HEAD POSE ESTIMATION THROUGH MULTI-CLASS FACE SEGMENTATION

Khalil Khan[1]*, Massimo Mauro*[2]*, Pierangelo Migliorati*[1]*, Riccardo Leonardi*[1]

[1]Department of Information Engineering, University of Brescia, Brescia, Italy
[2]Yonder s.r.l, Trento, Italy, http://www.yonderlabs.com

ABSTRACT

The aim of this work is to explore the usefulness of face semantic segmentation for head pose estimation. We implement a multi-class face segmentation algorithm and we train a model for each considered pose. Given a new test image, the probabilities associated to face parts by the different models are used as the only information for estimating the head orientation. A simple algorithm is proposed to exploit such probabilites in order to predict the pose. The proposed scheme achieves competitive results when compared to most recent methods, according to mean absolute error and accuracy metrics. Moreover, we release and make publicly available a face segmentation dataset[1] consisting of 294 images belonging to 13 different poses, manually labeled into six semantic regions, which we used to train the segmentation models.

Index Terms— Face segmentation, feature extraction, classification, head pose estimation

1. INTRODUCTION

Head pose estimation aims at predicting the orientation of human head from a facial image. More specifically, the output of a head pose estimator consists of the yaw and the pitch angles and, optionally, the roll angle in 3D space. Head pose estimation has become an important topic in computer vision and pattern recognition[1]. The main reason is that head pose is a key information for many other applications, such as face and expression recognition, gaze estimation, augmented reality, computer graphics, 3D animation, etc.

The estimation of head orientation is problematic because of the sparsity of data and ambiguity of labels. In fact, available datasets are limited and a precise ground truth of head pose is difficult to obtain. The widely used Pointing'04 head pose database [2] - which we also adopt for comparison - is collected by asking humans sitting at the same position of the room to look at markers. This approach leads to approximate results: both the starting 3D head location and the direction of the head toward the markers are not perfectly aligned. As a result, the dataset contains a finite number of coarse poses, and considers only yaw and pitch angles.

[1]http://massimomauro.github.io/FASSEG-repository

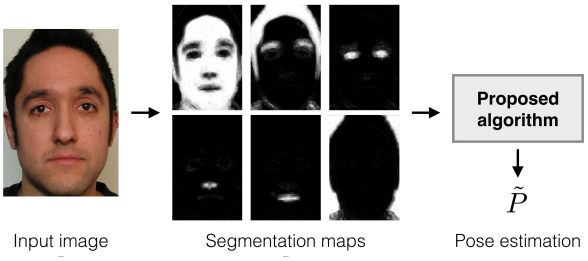

Fig. 1: Overall scheme of our approach. Probability maps are used as the only information for predicting the head pose.

In this paper we explore the usefulness of face semantic segmentation for estimating the head pose. Arguably, a strong interaction exists between facial parts and its corresponding pose. Relationship between face parts and pose recognition is confirmed by psychology literature as well, as facial features are informative for human visual system to recognize the face identity [3, 4]. Using a multi-class face segmentation algorithm [5], we train a model for each considered pose. The model parses the face into six parts and produces probability maps for skin, nose, eyes, mouth, hair, and background classes. Given a new image, the probabilities associated to face parts by the different models are used as the only information for estimating the head pose. A scheme of the proposed method is in Figure 1.

We evaluate our algorithm on the Pointing'04 database, and we show that the method achieves state-of-the-art performance according to both mean absolute error (MAE) and accuracy metrics. As an additional contribution, we release a dataset[1] which we used to train the segmentation models, that consists of 294 manually labeled ground-truth images belonging to 13 different poses. The dataset will be made publically available for research purposes.

The paper is arranged as follows: in Section 2 we present a short summary of related works, in Section 3 we introduce our pose estimation method, in Section 4 we explain the experimental setup and analyze the results, while Section 5 concludes the paper discussing some future directions.

2. RELATED WORK

A large amount of literature exists on the topic of pose estimation. A good historical survey and taxonomy of approaches can be found in [1]. We focus here only on methods most related to our work or that are evaluated on the Pointing'04 dataset as we do. Stiefelhagen et al. [6] extract horizontal and vertical image derivatives of the first order, then train a neural network to discriminate between poses. Bingpeng et al. [7] propose an algorithm for yaw estimation using the symmetry principle. Their method considers an effective combination of Gabor filters and covariance descriptors, and exploits the existing relationship between the head pose and the symmetry of the face image. A combination of Multi-scale Gaussian Derivatives (MGD) and Support Vector Machine is used by Jain et al. [8]. Face images are represented by MGDs, reducing the dimension with PCA. Due to low memory footprint and fast processing time, authors claim that their algorithm is suitable for hand-held devices. Huttunen et al. [9] summarizes outcome of a competition. The paper is using a collection of several machine learning methods for head pose estimation. The cropped face images are transformed into a dense histogram of oriented gradient (HOG) [10] features and used for classification. Authors of the paper claim better results as compare to state of the art. Kota et al. [11] propose a regression based method for head pose and car direction estimation. According to authors of the paper the proposed model is more flexible in the sense that the method does not rely on trial and error process for finding best spliting rules from already defined set of rules. Gaoli et al. [12] introduces feature weighting along with tree screening into the random forest trees in training stage. Features with high discriminative power are most likely to be chosen for building trees of the random forest. In this way, the diversity of the forest is not deteriorated and high pose estimation accuracy is obtained. The The proposed model is evaluated on head pose as well as survelance datasets.

Multivariate Label Distribution (MLD) [13] is a state of the art algorithm which uses HOG features and introduces the idea of soft labels: rather than explicit hard labels, every image is associated with a label distribution. To the best of our knowledge, this is the best performing method to date. While all previous methods mainly use appearance-based descriptors, we exploit face segmentations as the only feature in our work.

Vatahska et al. [14] algorithm initially detects facial features such as tip of the nose and both eyes. Based on the position of these features neural network estimates three rotation angles i.e., frontal, left and right profile images. Hatem et al. [15] method also uses facial features for head pose estimation. Haar like features are used initially for face localization, than the coordinates of eyes and mouth with respect to the nose are located. Authors of the paper perform their experimental work on a limited set of Pointing'04 poses (from $-60°$ to $+60°$). These methods face some problems possibly leading to a failure of the framework in some cases. For example, since feature extraction includes eye localization, a subject wearing glasses can make such localization problematic. In Pointing'04 dataset, 7 subjects out of 15 are wearing glasses. Similarly, if image resolution is poor, extraction of these features is almost impossible. In present work complicated face images such as images with glasses, moustaches and beared are also included in the experimental work.

Fig. 2: Examples of segmentation results on 7 poses (from $-90°$ to $+90°$ with a step size of $30°$). Labelled ground truth on the second row, PB algorithm output on the third row, and SPB algorithm output on the fourth row.

Huang et al. [16] approach the pose estimation problem through face segmentation. Authors of the paper argue that mid-level features such as pose, gender, and expression can be predicted easily from a well segmented facial image. Through experimental results, they prove that a relation exists between face segmentation and pose estimation. Experiments are performed on a small database of 100 images, considering only three poses (frontal, right, and left) and using a 3-class face segmentation (skin, hair and background). Their algorithm trains a pose regressor based on simple descriptors like the first moments of hair and skin about a centered vertical line. We exploit instead a 6-class segmentation algorithm and use probability maps as features. Moreover, we consider 13 total poses and make experiments on a larger image set taken from Pointing'04 database.

3. HEAD POSE ESTIMATION METHOD

We analyze separately the two main steps of our algorithm: multi-class face segmentation and head pose estimation.

3.1. Multi-class face segmentation

We use the algorithm already proposed in [5] for face segmentation. There, we analyzed two strategies - feature con-

Algorithm 1 Pose estimation algorithm

Input: a test image $\boldsymbol{I}_{test}$, n segmentation models S_1^n (n = number of poses).

1. **if** *PB segmentation*:

Extract patches from $\boldsymbol{I}_{test}$ using a step size 1

else if *SPB segmentation*:

Extract patches from the center of each super-pixel.

2. For each patch R:

2a. predict with S_1^n obtaining a set of class-probability pairs $(c, p)_1^n$

2b. evaluate $p_{max} = \max_{i=1\ldots n}(p_i)$

2c. **if** $c_{max} \in \{$mouth, eyes, nose$\}$:

assign the pixel to the corresponding pose P_R.

3. Count the number C_i of pixel assignments for each pose:

$$C_i = \sum_R (P_R == i)\,, i = 1 \ldots n$$

4. The predicted pose is the one for which C_i is maximum:

$$\tilde{P} = \arg\max_{i=1\ldots n} C_i$$

Output: Predicted pose $\tilde{P}$

Table 1: Pose estimation algorithm.

catenation and spatial prior - to exploit location feature. We adopt the former since it has shown to perform better. Such a method parses a frontal face into six classes: skin, eyes, mouth, nose, hair and background. We extend it by training models for all considered poses. For training, 14 images from each pose are manually labeled.

We analyse two possible solutions by using both pixel and super-pixels as base processing unit for segmentation. In the pixel-based approach (PB), squared patches are extracted and classified with a step size 1. In such a way, every pixel gets a class label and probabilities associated to each class. Some images segmented by PB method from Pointing'04 database are shown in figure 2 on the 3^{rd} row. In the superpixel-based approach (SBP), we use SEEDS [17] algorithm to over-segment an image into small superpixels. We extract and classify patches considering the center of each superpixel. As a result, all pixels belonging to the same superpixel share the same label and class probabilities. Some of the images segmented by the SPB method are shown in figure 2 on the 4^{th} row.

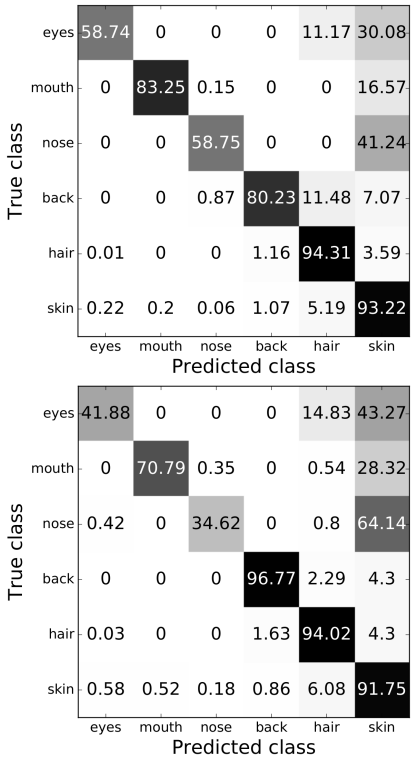

Fig. 3: Segmentation confusion matrices for pose 0° with the PB (upper) and SPB (lower) approaches.

3.2. Pose Estimation

The structure of our algorithm is outlined in Table 1. The working principle is as follows: given a test image, we run the segmentation models for all different poses, obtaining, for every pixel, a set of class predictions with associated probabilities. The set has the same size of the number of poses. We consider the maximum of such probabilities, and we "assign" the pixel to the respective pose. The maximum probability is noted as p_{max} in Table 1. At the end, the predicted pose is the one with the maximum number of assigned pixels. When using the SPB segmentation, the pipeline is the same, except that the pose is assigned to super-pixels. The underlying idea of our method is that when test data are most similar to training data, the segmentation model is more confident in its predictions, thus it classifies the image patch with a higher probability.

Additionally, we investigate which parts of the face are the most discriminative for understanding pose. To do this, we decide to assign pixels or super-pixels to a given pose only when the predicted semantic class belongs to a selected subset. After a detailed experimental analysis, we find that only eyes, nose, and mouth regions are really helpful. This fact is explained intuitively by observing that skin and hair patches from different poses can appear very similar.

Method	Reference	MAE (Deg.)	Accuracy (%)
Pixel-based approach	**–**	**3.75**	**77.4**
ML Distribution	MLD-wJ [13]	4.24	73.30
Super-pixel based approach	**–**	**5.69**	**67.3**
MG Derivative	Jain et al. [8]	6.9	64.51
kCovGa	Bingpeng et al. [7]	6.24	–
CovGa	Bingpeng et al. [7]	7.27	–
Neural Networks	Stifelhagen et al. [6]	9.5	52.0
Random forest regression	Li et al. [18]	9.6	–
Human performance	Gourier et al. [19]	11.4	40.7
High-order SVD	Tu et al. [20]	12.9	49.25

Table 2: Head pose estimation comparison of various methods on the Pointing'04 database. Performance data taken from [13].

4. EXPERIMENTS

The experimental section is divided into two parts. Subsection 4.1 is about the image setup used for experimental work and 4.2 explain the results and its discussion.

4.1. Experimental Setup

All images in the training and testing phases are the bounding boxes which are obtained through manual face localization method. We test our method on the Pointing04 dataset. Pointing04 contains images of 15 people who were asked to gaze at the markers marked on a sphere in a measurement room. The head orientation is determined by pan and tilt varying from -90° to $+90^\circ$ in both horizontal and vertical orientation. The phase shift for horizontal orientation is 15° while for vertical orientation is 30°. Thus, the "pitch dimention" is divided into 9 angles (-90°, -60°, -30°, 0°, 30°, 60°, 90°) while the "yaw dimension" is divided into 13 angles (-90°, -75°, -60°, -45°, -30°, -15°, 0°, $+15^\circ$, $+30^\circ$, $+45^\circ$, $+60^\circ$, $+75^\circ$, $+90^\circ$).

We consider all yaw poses from -90° to $+90^\circ$ with step size 15°, while we fix the pitch to 0°. In this setting, the total number of considered poses is 13. We do not perform experiments with vertical orientation poses but much better results are expected in that case as the phase difference between consecutive pose is 30° in vertical orientation poses.

To divide training and testing data, persons 1-7 are used for training the segmentation models while persons 8-15 are used for testing both segmentation and pose estimation performance. Original images are rescaled both in training and testing phases to a constant height $H = 512$, while width W is varied accordingly. The size of the extracted patches for segmentation is 16×16 for HSV and 64×64 for HOG features resulting a feature vector $f \in R^{1862}$.

4.2. Results and Discussion

Face Segmentation. In Figure 2 some examples of segmentation results on different poses of the same face are shown. In Figure 3 the confusion matrices are reported for both PB and SPB approaches. We use pixel-wise accuracy as the evaluation criterion for segmentation.

We note that the PB method outperforms SPB. The pixel-based method shows better results for all classes except background. In particular, accuracy for eyes, mouth, and nose classes dropped significantly with SPB. From confusion matrices it is clear that pixel-wise accuracy of the background class increases appreciably with SPB method. Although average accuracy of the pixel-based segmentation is greater, a substantial increase in speed - of an order of magnitude - is obtained by using super-pixels, since the number of patches to be classified by the model is greatly reduced.

Pose Estimation. The method is evaluated by two commonly used metrics. The first the MAE between the predicted pose and the "ground truth" pose. The second is the accuracy of pose predictions. MAE is a regression measure while pose estimation accuracy is a classification measure.

In Table 2 we list the performance of other recent methods, as reported in [13] for yaw angles. The comparison shows that we achieve state-of-the art performance with the pixel-based segmentation method for both MAE and accuracy metrics, while we rank third when switching to the superpixel-based approach.

In Figure 4 and 5 we show the confusion matrices for pose estimation when using PB and SPB segmentation respectively. Results are from good to perfect in most poses. The PB method achieves 100% accuracy for the largest angles (-90° and $+90^\circ$) that were declared as critique in [13]. We obtain instead poor performance only for 60° and -60°, probably because mouth, eyes, and nose classes are less discriminative than for other poses. Anyway, most incorrect predictions are adjacent to the ground truth angles.

An important observation derived from our results is that a correlation exists between segmentation and pose estima-

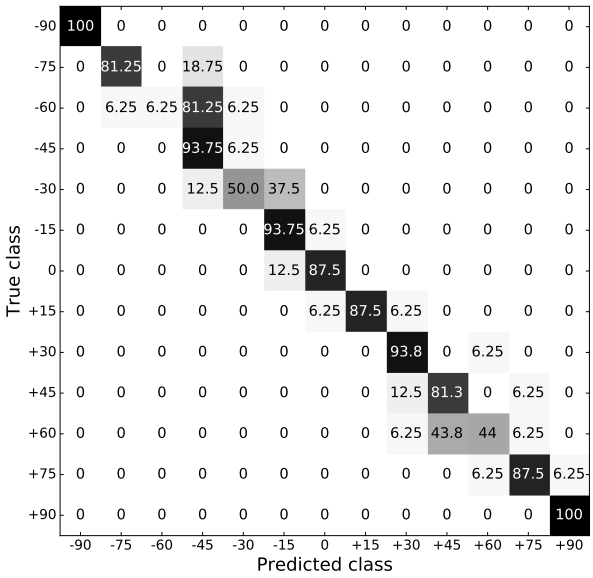

Fig. 4: Pose estimation confusion matrix for PB approach.

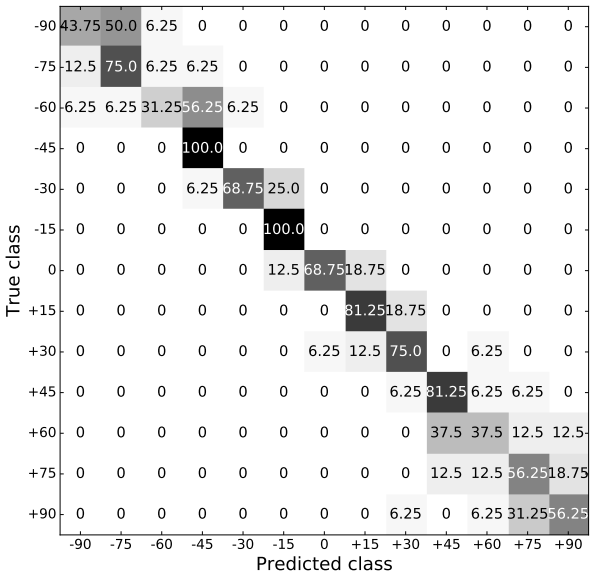

Fig. 5: Pose estimation confusion matrix for SPB approach.

tion accuracies. This effect can be observed from the performance gap among the PB and SPB methods in both tasks, and it was confirmed also in all our validation tests, where higher segmentation performance lead to more accurate pose predictions.

We observe that PB algorithm has much better results than SPB in pose estimation. Our proposed algorithm for pose estimation is mainly based on good segmentation of eyes, mouth, and nose. Due to low accuracies of SPB algorithm in these three classes, we have comparatively lower results for pose estimation.

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed an algorithm for pose estimation through face semantic segmentation. Using a multi-class face segmentation algorithm [5], we exploit the probabilities associated to face parts for estimating the head orientation. The proposed algorithm achieves state-of-the-art performance, proving that the coarse representation provided by the segmentation is highly informative for estimating the head pose. We also release a face segmentation dataset with labeled images belonging to 13 different poses.

We observed that a correlation exists between segmentation and pose estimation accuracies. A first direction for future work is thus to improve the segmentation models. Performance can be enhanced by integrating pixel probabilities into a conditional random field model to improve local labeling consistency, or by exploiting recent deep-learning approaches based on convolutional neural networks for image patch representation [21, 22].

A natural second direction for future work is to improve pose estimation. A possible idea is to use a more typical learning-based approach by training a pose estimation model using our probability maps as descriptors, eventually combined with other features.

Finally, a third interesting direction is to explore the use of the segmentation maps for other face analysis tasks, such as the estimation of gender, expression, age and ethnicity.

6. REFERENCES

[1] Erik Murphy-Chutorian and Mohan Manubhai Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607–626, 2009.

[2] Nicolas Gourier, Daniela Hall, and James L Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*. FGnet (IST–2000–26434) Cambridge, UK, 2004, pp. 1–9.

[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] Khalil Khan, Massimo Mauro, and Riccardo Leonardi, "Multi-class semantic segmentation of faces," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 827–831.

[6] Rainer Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data," in *Pointing04 ICPR Workshop of the Int. Conf. on Pattern Recognition*, 2004.

[7] Bingpeng Ma, Annan Li, Xiujuan Chai, and Shiguang Shan, "Covga: A novel descriptor based on symmetry of regions for head pose estimation," *Neurocomputing*, vol. 143, pp. 97–108, 2014.

[8] Varun Jain and James L Crowley, "Head pose estimation using multi-scale gaussian derivatives," in *Image Analysis*, pp. 319–328. Springer, 2013.

[9] Heikki Huttunen, Ke Chen, Abhishek Thakur, Artus Krohn-Grimberghe, Oguzhan Gencoglu, Xingyang Ni, Mohammed Al-Musawi, Lei Xu, and Hendrik Jacob Van Veen, "Computer vision for head pose estimation: Review of a competition," in *Scandinavian Conference on Image Analysis*. Springer, 2015, pp. 65–75.

[10] 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.

[11] Kota Hara and Rama Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *European Conference on Computer Vision*. Springer, 2014, pp. 552–567.

[12] Ronghang Zhu, Gaoli Sang, Ying Cai, Jian You, and Qijun Zhao, "Head pose estimation with improved random regression forests," in *Biometric Recognition*, pp. 457–465. Springer, 2013.

[13] Xin Geng and Yu Xia, "Head pose estimation based on multivariate label distribution," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1837–1842.

[14] Teodora Vatahska, Maren Bennewitz, and Sven Behnke, "Feature-based head pose estimation from images," in *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*. IEEE, 2007, pp. 330–335.

[15] Hiyam Hatem, Zou Beiji, Raed Majeed, Jumana Waleed, and Mohammed Lutf, "Head pose estimation based on detecting facial features," 2015.

[16] 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.

[17] Michael Van den Bergh, Xavier Boix, Gemma Roig, and Luc Van Gool, "Seeds: Superpixels extracted via energy-driven sampling," *International Journal of Computer Vision*, vol. 111, no. 3, pp. 298–314, 2015.

[18] Yali Li, Shengjin Wang, and Xiaoqing Ding, "Person-independent head pose estimation based on random forest regression," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1521–1524.

[19] Nicolas Gourier, Jérôme Maisonnasse, Daniela Hall, and James L Crowley, "Head pose estimation on low resolution images," in *Multimodal Technologies for Perception of Humans*, pp. 270–280. Springer, 2007.

[20] Jilin Tu, Yun Fu, Yuxiao Hu, and Thomas Huang, "Evaluation of head pose estimation for studio data," in *Multimodal Technologies for Perception of Humans*, pp. 281–290. Springer, 2007.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.