A CNN Model for Head Pose Recognition using Wholes and Regions

Ardhendu Behera, Andrew G Gidney, Zachary Wharton, Daniel Robinson and Keiron Quinn
Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, L39 4QP, UK
{beheraa,23179601,23280425}@edgehill.ac.uk, {zachary.wharton,
kieron.quinn}@go.edgehill.ac.uk

***Abstract*— Head pose recognition and monitoring is key to many real-world applications, since it is a vital indicator for human attention and behavior. Currently, head pose is often computed by localizing landmarks on a targeted face and solving 2D to 3D correspondence problem with a mean head model. Recent research has shown that this is a brittle approach since it relies entirely on the accuracy of landmark detection, the extraneous head model and an ad-hoc alignment step. Recent work has also shown that the best-performing methods often combine multiple low-level image features with high-level contextual cues. In this paper, we present a novel end-to-end deep network, which is inspired by these ideas and explores regions within an image to capture topological changes due to changes in viewpoint. We adapt the existing state-of-the-art deep CNNs to use more than one region for accurate head pose recognition. Our regions consist of one or more consecutive cells and is adapted from the strategies used in computing HOG descriptor. Extensive experimental results on head pose recognition using four different large-scale datasets, demonstrate that the proposed approach outperforms many state-of-the-art deep CNN models. We also compare our pose recognition performance with the latest OpenFace 2.0 facial behavior analysis toolkit. In addition, we contribute head pose annotation to a large-scale dataset (VGGFace2).**

I. INTRODUCTION

Automatic detection and analysis of faces in images and videos is a fundamental problem in computer vision and has an important role in various applications: person identification, face verification, social robotics, activity recognition, modeling attentions, fitting 3D models and many more. Although significant advancement has been made in face detection, the reliable estimation of head pose and landmarks is still a challenging problem, particularly in unconstrained "in the wild" images. Uncertainty in head pose estimation is seemed to be a key factor for face recognition [10] and landmarks estimation [48], [32]. In extreme poses, face detection is arguably still a difficult problem to address.

Traditionally, head pose estimation is computed by locating 2D facial landmarks (also known as keypoints) in the target face and establishing the correspondence between landmarks and a head template by performing alignment [48], [11], [49], [36]. However, many real-world applications require an approximate estimation of the coarse head poses. In such cases, are the landmarks-based approaches still the best way forward? This paper address this question by exploring the latest deep Convolutional Neural Networks (CNNs) models.

Recently, there has been a significant progress in detecting and localizing facial landmarks using modern deep learning models [23], [32], [33], [48]. This has significantly influenced the way facial expression analysis is carried out. This is mainly due to their flexibility and robustness to extreme poses and occlusions, encouraging improvements in performance. These models are aimed to jointly predict head poses and facial landmarks. However, the primary goal of the head pose estimation is to improve the accuracy of the landmarks predictions. As a result, head pose estimation itself is not sufficiently accurate on its own.

In this work, we propose a novel holistic approach to estimate head poses from image intensities using CNNs. The proposed approach is inspired by the recent success in contextual action recognition using region-based CNN (R*CNN) [13] that delivers superior accuracy. There is no doubt about the significant improvement in facial landmarks detection accuracy due to the recent success of deep learning. However, there are many possibilities for introducing error for estimating head poses from these detected facial landmarks [34]. This is mainly due to the detection of sufficient numbers of facial landmarks and the quality of the head model. Its adaptation to each individual also depends on the model deformation, which is computationally expensive.

A. Related Work

Head pose estimation is to infer the orientation of person's head relative to the camera view. It is widely studied with very diverse approaches by exploring explicit 3D models [3], [15] or 2D view-based models [48], [49], [7].

Recently, there has been some progress in head pose estimation using CNNs. A study involving relatively shallow networks trained using a regression loss is presented in [31]. In OpenFace 2.0 [1], authors use simplified deep Convolutional Experts Constrained Local Model (CE-CLM) for facial landmarks detection. Head pose estimation is carried out using a 3D representation of facial landmarks and projects them to the image using orthographic camera projection. The tool could also recognize facial action unit and estimate eye-gazes. Hyperface [32] is a CNN that combines R-CNN [12] and AlexNet to predict four different sub-tasks (detect faces, determine gender, detect facial landmarks and estimate head pose) at once. KEPLER [23] uses Heatmap-CNN (H-CNN), which is a modified GoogleNet architecture to predict facial landmarks and pose, jointly. In order to improve the facial

Fig. 1: Examples of facial landmarks (5 and 68) detection using Tasks-Constrained Deep Convolutional Network (TCDCN) [48]. Traditionally, these landmarks are often used for face alignment to estimate head pose. Figures are taken from [48].

landmarks detection, it uses the coarse pose supervision. All-In-One CNN [33] is another deep model, which is aimed for simultaneous face detection and alignment, face recognition, smile detection, pose estimation, gender recognition and age estimation using a single CNN. The model uses a multi-task learning concept that regularizes the shared parameters.

Ruiz *et al.* [34] urge for landmarks-free head pose estimation using image intensities. They regress head pose Euler angles by applying multi-loss objective function to ResNet-50 and AlexNet. They demonstrate the success of their approach by improving head pose estimation using synthetically expanded 300W-LP dataset. This differs from our work since we focus on classification of head poses targeting the existing large-scale datasets used for face recognition.

Chang *et al.* [6] also advocate for landmarks-free head pose estimation. They use a simple CNN to regress 3D head poses, focusing on facial alignment using the predicted head pose. The facial alignment pipeline is targeted towards improving the face recognition accuracy. There is no direct evaluation of the head pose estimation and is different from our work since we evaluate and compare the head pose results.

De *et al.* [9] use VGG16 model to regress the head pose Euler angles by adding an additional `FC` and RNN layer. The pose estimation is improved by leveraging the time dimension captured by the RNN. Our approach focuses on classification and improving the pose estimation accuracy from a single image by modifying the state-of-the-art deep network architectures. Moreover, our model can be easily integrated into most of the deep CNN architectures.

B. Deep CNN for Estimating Head Poses

Recently, deep CNNs [37], [16], [42], [43], [18] have been widely used for solving visual recognition (e.g. faces, objects, scene, action, etc.) problems and have achieved significant improvements in comparison to traditional approaches involving hand-crafted features (e.g. HOG [8], SIFT [25]). However, head pose estimation using such networks is still in its infancy. Most of the existing head pose estimation approaches focus on depth images [38], [30], [27], [29] and if no depth information exists, facial landmarks are detected and pose is estimated [48], [23], [32], [33]. Recently, there has been some progress to explore such networks for estimating head poses [34], [38], [9]. However, these approaches use these networks to regress the head pose Euler angles (commonly known as Yaw, Pitch and Roll). This raises an important question: *Is very fine-grained (e.g. change in Euler angles within a few degrees) head pose necessary for many real-world applications?*

A wide range of applications such as human-robot interactions, social scene understanding from videos and human computer interactions, do not necessarily require fine-grained head pose estimation. In such scenarios, the user attention is often measured as head pose directed towards the target objects or scene. Therefore, the pose within a certain range (coarse) of Euler angle would be enough to measure attention. For example, head pose is often used for drivers alertness monitoring [20]. This alertness is often measured if the driver head pose is directed towards windscreen (frontal view), left mirror or right mirror (quarter to three quarter profile view). Moreover, we looked into the existing datasets (listed in Table I), which are widely used for facial expression analysis and recognition, landmarks detection and head pose estimation. Most of these datasets contain fewer head poses (i.e. binned into 5 to 8 poses), representing a range of angles. Furthermore, the commonly used state-of-the-art face databases such as VGGFace2 [5], Annotated Faces in the Wild (AFW) [49] and Annotated Facial Landmarks in the Wild (AFLW) [22] also use the coarse head poses, influencing our motivation. Therefore, this paper focuses on pose estimation as a classification problem to recognize five different head poses (classes): 1) frontal ($0°$), 2) half profile - left ($-45°$), 3) full profile - left ($-90°$), 4) half profile - right ($+45°$) and 5) full profile - right ($+90°$). A given image will be assigned the nearest pose label with a $\pm 15°$ error tolerance as in [49].

Depth (RGB-D) cameras (e.g. Kinect) and magnetic sensors are widely used in applications, where accurate head pose estimation is needed. The estimation could be very accurate but they suffer form the following limitations: 1) such devices use active sensing and therefore, they can be difficult to use in uncontrolled environments and outdoors since sunlight or ambient light can interfere with the active illumination. 2) There is a constraint on power, size, area and weight for real-world mobile applications (e.g. robotics and autonomous vehicles). In such scenarios, RGB cameras are more suitable since depth cameras draw more power, heavier and bigger. 3) The data rate for depth cameras are higher than the RGB ones, resulting increasing storage, data transfer and processing time.

One of the reasons for not exploring deep CNNs for head pose estimation could be due to the limited number of images and identities in a given dataset. For example, BIWI [11] dataset consists of 24 sequences (RGBD) from 20

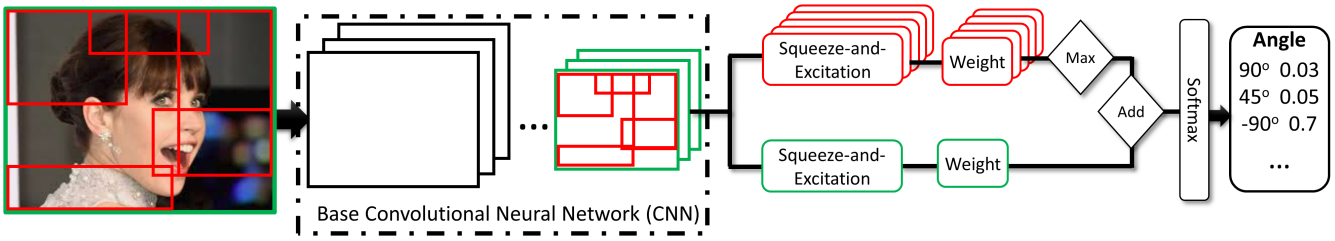

Fig. 2: Illustration of our proposed approach. Given an image, we select a set of candidate regions (bounding boxes). The image is passed through a base CNN (*e.g.*VGG [40], ResNet [16], Inception [43], etc.). For each pose p, the most informative region is selected (*max* operation) and its weight is added to the weight vector representing whole image. The weight vector is computed using specialised Squeeze-and-Excitation [17] layer, which responds to different inputs in a highly *pose-specific* manner. The softmax operation transforms weight vector into probabilities and forms the final prediction.

identities. Similarly, DriveAHead [38] contains 21 sequence (IR image and depth) from 20 subjects. Whereas, the deep CNNs for solving face recognition tasks are often evaluated on large datasets such as VGGFace2 [5], CelebFaces+ [41], CASIA-WebFace [46] and Labelled Faces in the Wild (LFW) [19], containing many thousands to a few millions images with over a thousand identities. One could argue for using *transfer learning* approach [47] for adapting such models to smaller datasets. These models' weights are learned from RGB intensities and adapting them to recognize depth and IR images is yet to be explored. More recently, it has also been shown that head pose has a significant influence on face recognition accuracy [5], [6].

The above discussion suggests that there is a need for head pose estimation using image intensities. In this paper, we aim to address this by proposing a novel approach, which is motivated by the success of the state-of-the-art deep CNNs for solving visual recognition problems. Our key contributions are:

- A novel end-to-end approach to combine regions with whole image to predict head pose directly from image intensities. We validate the significance of the regions leading to sizeable improvements in performance.
- Demonstrating the generalization capacity of our approach by integrating it with various state-of-the-art deep CNN models. The proposed model outperformed most of these models.
- To cope with the need for large training data with accurate annotations, we annotate the head pose of 63,016 images from 200 identities in the VGGFace2 [5] training set. Currently, 10,750 images in testing set are annotated. We are the first one to report the pose recognition performance on this dataset.
- We also create another large dataset for head pose recognition by combining various existing datasets listed in Table I.

II. PROPOSED APPROACH

This section describes the proposed CNN architecture, inspired by the recent advances in deep learning approaches to solve image recognition problem. The overview of the proposed architecture is shown in Fig. 2. We aim to estimate the head pose using deep CNNs directly from image intensities. We argue that it should be preferred to landmark-to-pose approaches [23], [32], [33] that require: 1) accurate 2D landmarks detection, 2) assumption of 3D human mean face model, 3) approximation of camera intrinsic parameters and 4) need to solve 2D-3D correspondence problem. There is a chance of introduction of error at each step which has been well-explored in [34]. We explain how combined image and region CNN features can be used for end-to-end training resulting in improved performance while using state-of-the-art CNN models. We also discuss about the creation of a large dataset, which is a collection of smaller datasets that are publicly available.

The proposed approach is inspired by R*CNN [13] and Histogram of Oriented Gradient (HOG) [8] for combining multiple cues in a given image. These cues are extracted using regions (similar to cells and blocks in HOG [8]). For a given region r in an image I, we define weight $\mathbf{W}(p; I, r)$ of pose p as:

$$\mathbf{W}(p; I) = \mathbf{W_I}^p \cdot \mathbf{F}(I) + \max_{r \in R(I)} \mathbf{W_r}^p \cdot \mathbf{F}(r; I) \quad (1)$$

where $\mathbf{F}(r; I)$ is a feature vector representing region r in I and $R(I)$ is the set of candidates region. Similarly, feature $\mathbf{F}(I)$ represents the whole image I. The weight vectors $\mathbf{W_I}^p$ and $\mathbf{W_r}^p$ are the weights of whole image I and region r for a given pose p, respectively. Given weights of each pose ($\mathbf{W}(p; I, r)$), we compute the probability of a given pose p in image I by using a softmax layer:

$$Prob(p; I) = \frac{e^{\mathbf{W}(p;I)}}{\sum_{p' \in P} e^{\mathbf{W}(p';I)}} \quad (2)$$

The feature $\mathbf{F}(.)$ and weights $\mathbf{W_I}^p$ and $\mathbf{W_r}^p$ are all trainable parameters and *learned jointly* for all poses $p \in P$ using a CNN, trained with gradient-based optimization of stochastic objective function.

A. Candidates Region Selection

Computer vision research has a long history of patch- or component/region-based approaches to visual recognition problem. This is mainly due to 1) different objects often

share some of their parts, 2) deal with partial occlusions and cluttered scenes, and 3) changes in the geometrical relation between parts can be modeled to be flexible to tolerate some deformations. Human head pose exhibits most of these characteristics.

Various hand-crafted features such as HOG [8] and SIFT [25] consider patches around keypoints or facial landmarks to extract features. Often the number of patches and their sizes are pre-defined. Not long ago, this patch-based approach has been adapted into the deep learning models such as R-CNN (Regions with CNN features) [12], which led to a significant impact on the simultaneous detection and localization problem involving objects and people. Our approach is inspired by this. In R-CNN, selective search is used to find 2K region proposals per image. Each region is passed through the same network to compute its objectness and is more suitable for the detection of distinct objects. Our aim is to recognize face orientation, which can be seen as the deformation of the same object (fine-grained task) and therefore, learning separate region-specific features is more suitable. To achieve this using a CNN, each region has to be modeled separately and will be difficult to fit to a large number (e.g. 2K in R-CNN) of regions. Thus, we adapted the strategies (cells and blocks) used in HOG [8] for our region proposals. We divide a given image into $C \times C$ cells. Our region consists of one or more consecutive cells, resulting regions of different aspect ratios and areas. For example, there are $|R(I)| = 35$ possible regions for $C = 3$. Moreover, the proposed region-specific computation layers are added towards the end layers of our network (Fig. 2) and therefore, most computational time is spent in the base CNN, which considers the whole image. One of the main advantages of the proposed region-based approach is that it can be added on the top of any existing CNN models. We compare our performance using different state-of-the-art CNN models as base CNN.

B. Weight Computation

An image is fed into a base CNN (Fig. 2). The output of the base CNN is reduced spatial resolution with increased number of channels/filters. This output is up-sampled and is fed into an adaptive max pooling layer, which also takes input as a list of regions of interest (ROIs) with spatial location (x, y) and size (width and height) information. This max pooling layer provides a fixed size (e.g. 7×7 for the FaceNet [37] as a base CNN) feature map for each ROI by manipulating the pooling size. These ROIs are computed from all possible combinations by considering one or more consecutive cells from a given $C \times C$ cells as described earlier. Therefore, our network does not require the cropped region or region annotations. Subsequently, the ROI-pooled features are passed through the corresponding Squeeze-and-Excitation [17] layer (red layers in Fig. 2) for computation of weights. The weights, W refers to convolutional kernel weights and the proposed network does not have any `FC` layer. We avoided the use of `FC` layer since it is the main cause for big memory footprint of CNNs. The weights are passed through a `Max` layer to identify ROIs with pose-specific contribution. Similarly, the output of the convolutional layer is also passed through a Squeeze-and-Excitation layer (green layer) for computing the respective weights. The green/red weights are indeed outputs of the respective Squeeze-and-Excitation [17] layer. Region-specific weights are added with the image-specific weights before the `Softmax` layer for head pose prediction. The image-specific weights (green) representing feature map of whole image as a output of the base network. This feature map is resized using bilinear interpolation, and then pooling is applied to provide fixed feature map (like in ROIs). This enables the use of `Add` layer before the `Softmax` layer. This network architecture is efficient, since the computationally intense convolutions are performed at an image-level within the base CNN and are eventually being reused by the ROI-specific operations.

The motivation for selecting Squeeze-and-Excitation [17] layer is to improve the representational power of our architecture by explicitly modeling the inter-dependencies between the channels of ROI-pooled features. In order to achieve this, one needs to perform feature re-calibration in which the network learns to use global information to selectively suppress less useful features and emphasize on the more informative ones. As a result, our model will be able to emphasize ROIs with pose-specific features. The feature re-calibration capability within the Squeeze-and-Excitation layer is computed as: 1) First, the RIO-pooled features are passed through a *squeeze* operation (*channel-wise scaling*), which aggregates the feature maps across ROIs spatial dimension (e.g. 7×7 for the FaceNet [37]) to produce a channel descriptor. This embeds the global distribution of channel-wise feature responses. 2) Second, this is followed by an *excitation* operation (*element-wise summation*) in which ROI-specific activations are learned for each channel by a self-gating mechanism based on channel dependence and governs the excitation of each channel. As a result, the Squeeze-and-Excitation layer becomes increasingly specialized and responds to different ROIs in a highly *pose-specific* manner.

C. Learning

The proposed approach is experimented with various state-of-the-art CNN as a base network. The main contribution involving features and weight vectors representing the respective region and whole image in (1) is added on top of the base network. All layers in the base network are initialized with pre-trained ImageNet's [35] weights except FaceNet [37], which is initialized with pre-trained CASIA-WebFace [46] and VGGFace2 [5] weights for the respective experiments on two different datasets (section III). The ImageNet consists of 1.2M natural images with 1K categories. The CASIA-WebFace contains 0.5M images of 10K subjects whereas VGGFace2 consists of 3.31M images from 9131 identities.

The network is trained in an end-to-end fashion with default image size of 224×224 and is randomly selected from 256×256 with data augmentation of height and width shift of up to 20%. The model is trained with a batch size of 32 using a Linux (Ubuntu) machine fitted with 24GB GPU (NVIDIA Quadro P6000) card. During training, we minimize

TABLE I: A **MultiLab** head pose dataset (24,334 images from 1288 identities) is created from the following existing publicly available datasets.

Dataset	Identities	# Images with 5 different poses				
		$-90°$	$-45°$	$0°$	$+45°$	$+90°$
IST-EURECOM [39]	50	100	100	1600	100	100
VGGFace2 Test [5]	500	1750	1930	3595	1648	1827
EURECOM Face [28]	52	104	0	728	0	104
Pointing [14]	15	630	420	690	420	630
KDEF [26]	70	243	245	245	245	245
BrazilFEI [44]	200	374	373	1135	378	373
Pain [45]	23	0	0	530	23	23
GUFD [4]	303	295	301	318	276	284
Iranian Women [2]	8	15	13	8	16	15
Radboud Faces [24]	67	377	377	377	377	377

the softmax probability $Prob(p; I)$ representing pose p is appeared in image I computed in (2). The loss over a batch of training images $B = \{I_i, y_i\}_{i=1}^{M}$ is given by

$$\text{loss}(B) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{p' \in P} \log\ Prob(p = y_{i,p'} \mid I_i) \qquad (3)$$

where p are the pose predictions, y are the actual labels, i denotes the training images and P represents a set of head poses. We use the Adam [21] optimizer with learning rate of 0.001 to minimize the objective function in (3) and train the model for 50 epochs. A region can be a single cell or combination of two or more adjacent cells. The number of regions $N = |R(I)|$ is a function of GPU memory limit. Therefore, we divide the whole image into 3×3 cells and consider all possible combinations, resulting in $N = 35$ regions.

The architecture of our network is based on the existing models. As a result, most computation time is spent in the base networks (Fig. 2), which consider the whole image.

III. Experiments

In order to validate our model for pose estimation using image intensities, we aim to consider a dataset consisting of a large number of identities. Such datasets (e.g. VGGFace2 [5], CelebFaces+ [41] and CASIA-WebFace [46]) exist but are mainly focused on identity recognition. We consider three different kinds of datasets for our experiments.

A. Datasets

We notice that there are a number of publicly available datasets for pose estimation and related research on face analysis. The list is presented in Table I. However, these datasets consist of a relatively small number of images and/or identities and are not large enough for developing deep models. Therefore, we combine these datasets to create a larger dataset (called MultiLab head pose). The MultiLab dataset consists of 24,334 images from 1288 identities. The dataset is randomly split into training set (80%) and testing set (20%) to evaluate our proposed architecture.

We also consider the VGGFace2 [5] dataset for our evaluation. However, it provides pose annotation of a subset of images ($10,750$ from 500 identities) within the test set. Using their head pose templates, we contribute to the pose annotation of 200 identities ($63,016$ images) within the training set. We evaluate our architecture using this dataset.

We also evaluate our approach on the Multi-Task Facial Landmark (MTFL) dataset [48] consisting five different head poses $(0, \pm 30, \pm 60)$. The dataset consists of $13,466$ faces in which $5,590$ are from LFW [19].

B. Results and Discussion

We have evaluated the proposed approach on the above-mentioned three challenging datasets. We use the metric as accuracy in percentage and is presented in Table II for MultiLab, VGGFace2 [5] and MTFL [48] dataset. The proposed region-specific computation layers are added on top of a base CNN model (Fig. 2). The weights for region-specific layers are initialized randomly whereas in base models, these are taken from pre-trained models. These pre-trained models are created using ImageNet dataset [35] except the FaceNet [37]. For FaceNet, the pre-trained model using CASIA-WebFace [46] is used to evaluate MultiLab dataset and VGGFace2 [5] for the rest of the two datasets. For the baseline, only the last layer (`Softmax`) of the base CNNs is adapted to the target number of classes (five in our case) and randomly initialized. For both baseline and proposed approach the networks are trained on the target datasets in an end-to-end fashion.

A common observation in performance of both the baselines and proposed approach is that the overall performance decreases as we move from MultiLab to MTFL dataset (Table II). This is mainly due to the increasing clutter in images. For example, most images in the MultiLab dataset are captured in a laboratory setup and thus, often exhibit clean background. In both MTFL [48] and VGGFace2 [5], we observed that both datasets contain images with mixed difficulty (e.g. occlusion, multiple faces, hand-over-faces, etc.) since they are collected from web. However, it is more often in MTFL than VGGFace2. Moreover, the size of VGGFace2 head pose train set is larger ($\sim$63K) than the MTFL (10K), resulting an impact on the performance because deep models learns better from large datasets.

Among baselines in all three datasets (Table II), the FaceNet[37] performed better than the rest (except VGG16). This is mainly due to the fact that the pre-trained model used for the initialization, is trained on large-scale face only dataset. Whereas, the rest are initialized with pre-trained models that are trained on ImageNet [35], which is targeted for general visual recognition challenge. This demonstrates the benefits of *transfer learning* [47] when base and target dataset containing images, which are of similar content.

The other observation is that among all models, Inception-V3 [43] and Inception ResNet-V2 [42] have the most improvement (i.e. gain in accuracy) when our proposed region-specific layers are added. These two models with our region-specific layers, are also best performer (except ResNet-V2 in VGGFace2 dataset) in all three datasets even better than the FaceNet [37]. The exception in the VGGFace2 dataset is not very far from the respective ResNet-50 and FaceNet.

TABLE II: Performance as recognition accuracy in percentage using MultiLab (Table I), VGGFace2 [5] and MTFL [48] head pose datasets. The proposed approach is outperformed all baselines except VGG16 [40] in MultiLab and MTFL datasets. For a given dataset and method, the best and the second best performances are in **bold** and *italic*, respectively.

Deep CNN Model	**MultiLab dataset** (Table I)			**VGGFace2 dataset** [5]			**MTFL dataset** [48]		
	Baseline	**Ours**	**Gain**	**Baseline**	**Ours**	**Gain**	**Baseline**	**Ours**	**Gain**
FaceNet [37]	*92.23*	94.87	2.64	**86.33**	*92.60*	6.27	*75.50*	79.71	4.21
ResNet-50 [16]	91.50	94.17	2.67	85.40	91.40	6.00	75.47	77.67	2.22
Inception ResNet-V2 [42]	88.82	*94.89*	**6.07**	82.43	90.50	**8.07**	69.12	**81.45**	**12.33**
Inception-V3 [43]	90.97	**95.01**	*4.04*	85.54	**93.35**	*7.81*	70.70	*80.85*	*10.15*
DenseNet-121 [18]	90.89	92.23	1.34	85.57	87.13	1.56	**77.11**	78.09	0.98
DenseNet-169 [18]	91.56	92.09	0.53	85.23	85.55	0.32	75.23	76.20	0.97
DenseNet-201 [18]	90.75	92.56	1.81	*85.85*	86.30	0.45	72.06	76.57	4.51
VGG16 [40]	**92.84**	92.11	-0.73	85.16	90.39	5.23	71.50	64.84	-6.66
NASNet mobile [50]	-	-	-	-	-	-	64.72	66.10	1.38

TABLE III: The recognition accuracy in percentage using the state-of-the-art OpenFace 2.0 [1] facial behavior analysis toolkit for head pose recognition using MultiLab, VGGFace2 [5] and MTFL[48] datasets.

Dataset	$-90°$	$-45°$	$0°$	$+45°$	$+90°$	**Overall**
MultiLab	12.31	43.82	99.25	35.86	16.87	54.14
VGGFace2 [5]	3.26	25.03	99.69	20.87	4.21	42.47
MTFL [48]	0.00	23.56	99.62	31.42	0.0	68.45

TABLE IV: The recognition accuracy in percentage using FaceNet [37] as baseline and the proposed region-based approach that uses FaceNet as a base network for head pose recognition using MultiLab and VGGFace2 [5] datasets.

Dataset		$-90°$	$-45°$	$0°$	$+45°$	$+90°$	**Overall**
MultiLab	baseline	92.84	85.57	95.82	86.36	91.94	91.76
	ours	98.37	85.44	96.84	93.95	93.92	94.42
VGGFace2	baseline	80.29	80.57	95.58	83.74	83.20	86.47
	ours	94.40	89.27	96.83	88.41	95.84	93.61

We believe the performance would further improve if we initialize these networks with pre-trained models, which are trained on a large-scale face only dataset (e.g. VGGFace2 [5], CASIA-WebFace [46], etc.) like in FaceNet instead of ImageNet [35].

The VGG16 [40] network did not do well in both MultiLab and MTFL [48] (Table II) datasets when our region-specific layers are added. However, it performed pretty well (improvement of 5.23%) in VGGFace2 [5] dataset. The only reason we could argue is that the VGGFace2 dataset is larger (training images $\sim$63K) than MTFL (10K) and MultiLab ($\sim$24K). The other unique characteristics of the VGG16 network is that it has two `FC` layers before the `Softmax` layer in the baseline model. These `FC` layers contribute a significant portion of network parameters, making it the largest ($\sim$138M) in comparison to other baseline models. These `FC` layers are removed when we add our region-specific layers since input to our region-specific layer is from a `Convolution` layer. This could be the reason why VGG16 baseline model performed well on smaller datasets in comparison to our region-specific architecture. Nevertheless, the performance of other networks integrated with our region-specific layers, performed significantly well in comparison to VGG16 baseline, even on a smaller dataset.

Fig. 3 presents some example images in which the proposed approach using FaceNet [37] as a base network is able to recognize the correct pose but the baseline fails. Similarly, example images in which both baseline and our approach fail to recognize correctly are presented in Fig. 4. Example images in which the baseline is able to recognize correctly whereas the proposed approach is failed to do so is shown in Fig. 5.

We have also evaluated the proposed approach using the AFLW dataset [22] consisting 25K annotated face images with coarse head poses. The performances of various CNN models are: 1) FaceNet [37] - 85.18% (baseline) and 93.19% (proposed), 2) Inception-V3 [43] - 83.55% (baseline) and 92.14% (proposed), and 3) ResNet-50 [16] - 84.54% (baseline) and 92.80% (proposed). This confirms the similar trend in performance improvement on other datasets in Table II.

We are the first to provide the quantitative evaluation of head pose recognition on VGGFace2 and MTFL dataset. To the best of our knowledge, there is not direct evaluation of head pose recognition using these datasets. However, head pose has been used to improve the detection of facial landmarks [48], as well as the influence of head pose in identity recognition performance [5].

We also evaluate the head pose recognition performance using the state-of-the-art OpenFace 2.0 toolkit [1], which is developed for facial behaviour analysis. The recognition performance is presented in Table III. All the test images from three datasets are processed using this toolkit. The detected poses are binned into five different poses: 1) $0°$, 2) $-45°$, 3) $-90°$, 4) $+45°$ and 5) $+90°$, allowing $\pm 15°$ error tolerance so that all three datasets will have the identical topology. From the Table III, it is evident that the toolkit performed nearly perfect (100%) for the frontal view ($0°$). However, the performance is significantly dropped for both half profile ($\pm 45°$) and full profile ($\pm 90°$) face images. The overall performance is significantly lower than the proposed approach, as well as the baseline models in Table II. It is observed that the performance for both half profile and full profile is lower in VGGFace2 and MTFL dataset

in comparison to MultiLab. These two datasets contain a significant amount of clutter images (e.g. occlusion, multiple faces, hand-over-faces, etc.). This suggests that the toolkit does not work well for recognizing profile views when clutter in images increases.

For a fair comparison, we also carried out performance analysis involving individual poses (above-mentioned five poses) using the proposed approach. The recognition performance of various poses using FaceNet [37] as a baseline and the proposed approach using FaceNet as a base network is presented in Table IV. The performance of the baseline, as well as the proposed approach is far better than the OpenFace 2.0 toolkit [1] (Table III).

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In this work, we present a novel approach that combines pose-specific ROI-pooled features by exploring multiple regions. The network learns to suppress regions, which are less useful and emphasize on more informative ones for a given head pose. Therefore, the proposed multi-region architecture can directly, accurately and robustly predict the head pose using image intensities. The proposed region-specific layers are added on top of the existing CNN models and therefore, most computational processing is in the base CNN, which process the whole images.

The proposed approach is evaluated on three challenging datasets. Our method significantly improves the prediction accuracy in comparison to the state-of-the-art CNN models, as well as the latest facial behavior analysis toolkit (OpenFace 2.0 [1]). We have introduced a new head pose dataset (MultiLab) by combining the existing datasets. We have also provided head pose annotations to 200 identities ($\sim$63K images) in VGGFace2 dataset [5]. We strongly believe that this will help advance the field of face and facial expression analysis and recognition.

B. Future Works

This includes further investigation into the reason(s) behind the incorrect classifications presented in Fig. 4 and 5. We will aim to further improve the current performance by extending the proposed model to address these incorrect classifications.

We will also like to extend our head pose architecture for face recognition problem since the recent research [5] shows that the identity recognition is often influenced by the head poses.

V. ACKNOWLEDGMENTS

The research is supported by the Edge Hill University's Research Investment Fund (RIF). The GPU used in this research is generously donated by the NVIDIA Corporation.

REFERENCES

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proc. IEEE Automatic Face & Gesture Recognition*, pages 59–66, 2018.

[2] A. Bastanfard, M. A. Nik, and M. M. Dehshibi. Iranian face database with age, pose and expression. In *IEEE Int'l Conf. on Machine Vision (ICMV)*, pages 50–55, 2007.

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on PAMI*, 25(9):1063–1074, 2003.

[4] A. M. Burton, D. White, and A. McNeill. The glasgow face matching test. *Behavior Research Methods*, 42(1):286–291, 2010.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE Automatic Face & Gesture Recognition*, pages 67–74, 2018.

[6] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Faceposenet: Making a case for landmark-free face alignment. In *Proc. IEEE ICCVW*, pages 1599–1608, 2017.

[7] T. F. Cootes, K. Walker, and C. J. Taylor. View-based active appearance models. In *Proc. IEEE Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, pages 886–893, 2005.

[9] J. De, Y. Gu, Xiaodong, D. M. Shalini, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proc. IEEE CVPR*, 2017.

[10] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Trans. on Intell. Sys. and Tech.*, 7(3):37, 2016.

[11] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *IJCV*, 101(3):437–458, 2013.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE CVPR*, pages 580–587, 2014.

[13] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r*cnn. In *ICCV*, pages 1080–1088, 2015.

[14] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *ICPR Int'l Workshop on Visual Observation of Deictic Gestures*, 2004.

[15] L. Gu and T. Kanade. 3d alignment of face in a single image. In *Proc. IEEE CVPR*, pages 1305–1312. IEEE, 2006.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE CVPR*, pages 770–778, 2016.

[17] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. of the IEEE CVPR*, 2018.

[18] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *IEEE CVPR*, volume 1, pages 4700–4708, 2017.

[19] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in Real-Life Images: detection, alignment, and recognition*, 2008.

[20] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt. Driver behavior analysis for safe driving: A survey. *IEEE Trans. on Intel. Transp. Syst.*, 16(6):3017–3032, Dec 2015.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] 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 *2011 IEEE Int'l Conf. on Computer Vision (ICCV) workshops*, pages 2144–2151, 2011.

[23] A. Kumar, A. Alavi, and R. Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *IEEE Automatic Face & Gesture Recognition*, pages 258–265, 2017.

[24] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[26] D. Lundqvist, A. Flykt, and A. Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998.

[27] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz. Robust model-based 3d head pose estimation. In *Proc. IEEE ICCV*, pages 3649–3657, 2015.

[28] R. Min, N. Kose, and J.-L. Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Trans. on Sys., Man, and Cybernetics: Systems*, 44(11):1534–1548, 2014.

[29] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Trans. on Multimedia*, 17(11):2094–2107, 2015.

[30] C. Papazov, T. K. Marks, and M. Jones. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface

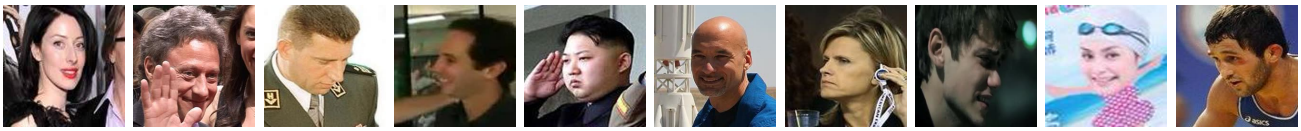

Fig. 3: Results of the proposed approach using FaceNet [37] as a base CNN. These images are from VGGFace2 dataset [5]. These examples are incorrectly recognized by the baseline (FaceNet) whereas the proposed approach (FaceNet + region-specific layers) did it with high probabilities (> 0.99). The recognized poses (G:ground-truth, P: proposed and B: baseline) from left-to-right are: (G:45, P:45, B:90); (G:45, P:45, B:$-$45); (G:90, P:90, B:0); (G:$-$90, P:$-$90, B:$-$45); (G:$-$45, P:$-$45, B:90); (G:$-$45, P:$-$45, B:45); (G:$-$90, P:$-$90, B:0); (G:90, P:90, B:$-$90); (G:0, P:0, B:$-$90); (G:$-$90, P:$-$90, B:0).

Fig. 4: Some examples of misclassification results of the proposed approach using FaceNet [37] as a base CNN. These images are from VGGFace2 dataset [5]. These are incorrectly recognized by both the FaceNet and the proposed approach (FaceNet + region-specific layers). The recognized poses (G:ground-truth, P: proposed and B: baseline) from left-to-right are: (G:0, P:45, B:45), (G:45, P:90, B:0); (G:$-$90, P:$-$45, B:$-$45); (G:90, P:45, B:45); (G:90, P:45, B:45); (G:$-$45, P:90, B:0); (G:$-$90, P:90, B:45); (G:$-$45, P:0, B:0); (G:90, P:45, B:0); (G:0, P:90, B:$-$45).

Fig. 5: Examples of incorrect results of the proposed approach using FaceNet [37] as a base CNN. These images are from VGGFace2 dataset [5]. The recognized poses (G:ground-truth, P: proposed and B: baseline) from left-to-right are: (G:0, P:$-$90, B:0); (G:0, P:90, B:0); (G:45, P:90, B:45); (G:$-$45, P:$-$90, B:$-$45); (G:$-$45, P:0, B:$-$45); (G:90, P:0, B:90); (G:$-$90, P:$-$45, B:$-$90).

patch features. In *Proc. IEEE CVPR*, pages 4722–4730, 2015.

[31] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017.

[32] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. on PAMI*, 2017.

[33] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *IEEE Automatic Face & Gesture Recognition*, pages 17–24, 2017.

[34] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *IEEE CVPR*, 2018.

[35] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[36] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*, pages 815–823, 2015.

[38] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen. Driveahead a large-scale driver head pose dataset. In *IEEE CVPR Workshops*, pages 1165–1174, 2017.

[39] A. Sepas-Moghaddam, V. Chiesa, P. L. Correia, F. Pereira, and J.-L. Dugelay. The ist-eurecom light field face database. In *5th IEEE Int'l Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE CVPR*, pages 1891–1898, 2014.

[42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, pages 2818–2826, 2016.

[44] C. E. Thomaz and G. A. Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902–913, 2010.

[45] University of Stirling. Pain expression, http://pics.stir.ac.uk/2D_face_sets.htm.

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[47] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in NIPS*, pages 3320–3328, 2014.

[48] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, pages 94–108, 2014.

[49] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE CVPR*, pages 2879–2886, 2012.

[50] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proc. IEEE CVPR*, pages 8697–8710, 2018.