Computers, Materials & Continua
DOI:10.32604/cmc.2020.013590
Article

Tech Science Press

3D Head Pose Estimation through Facial Features and Deep Convolutional Neural Networks

Khalil Khan[1], Jehad Ali[2], Kashif Ahmad[3], Asma Gul[4], Ghulam Sarwar[5], Sahib Khan[6], Qui Thanh Hoai Ta[7], Tae-Sun Chung[8] and Muhammad Attique[9,*]

[1]Department of Computer Science and Software Engineering, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur-KPK, Pakistan
[2]Department of Computer Engineering and Department of AI Convergence Network, Ajou University, Suwon, 16499, South Korea
[3]Hamad Bin Khalifa University, Doha, Qatar
[4]Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan
[5]Department of Software Engineering, University of Azad Jammu and Kashmir, Pakistan
[6]Department of Electronics and Telecommunications, Politecnico di Torino, Torino, 10129, Italy
[7]Institute of Research and Development, Duy Tan University, Danang, 550000, Vietnam
[8]Department of Computer Engineering, Ajou University, Suwon, 16499, Korea
[9]Department of Software, Sejong University, Seoul, 05006, South Korea
*Corresponding Author: Muhammad Attique. Email: attique@sejong.ac.kr

Received: 12 August 2020; Accepted: 28 September 2020

Abstract: Face image analysis is one among several important cues in computer vision. Over the last five decades, methods for face analysis have received immense attention due to large scale applications in various face analysis tasks. Face parsing strongly benefits various human face image analysis tasks inducing face pose estimation. In this paper we propose a 3D head pose estimation framework developed through a prior end to end deep face parsing model. We have developed an end to end face parts segmentation framework through deep convolutional neural networks (DCNNs). For training a deep face parts parsing model, we label face images for seven different classes, including eyes, brows, nose, hair, mouth, skin, and back. We extract features from gray scale images by using DCNNs. We train a classifier using the extracted features. We use the probabilistic classification method to produce gray scale images in the form of probability maps for each dense semantic class. We use a next stage of DCNNs and extract features from grayscale images created as probability maps during the segmentation phase. We assess the performance of our newly proposed model on four standard head pose datasets, including Pointing'04, Annotated Facial Landmarks in the Wild (AFLW), Boston University (BU), and ICT-3DHP, obtaining superior results as compared to previous results.

Keywords: Face image analysis; face parsing; face pose estimation

1 Introduction

Face pose estimation, also known as head pose estimation is a challenging task in the field of computer vision. Head pose estimation plays an important role in many real-world applications, including gaze estimation [1], human computer interaction [2], and augmented reality [3]. However, face pose estimation is still a difficult task for various reasons such as variations in facial appearance, complex and unconstrained background, and different facial expressions. Head pose estimation is particularly confronted with problems in the uncontrolled and wild conditions. Some of the applications that rely heavily on an accurate head pose estimation system are human behavior analysis, safety during driving, surveillance applications, and targeted advertisements.

Head pose estimation is linked with gaze estimation; it is confirmed by the research conducted in the 19th century [4]. The relationship between head pose estimation and gaze prediction is also confirmed by later stage in research [5]. The research conducted in Langton et al. [5] suggests that gaze estimation comes from both eyes and head pose direction. Apart from eyes, this mutual relationship between various face parts and head pose is also confirmed by Huang et al. [6]. The work proposed by Huang et al. [6] suggests that relationship between face parts can be exploited for several mid-level vision tasks along with head pose estimation. Our work is also impressed by this idea of mutual relationship.

The face parsing method proposed in Khan et al. [7], segments a face image into face classes including, mouth, nose, hair, skin, back, and eyes. We also use face parts information by first developing a face segmentation framework. As can be seen these days, a shift in state-of-the-art methods from traditional machine learning algorithms towards recently introduced deep learning methods is prevalent. We also develop a DCCNs based face parsing method for seven different classes.

Our work is inspired from Huang et al.'s [6] idea. We argue that all face analysis tasks are related and can assist each other in specific applications. The performance of the face pose prediction can be improved if a prior efficiently parsed image having information about various face features is provided as input. The same fact is also confirmed by psychology literature, for example, [8,9]. In a nutshell, the performance of the face pose estimation can be improved if the information from various face features is extracted from a segmented image and given as input to the face pose estimation framework.

Face pose estimation is already being predicted through various face parts information in literature [10,11]. These methods involve landmarks localization before face pose estimation. However, the performance of the system in such cases depends on this method [12,13], which is itself another challenge. The landmarks localization algorithms are greatly affected in certain cases such as complicated facial expressions, changes in face rotation and lighting conditions, occlusions, and far field imagery conditions. All these factors make this method a rather challenging task, which ultimately drops the performance of the face pose estimation system; if it depends on it. Unlike the landmarks localization method, we introduce a face pose estimation method which does not need landmarks information but rather depends on various face parts information.

We propose a face segmentation method based on deep learning that segments a face image into seven different classes. For building a face parsing framework, we labeled 200 face images from each database through image editing software. The deep learning-based model extracts features through Convolutional Neural Networks (CNNs) and build a Soft-Max classifier. When a testing image is provided as input to the face parsing framework, it is segmented into seven face classes. We use a probabilistic classification method and create probability maps (PMAPS) along with segmentation results. We use five different face features and extract information through CNNs to build another Soft-Max classifier. To summarize, contributions of this paper are:

- We propose a face parsing method that segments a face image into seven different classes. The face parsing method is based on DCNNs.
- We develop a new face pose estimation algorithm. The proposed face pose estimation method is based on a prior face parsing method.
- We conduct experiments on state-of-the-art (SOA) datasets for face pose estimation and obtained much better results compared to previous results.

The structure of the paper is as follows: Section 2 presents related work for head pose estimation. Several datasets are reported by previous literature for head pose estimation. The datasets used in this paper are presented in Section 3. The face segmentation part is presented in Section 4, whereas the proposed head pose estimation algorithm is discussed in Section 5. The obtained results are discussed and then a comparison with SOA is shown in Section 6. We summarize the article with future directions in Section 7.

2 Related Work

2.1 Genetic Algorithm (GA)

Face parsing algorithms can be categorized into local and global based parsing methods. These methods are described in the following paragraphs.

Local methods: Local methods adopt a specific strategy of coarse-to-fine. Local based methods consider both local precision and global consistency. In local-based algorithms, separate models are trained for various face components, e.g., mouth, nose, eyes, etc. A method proposed by Luo et al. [14] trains a model that segments each face part individually. Zhou et al. [15] propose an interlinked CNNs based model after localizing face parts. The interlinked based method passes information bidirectionally, i.e., coarse and fine levels. The computational cost and memory consumption of the proposed method is large due to the bidirectional level information exchange. Another approach [16] combines the CNNs with Recurrent Neural Networks (RNNs) in two successive stages achieving SOA results on challenging.

Global methods: Global based methods treat different face parts information globally. Accuracy of these algorithms is less, as single face parts are not targeted. These methods estimate a label for each pixel over the entire face image. Some earlier works represent the spatial relationship between face parts through different models, for example, [17] and exemplar-based model [18]. The CNNs structure and loss function were processed by Liu et al. [16], which encodes the underlying layouts of the face image. This method integrates conditional random fields with CNNs, which the authors named Multi-Objective learning method. Jackson et al. [19] integrated CNNs with boundary cue to confine face regions. This method utilizes facial landmarks in the first step. Super-pixel information, Conditional Random Fields (CRFs), and CNNs are combined by Zhou et al. [20]. The method proposed in Zhou et al. [20] employs fully convolutional networks, therein obtaining better performance compared to SOA. The method proposed by Wei et al. [21] regulates receptive fields in a face parsing network. To achieve good performance on real time scenario, Saito et al. [22] propose another algorithm. The computational cost of this method is much lower than other methods.

2.2 Face Pose Estimation

Before describing the proposed face pose estimation model, we review related work on face pose estimation algorithms in this Section of the paper. A rich literature already exists for head pose estimation; however, in this Section of the article, we will try to provide maximum information about the recently introduced algorithms for face pose estimation.

Face pose can be classified into three categories, including yaw, roll, and pitch. The horizontal orientation is represented with yaw, vertical orientation with pitch, and the image plane by roll angle. We evaluate our proposed face pose estimation method with four large scale datasets, including Pointing’04 [23], AFLW [24], ICT-3DHPE [25], and Boston University (BU) [26] datasets. We classify the face pose estimation methods into three categories, including holistic approaches, geometric, and deep learning-based methods. These methods are described below.

2.2.1 Holistic Methods

In holistic approaches, the face image is considered as one-dimension vector, and certain features are extracted. Holistic methods assume a certain relationship between 2D face image properties and their

3D pose. A large number of face images are used for training purposes, and various statistical learning methods are exploited with different classification tools. The trained model then differentiates between various face poses. Some methods which used holistic approaches to address head pose estimation can be explored in [27–30].

Holistic methods show some advantages over its competitive methods. These approaches are comparatively simple and easy to implement. These methods are fit both for high- and low-resolution images. Moreover, no negative training data is needed in the training stage. Extension of the template models is also easy and can be extended any time without doing sufficient changes in the architecture.

Holistic methods also face some serious weaknesses. Like other methods, these algorithms also suffer from the limitation that the system must detect the face part. The system accuracy also degrades drastically with localization errors. Holistic methods become unreliable with variations in face appearance, changing in lighting conditions, and occlusions. A significant problem faced by these methods is pair wise similarity, which is the faulty assumption of the images of the same candidate in two different positions.

2.2.2 Geometric Methods

Geometric methods are also known as model-based methods. These methods require the localization of certain facial key points such as eyebrow, eyes, nose, the tip of the nose, lips, etc. A single feature vector is extracted from the located facial key points, and then the desired pose is predicted based on the relative position of the extracted key points. These methods are almost similar to how the human brain estimates the head pose of a face image.

The literature reports different face features in different combinations for head pose estimation. The intraocular distance and eyes are commonly used for head pose discrimination due to their easy detection [31]. In some cases, the mouth is also used, but mouth detection is comparatively difficult if the facial expressions are complicated. The tip of the nose and hair is another discriminative cue which is used for modeling of an efficient head pose estimation system.

Geometric methods are very fast, as very few features are needed for modeling. However, if the number of features is increased, the computational cost also raises; for example, Cootes et al. [32] uses a combination of fifteen different feature points around the mouth, eyes, and nose regions. Another main advantage noted for these methods is; the extracted points are robust to rotation and translation changes.

For different key points localization Active Shape Modeling (ASM) [33] is frequently used. However, ASM fails drastically in some critical conditions, for example, changing in lighting conditions, complicated facial expressions, and occlusions. If ASM does not perform well, the performance of the head pose estimation drops significantly. Some methods which use geometric based methods can be explored in references [34–36].

2.2.3 Deep Learning Methods

The performance of various visual recognition tasks has been greatly improved with deep learning architecture. Some very complex scenarios of computer vision tasks are improved with these deep learning methods. Major weaknesses of the conventional machine learning techniques are mitigated with this transition of deep learning architecture. The same is the case with face pose estimation.

Ruiz et al. [37] propose a deep learning-based method that does not depend on prior landmarks estimation. Some hybrid models are proposed in [38–40], which address head pose estimation along with other face image analysis tasks such as gender and race classification, face recognition, and detection, etc. Similarly, Hsu et al. [41] propose a method that combines regression-based function with deep learning modules. The lastly proposed method is named QuatNet (Quaternions) by the authors. A new deep learning-based method is proposed by Lee et al. [42], which is evaluated with Pointing'04. The method proposed in Lee et al. [42] is fast and comparatively robust to certain environmental factors; therefore, presents a better choice for the datasets collected in the wild conditions.

The performance of conventional machine learning methods is satisfactory on the datasets collected in constrained conditions. However, when these traditional machine learning methods are exposed to such unconstrained datasets, performance drops drastically. Unlike conventional machine learning methods, deep learning methods show much improvement with the challenging database. In a nutshell, some research work on face pose estimation already exists, but still, face pose estimation is an open challenge for researchers.

3 Databases

We evaluate our face pose estimation framework with five datasets, including Pointing'04, BU, AFLW, and ICT-3DHPE. This Section presents details about the databases used in our proposed work.

- **Pointing'04:** Images in the Pointing'04 are manually annotated. Although it is an old dataset, it is still used by researchers [43–45] due to the diversity in the images and its challenging nature. The database consists of fifteen sets of low-resolution images. Each individual set has a further two subsets having 93 face images for every subject in different orientations. The age of all participants in the dataset is between 20–40 years. Some participants in the datasets also include facial hair and some wearing glasses. The head pose of a subject is determined with the pan and tilt angle. Each participant is asked to look into 93 different markers marked on a wall in a measurement room. Each point represents a specific head pose. Due to manual labeling, the given face localization may not be accurate in some cases. The head orientation varies between -90^{o} to $+90^{\mathrm{o}}$ for yaw pose. The step size between two consecutive poses is 15. Similarly, for pitch, the top poses are represented with positive values and bottom poses with negative values. The difference between the two poses for pitch is 30^{o}.
- **AFLW:** These images are collected in the unconstrained condition with large variations in lighting conditions, facial expression, appearance, and some environmental factors. All images in AFLW are collected from the Internet. These images are collected in 9 different lighting conditions. The total number of face images is 13,000, whereas the number of participants is 5,749. The head pose is manually annotated in AFLW, where the yaw angle varies between $\pm 120^{\mathrm{o}}$ and pitch and roll in the range $\pm 90^{\mathrm{o}}$.
- **BU Data-set:** This dataset has two sequences; images exposed to changing lighting conditions and those collected in controlled and uniform lighting conditions. The database consists of both RGB and depth images. We use only RGB images in our experiments. We consider all three rotation angles, i.e., pitch, yaw, and roll. The number of participants in the database is only five. To collect ground truth data, magnetic trackers are attached to every subject's head.
- **ICT-3DHPE:** These images are collected through Kinect sensor. Both RGB and depth images are included; however, we use only RGB for our experiments. The total number of participants in the dataset is ten, with six male and four females. The ground truth data, in this case, is also accurate as again magnetic tracker is attached to each subject's head.

4 Proposed Face Parsing Framework

The face parsing module of the proposed framework is presented in this Section of the paper. We make this face parsing model for each dataset separately. Some of the datasets do not provide cropped face images; we apply a face detection method in the initial phase. As face detection is a mature research area, we use a face detection algorithm already proposed in Wei et al. [46] to each image. We re-scaled each face image to a fixed size of 227 × 227 after face detection. The proposed DCNNs based face segmentation model and its architecture is presented in Tab. 1. The Fig. 1 shows the proposed face parsing module.

Table 1: Information about each CNNs layer

Layer	Kernel Size	Stride	Feature Maps	Output Size
Input image	–	–	–	250 × 250
C-1	5 × 5	2	96	124 × 124
P-1	3 × 3	2	96	62 × 62
C-2	5 × 5	2	256	30 × 30
P-2	3 × 3	2	256	15 × 15
C-3	5 × 5	1	316	12 × 12
P-3	3 × 3	2	316	6 × 6
C-4	5 × 5	1	512	4 × 4
P-4	3 × 3	2	512	2 × 2

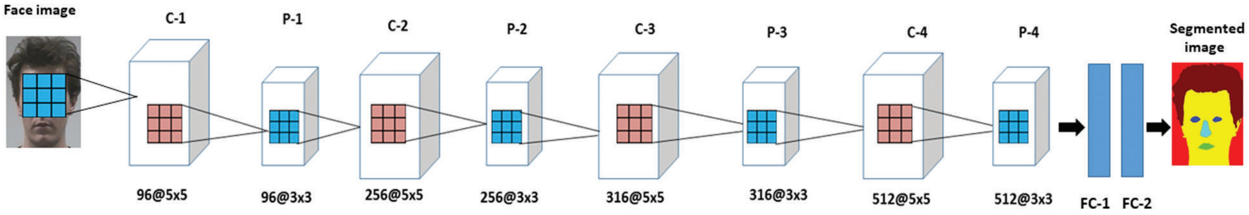

Figure 1: Proposed deep CNNs face parsing framework

Architecture: There are some parameters that greatly affect the performance of the DCNNs based model. For example, the size of the kernel used for CNNs and the pooling layer, the number of layers used for convolution, and the number of filters in every layer. In our face parsing module, we use four sets of convolutional layers (C1–C4), followed by maximum pooling layers (P1–P4) and, at the end, two fully connected layers. We fix the size of the kernel in convolutional as 5 × 5. We also fix the down sampling stride, as can be seen in Tab. 1. Details about the convolutional layer, feature map, and kernel size are shown in Tab. 1. Some other parameters of the proposed DCNNs are presented in Tab. 2.

Table 2: CNNs Parameters setting for training

CNNs parameters	Values
Epochs	40
Base learning rate	10^{-4}
Momentum	0.9
Batch size	150

We use a rectified linear unit for activation function. We embed the maximum pooling layer after each convolutional layer. Our proposed DCNNs face parsing model has main three parts, i.e., convolutional layers, maximum pooling layers, and two fully connected layers. We represent the convolutional layer kernel by $N * M * C$. Where height and width of the kernel is represented by N and M and the channel by C. The maximum pooling layer kernel is represented by $P * Q$, where P are representing height and

Q width of the kernel. The fully connected layer performs the task of classification. For optimization of the deep learning architecture and more details, Goodfellow et al. [47] can be explored further. The overview of the face parsing module is summarized in Tab. 1. We train a face parsing module for each database individually.

5 Proposed Features Based Face Pose Estimation Algorithm

Our proposed face pose estimation model is summarized in Algorithm 1. Initially, we develop a face segmentation model through DCNNs. The face parsing model outputs the most likely class for each pixel in a face image. We created PMAPS during the segmentation phase, which we further use for face pose estimation. We investigate different combinations of these PMAPS to know which face parts help face pose differentiation. We represent these PMAPS as: PMAPS$_{\text{nose}}$, PMAPS$_{\text{back}}$, PMAPS$_{\text{eyes}}$, PMAPS$_{\text{eyebrows}}$, PMAPS$_{\text{skin}}$, PMAPS$_{\text{mouth}}$, and PMAPS$_{\text{hair}}$. Fig. 1 shows some face images from Pointing'04 along with PMAPS for all five face classes, which we use in our face pose prediction model. PMAPS are grey scale images where higher intensity shows more probability of estimation for a face class and vice versa.

Algorithm 1: Proposed face pose estimation algorithm

```
Input: $G_{training} = \{(M_n, T_n)\}_{n=1}^{m}$ , $G_{testing}$
Where $G_{training}$ is the training data for DCNNs model, $G_{testing}$ is the testing data, $M$ is the training image
and $T(i,j) \in \{0,1,2,3,4,5,6\}$ is the data used as ground truth.
a. Face parsing part:
Step a.1: Training a face parsing DCNNs model through training data.
Step a.2: Producing face parts segmentation and probability maps for each dense semantic class
Step a.3: Using the DCNNs model to create PMAPS for the seven classes listed as;
PMAPS$_{nose}$, PMAPS$_{back}$, PMAPS$_{eyes}$, PMAPS$_{eyebrows}$, PMAPS$_{skin}$ , PMAPS$_{mouth}$, and PMAPS$_{hair}$
b. Face parsing part:
Step b.1: Extracting information from the PMAPS through DCNNs from five semantic classes including
nose, hair, mouth, eyes, and eyebrows.
Step b.2: Training a Soft-Max classifier A by creating feature vector such that;
$f = PMAPS_{eyes} + PMAPS_{mouth} + PMAPS_{eyebrow} + PMAPS_{nose}$
Output: estimated face pose.
```

For head pose estimation, we used extracted features from PMAPS images through CNNs. After extracting the features, we train another Soft-Max classifier for each dataset.

We manually labeled 200 face images from each dataset for seven face classes. The manually labeled images are used to build a face parsing model. For all images of every dataset, the PMAPS are generated. When a testing image is provided as input, the face parsing model creates the PMAPS for seven face classes. We conduct detailed experiments to investigate which face features can help face pose estimation.

After detailed analyses, we decided to use PMAPS for five classes, including the eyes, nose, mouth, eyebrow, and hair. After extracting features through CNNs we concatenate five feature vectors with each other to build a single unique feature vector. We trained another Soft-Max classifier using the feature vector. To validate our model more precisely, we use ten-fold cross-validation experiments. However, we exclude those 200 images which we previously used to build our classifier. The PMAPS of a subject from Pointing'04 can be seen in Fig. 2. From Fig. 2, it can be noticed that changes in PMAPS are occurring as we move from one pose to another.

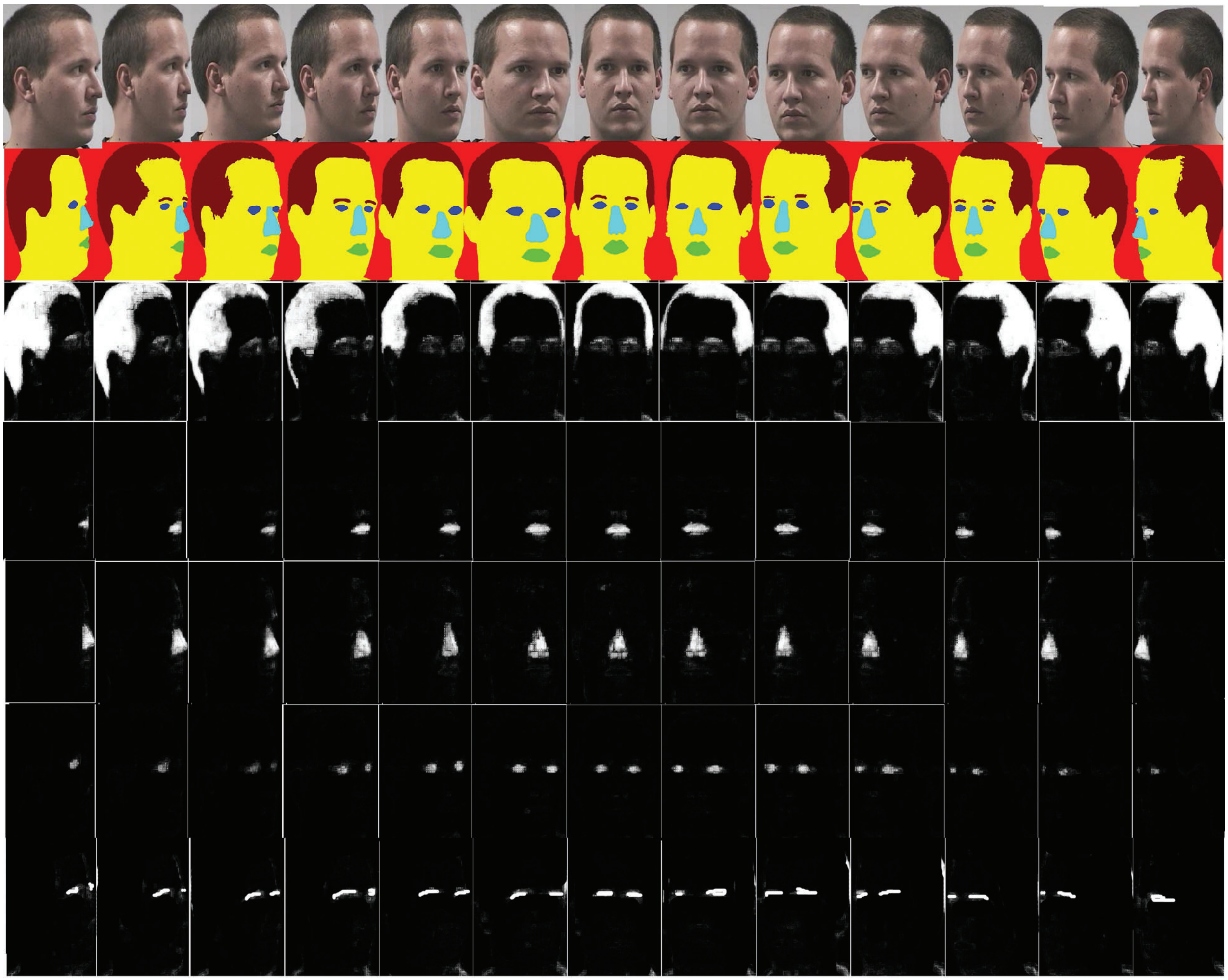

Figure 2: Segmentation results and probability maps from images of Pointing’04 dataset, where pose varies from −90° to +90° with 15° difference between two adjacent poses. The images are in the order where: row 1: original testing images, row 2: segmentation results with proposed face parsing module, row 3: probability maps for hair, row 4: probability maps for mouth, row 5: probability maps for nose, row 6: probability maps for eyes, and row 7: probability maps for eye brows

We investigate some interesting points during our experiments. We came to know that minor classes have more contribution towards face pose estimation system as compared to major classes (except hair). Hence, we use PMAPS of the four small classes (nose, eyes, brow, and mouth) and one major class hair. We ignore two major classes, skin and back. It can be seen from Fig. 2, that PMAPS for minor classes highly differs from one pose to another. For example, considering the fifth row (i.e., PMAPS for the nose), in frontal face images, the nose is more exposed to the camera, and as a result, the nose is almost in the middle of the face image. As pose of the image changes from the center to left (0^o to -90^o) and right profile (0^o to $+90^o$), the PMAPS of the nose also moves accordingly. We use this information as a feature and encode it in a unique feature vector.

The same difference can be noticed from Fig. 2 for four other PMAPS, including eyes, mouth, eyebrows, and hair. In extreme left and right profile images in some cases, the class information is entirely missing. This also clearly shows that our feature-based face pose estimation method highly depends on accurate face parts segmentation.

Hair class has a very complex geometry that varies from person to person. Our proposed face parsing part reports excellent labeling accuracy for hair. From the segmentation results in Fig. 2, it can be observed how efficiently our face parsing module segments a hair class. The borderline for hair is detected by our face segmentation part in a much better way.

For face pose estimation, we label 200 images from each dataset. We build a face parsing model for each database using 200 manually labeled images. We train a Soft-Max classifier for face parsing. When a test face image is given to the face parsing module, segmentation results are produced, as some of these are shown in Fig. 2. We use the probabilistic classification strategy and created PMAPS for each face class. We concatenate the five face classes' information by first extracting information through CNNs and create a unique feature vector, with which we again train a Soft-Max classifier. In experiments, we adapt 10-fold cross-validation experiments and report average results in the paper.

6 Results and Discussion

6.1 Experimental Setup

We use the Intel i7 CPU for our experiments. We use 16G RAM with NVIDIA 840 M graphical processing unit. We use Tensor flow and Keras as experimental tools. We train our model for 40 Epochs and batch size 150. We keep this setting for all face parsing models developed for all four datasets.

6.2 Face Segmentation Results

Some remarks for face parsing results are summarized in the following paragraphs:

Some qualitative results are shown in Fig. 2. The results show that face segmentation is better for the frontal poses as compared to profile. In contrast, labeling accuracy drops as the pose moves to right or left profile, which was expected as well, as minimal information are provided for training in case of extreme profile face images.

We also observe that as the pose moves to the right or left, labeling accuracy drops particularly for minor classes (nose, brow, eyes, and mouth). For extreme right or left profile face images, in some cases, the minor classes in some cases are completely missing. This can be seen from the images in Fig. 2. In such cases, the PMAPS produced are also unclear and the segmentation part provides minimal information.

The performance of the face parsing part also highly depends on image quality. For example, for low quality images (AFLW), comparatively poor results are reported by our proposed method. While for images from high quality datasets, such as Pointing'04 we obtained better results and also surpassed previously reported results.

We labeled all face images through image editing software. We used the manually labeled images to build a face parsing model. We used no automatic segmentation tool in all this process. This kind of labeling strictly depends on the subjective perception of a single human involved in manual labeling. To provide accurate face labels to face images with such type of labeling is very difficult. Differentiating boundary regions of face parts in such cases is very difficult; for example, explicitly drawing a boundary region between skin and nose is not accurate. And lastly, this labeling method is very tedious and time-consuming task. To label large number of images, sufficient time is needed.

6.3 Face Pose Estimation

To investigate which face features contribute significantly towards face pose estimation, we exploit the feature importance measure as reported in Pedregosa et al. [48]. It is a Random Forest implementation that calculates how certain features contribute to a specific task. Fig. 3 shows the importance of each face feature in face pose estimation. From Fig. 3, it can be seen that the maximum contribution to face pose estimation is provided by five classes, including eyes, hair, nose, mouth, and brows. Therefore, we use PMAPS of only five features as feature descriptors and discarded the remaining two classes.

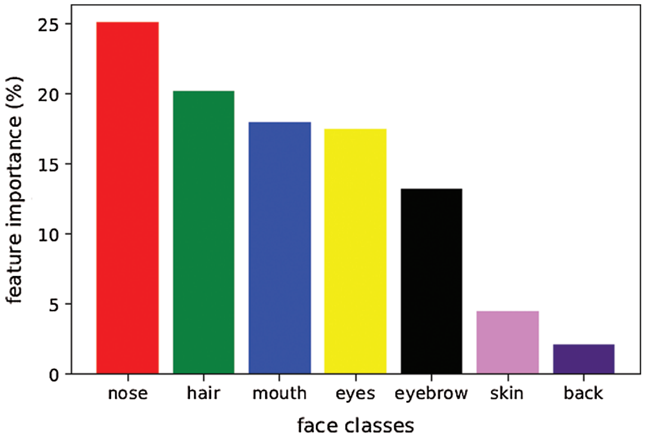

Figure 3: Feature importance of all the seven face classes including nose, hair, mouth, eyes, eyebrow, skin, and back mouth, and nose for face pose estimation

We evaluate our propose face pose estimation framework with two metrics, mean absolute error (MAE) and pose estimation accuracy (PEA). The MAE is a regression and PEA classification measures. MAE calculates the error between the estimated and actual pose, whereas PEA estimates how accurately a trained model estimates a pose.

The results obtained with the proposed face pose estimation framework on Poinitng'04 and its comparison with previous results are shown in Tab. 3. From Tab. 3, it can be seen that better results are obtained with the proposed model as compared to previous results. We explore all combinations of facial features and conclude to use just five classes. We obtain the best results as can be noticed from Tab. 3. Some of the previous techniques in Tab. 3 may use different validation protocols; for example, some methods use five-fold cross-validation during their experiments, whereas we perform our experiments with ten-fold cross-validation, which is more frequently used in the literature.

For AFLW, BU, and ICT-3DHPE results are reported only for MAE values. For a fair and exact comparison, we also report MAE values only. A complete summary of the results and then comparison with SOA is shown in Tab. 3. The results reveal that we have better results for two datasets BU and ICT-3DHPE. AFLW images are collected from the internet with a very complex background. Most of these images are in wild conditions with poor resolution as well. Our reported results for AFLW are less as compared to previous results. From the segmentation results, we note that face parsing results of the face segmentation module are weak for these complex images. One possible reason for poor results for AFLW, is comparatively poor performance of the face segmentation part.

The other two datasets, BU and ICT-3DHPE are also collected in real-world conditions. However, the quality of the images is comparatively better, and the background scenario is also not much complex. As a result, we obtain better results as compared to results reported in the literature.

Table 3: Head pose estimation results with proposed method and its comparison with SOA

Used database	Method used	MAE for yaw	Accuracy for yaw	MAE for pitch	Accuracy for pitch
Pointing'04	proposed feature-based method	**2.02°**	**89.2%**	**1.02°**	**96.5%**
	MSF [49]	3.7°	77.4%	–	
	MLD [50]	4.2°	73.3%	6.4°	89.2%
	CNNs [51]	5.2°	69.9%	5.4°	89.2%
	kCovGa [52]	6.3°	–	7.1°	–
	CovGA [52]	7.3°	–	8.7°	–
AFLW	QuatNet [41]	**4.3°**	**3.9°**	**2.6°**	**3.6°**
	proposed feature-based method	4.9°	4.2°	3.2°	4.1°
	HyperFace [38]	5.3°	6.2°	3.2°	4.9°
	Multi-Loss [37]	5.9°	6.2°	3.8°	5.3°
BU	proposed feature-based method	**2.4°**	**2.0°**	**2.0°**	**2.1°**
	OpenFace2.0 [53]	3.2°	2.4°	2.4°	2.6°
	OpenFace [54]	3.3°	2.8°	2.3°	2.8°
	Chehra [55]	4.6°	3.8°	2.8°	3.8°
	FLPD [56]	5.3°	4.9°	3.1°	4.4°
ICT-3DHPE	proposed feature-based method	**2.9°**	**2.2°**	**2.3°**	**3.0°**
	OpenFace2.0 [53]	3.5°	3.1°	3.1°	3.2°
	OpenFace [54]	3.6°	3.6°	3.6°	3.6°
	CLM [57]	4.2°	4.8°	4.5°	4.5°
	Reg. Forest [58]	9.4°	7.2°	7.5°	8.0°
	Chehra [55]	14.7°	13.9°	10.3°	13.0°

7 Conclusion

In the proposed work we introduced an end to end face parsing algorithm which tries to address a challenging problem of face pose estimation. We train a face parsing model through DCNNs by extracting useful information from different face parts. The face parsing model provides a class label for each pixel in a face image. We use a probabilistic classification technique and create PMAPS in the form of grey scale images for each face class. We perform a series of experiments to know which face feature helps in face pose estimation and conclude to use only five classes. We evaluate our proposed face pose estimation method with four datasets, including Pointing'04, BU, AFLW, and ICT-3DHPE obtaining better and competitive results.

Optimization of the face parsing part is one scenario to be addressed in the future. An important point to improve the performance of the face parsing system is by applying carefully well-managed engineering methods. For example, data augmentation [2] and foveated architecture [59] are some possible options to be adapted. Secondly, sufficient information is provided by the face segmentation part to address different visual recognition problems relating to the face. We provide a simple route towards some other complicated face analysis tasks, for example, gesture recognition, face beautification, and many more.

Funding Statement: This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01592) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant (2019R1F1A1058548) and Grant (2020R1G1A1013221).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Valenti, N. Sebe and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2011.

[2] K. Wang, R. Zhao and Q. Ji, "Human computer interaction with head pose, eye gaze and body gestures," in *Proc. of the 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, China, pp. 789, 2018.

[3] E. M. Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 300–311, 2010.

[4] W. H. Wollaston, "On the apparent direction of eyes in a portrait," *Philosophical Transactions of the Royal Society of London*, vol. 114, pp. 247–256, 1824.

[5] S. R. Langton, H. Honeyman and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception & Psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

[6] G. B. Huang, M. Narayana and E. Learned-Miller, "Towards unconstrained face recognition," in *Proc. of the IEEE Computer Vision and Pattern Recognition Workshops*, Anchorage, Alaska, pp. 1–8, 2008.

[7] K. Khan, M. Mauro and R. Leonardi, "Multi-class semantic segmentation of faces," in *Proc. of the IEEE Int. Conf. on Image Processing*, Quebec City, Canada, pp. 827–831, 2015.

[8] G. Davies, H. Ellis and J. Shepherd, *Perceiving and Remembering Faces*. Academic Press, 1981.

[9] P. Sinha, B. Balas, Y. Ostrovsky and R. 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.

[10] M. Dantone, J. Gall, G. Fanelli and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. of the Computer Vision and Pattern Recognition*, Rhode Island, pp. 2578–2585, 2012.

[11] R. Gross, I. Matthews and S. Baker, "Generic *vs.* person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080–1093, 2005.

[12] M. A. Haj, J. Gonzalez and L. S. Davis, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," in *Proc. of the Computer Vision and Pattern Recognition*, Rhode Island, pp. 2602–2609, 2012.

[13] E. M. Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2008.

[14] P. Luo, X. Wang and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. of the Computer Vision and Pattern Recognition*, Rhode Island, pp. 2480–2487, 2012.

[15] Y. Zhou, "Top-down sampling convolution network for face segmentation," in *Proc. of the IEEE Int. Conf. on Computer and Communications*, Chengdu, China, pp. 1893–1897, 2017.

[16] X. Liu, W. Liang, Y. Wang, S. Li and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. of the IEEE Int. Conf. of Image Processing*, USA, pp. 1289–1293, 2016.

[17] J. Warrell and S. J. Prince, "Label faces: Parsing facial features by multiclass labeling with an epitome prior," in *Proc. of the 16th in Proc. IEEE Int. Conf. of Image Processing*, Cairo, Egypt, pp. 2481–2484, 2009.

[18] K. Smith, S. O. Ba, J. M. Odobez and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1212–1229, 2008.

[19] A. S. Jackson, M. Valstar and G. Tzimiropoulos, "A CNN cascade for landmark guided semantic part segmentation," in *Proc. of the European Conf. on Computer Vision*, Cham: Springer, pp. 143–155, 2016.

[20] L. Zhou, Z. Liu and X. He, "Face parsing via a fully-convolutional continuous CRF neural network," arXiv preprint arXiv:1708.03736, 2017.

[21] Z. Wei, Y. Sun, J. Wang, H. Lai and S. Liu, "Learning adaptive receptive fields for deep image parsing network," in *Proc. of the Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2434–2442, 2017.

[22] S. Saito, T. Li and H. Li, "Real-time facial segmentation and performance capture from rgb input," in *Proc. of the European Conf. on Computer Vision*, Cham: Springer, pp. 244–261, 2016.

[23] N. Gourier, D. Hall and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Int. Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004.

[24] M. Koestinger, P. Wohlhart, P. M. Roth and H. Bischof, "Annotated facial landmarks in the wild: A large-scale real world database for facial landmark localization," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 2144–2151, 2011.

[25] T. Baltrusaitis, P. Robinson and L. P. Morency, "3D constrained local model for rigid and non-rigid facial tracking," in *Proc. of the Computer Vision and Pattern Recognition*, Rhode Island, pp. 2610–2617, 2012.

[26] M. L. Cascia, S. Sclaroff and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, 2000.

[27] V. Jain and J. L. Crowley, "Head pose estimation using multiscale gaussian derivatives," in *Proc. of the Scandinavian Conf. on Image Analysis*, Espoo: Springer, pp. 319–328, 2013.

[28] B. Ma, R. Huang and L. Qin, "Vod: A novel image representation for head yaw estimation," *Neurocomputing*, vol. 148, pp. 455–466, 2015.

[29] M. C. Burl and P. Perona, "Recognition of planar object classes," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 223–230, 1996.

[30] T. S. Jebara, "3D pose estimation and normalization for face recognition," Ph.D. thesis. McGill University, Canada, 1995.

[31] R. Stiefelhagen, J. Yang and A. Waibel, "A model-based gaze tracking system," *International Journal on Artificial Intelligence Tools*, vol. 6, no. 2, pp. 193–209, 1997.

[32] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[33] F. Fleuret and D. Geman, "Fast face detection with precise pose estimation," in *Proc. of the IEEE 16th Int. Conf. on Pattern Recognition*, Quebec City, Canada, pp. 235–238, 2002.

[34] A. Nikolaidis and I. Pitas, "Facial feature extraction and determination of pose," in *Noblesse Workshop on Non-Linear Model Based Image Analysis*, London: Springer, pp. 257–262, 1998.

[35] J. Wu and M. M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recognition*, vol. 41, no. 3, pp. 1138–1158, 2008.

[36] J. Sherrah and S. Gong, "Fusion of perceptual cues for robust tracking of head pose and position," *Pattern Recognition*, vol. 34, no. 8, pp. 1565–1572, 2001.

[37] N. Ruiz, E. Chong and J. M. Rehg, "Fine-grained head pose estimation without key points," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, Utah, USA, pp. 2074–2083, 2018.

[38] R. Ranjan, V. M. Patel and R. Chellappa, “HyperFace: A deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[39] T. Baltrusaitis, P. Robinson and L. P. Morency, “3D constrained local model for rigid and non-rigid facial tracking,” in *Proc. of the IEEE CVPR*, Providence, Rhode Island, pp. 2610–2617, 2012.

[40] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. P. Morency, “Open face 2.0: Facial behavior analysis toolkit,” in *Proc. of the 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, Xi’an, China, pp. 59–66, 2018.

[41] H. W. Hsu, T. Y. Wu, S. Wan, W. H. Wong and C. Y. Lee, “QuatNet: Quaternion-based head pose estimation with multiregression loss,” *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2019.

[42] S. Lee and T. Saitoh, “Head pose estimation using convolutional neural network,” in *Proc. of the IT Convergence and Security*, Singapore: Springer, pp. 164–171, 2018.

[43] Y. Liu, Z. Xie, X. Yuan, J. Chen and W. Song, “Multi-level structured hybrid forest for joint head detection and pose estimation,” *Neurocomputing*, vol. 266, pp. 206–215, 2017.

[44] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou *et al.,* “Compound rank-k projections for bilinear analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1502–1513, 2016.

[45] A. Schwarz, M. Haurilet, M. Martinez and R. Stiefelhagen, “Drive a head-a large-scale driver head pose dataset,” in *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, USA, pp. 1–10, 2017.

[46] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* “SSD: Single shot multibox detector,” in *Proc. of the European Conf. on Computer Vision*, Cham: Springer International Publishing, pp. 21–37, 2016.

[47] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.,* “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning*, vol. 12, pp. 2825–2830, 2011.

[49] K. Khan, M. Mauro, P. Migliorati and R. Leonardi, “Head pose estimation through multi-class face segmentation,” in *Proc. of the IEEE Int. Conf. on Multimedia and Expo*, Hong Kong, China, pp. 253–258, 2017.

[50] X. Geng and Y. Xia, “Head pose estimation based on multivariate label distribution,” in *Proc. of the Computer Vision and Pattern Recognition*, Columbus, Ohio, pp. 1837–1842, 2014.

[51] V. V. Jain and J. L. Crowley, “Head pose estimation using multi-scale Gaussian derivatives,” in *Proc. of the Scandinavian Conf. on Image Analysis*, Berlin, Heidelberg: Springer International Publishing, pp. 319–328, 2013.

[52] B. Ma, A. Li, X. Chai and S. Shan, “CovGa: A novel descriptor based on symmetry of regions for head pose estimation,” *Neurocomputing*, vol. 143, pp. 97–108, 2014.

[53] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L. P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision*, NY, USA, pp. 59–66, 2018.

[54] T. Baltrušaitis, P. Robinson and L. P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision*, NY, USA, pp. 1–10, 2016.

[55] A. Asthana, S. Zafeiriou, S. Cheng and M. Pantic, “Incremental face alignment in the wild,” in *Proc. of the Computer Vision and Pattern Recognition*, Columbus, Ohio, pp. 1859–1866, 2014.

[56] Y. Wu, C. Gou and Q. Ji, “Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, pp. 3471–3480, 2017.

[57] J. M. Saragih, S. Lucey and J. Cohn, “Deformable model fitting by regularized landmark mean-shift,” *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[58] G. Fanelli, T. Weise, J. Gall and L. V. Gool, “Real time head pose estimation from consumer depth cameras,” in *Joint Pattern Recognition Sym.*, Berlin, Heidelberg: Springer, pp. 101–110, 2011.

[59] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar *et al.,* “Large-scale video classification with convolutional neural networks,” in *Proc. of the Computer Vision and Pattern Recognition*, Columbus, Ohio, pp. 1725–1732, 2014.