIT-CBCL face databases. We use such database for training a discriminative model and propose a simple yet effective algorithm for segmentation. Exploring various configurations of color, shape and location features we can achieve a pixel labeling accuracy of 92.95%.

A few research directions are planned as future work. First, we aim to extend our dataset and method to support a higher level of variability, especially regarding face pose and orientation. Second, we intend to integrate our estimated pixel probabilities into a CRF with smoothness-based priors, in order to enhance the local labeling consistency. Lastly, we plan to combine our method with the result of rigid part detectors to improve the accuracy for the most problematic classes, such as eyes, nose and mouth.

5. REFERENCES

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[2] Gary B Huang, Manjunath Narayana, and Erik Learned-Miller, "Towards unconstrained face recognition," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[3] Graham Davies, Hadyn Ellis, and John Shepherd, *Perceiving and remembering faces*, Academic Press, 1981.

[4] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[5] Jonathan T Moon and Stephen R Marschner, "Simulating multiple scattering in hair using a photon mapping approach," in *ACM Transactions on Graphics (TOG)*. ACM, 2006, pp. 1067–1074.

[6] Kelly Ward, Florence Bertails, T-Y Kim, Stephen R Marschner, M-P Cani, and Ming C Lin, "A survey on hair modeling: Styling, simulation, and rendering," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 2, pp. 213–234, 2007.

[7] Changhyung Lee, Morgan T Schramm, Mireille Boutin, and Jan P Allebach, "An algorithm for automatic skin smoothing in digital portraits," in *Proceedings of the 16th IEEE international conference on Image processing*. IEEE Press, 2009, pp. 3113–3116.

[8] Chih-Wei Chen, Da-Yuan Huang, and Chiou-Shann Fuh, "Automatic skin color beautification," in *Arts and Technology*, pp. 157–164. Springer, 2010.

[9] Lin Xu, Yangzhou Du, and Yimin Zhang, "An automatic framework for example-based virtual makeup," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3206–3210.

[10] Yaser Yacoob and Larry S Davis, "Detection and analysis of hair," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 7, pp. 1164–1169, 2006.

[11] Kuang-chih Lee, Dragomir Anguelov, Baris Sumengen, and Salih Burak Gokturk, "Markov random field models for hair and face segmentation," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[12] Carl Scheffler and Jean-Marc Odobez, "Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps," in *British Machine Vision Association-British Machine Vision Conference*, 2011.

[13] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller, "Augmenting crfs with boltzmann machine shape priors for image labeling," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2019–2026.

[14] John Lafferty, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[16] SM Ali Eslami, Nicolas Heess, Christopher KI Williams, and John Winn, "The shape boltzmann machine: a strong model of object shape," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 155–176, 2014.

[17] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] MIT Center for Biological and Computational Learning (CBCL), "Mit-cbcl database," `http://cbcl.mit.edu/software-datasets/FaceData2.html`.

[19] Centro Universitario da FEI, "Fei database," `http://www.fei.edu.br/~cet/facedatabase.html`.

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

[21] Sergey Bochkanov, "Alglib," `http://www.alglib.net`.