An All-In-One Convolutional Neural Network for Face Analysis

Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo and Rama Chellappa
Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742
{rranjan1,swamiviv,carlos,rama}@umiacs.umd.edu

***Abstract*— We present a multi-purpose algorithm for simultaneous face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and face recognition using a single deep convolutional neural network (CNN). The proposed method employs a multi-task learning framework that regularizes the shared parameters of CNN and builds a synergy among different domains and tasks. Extensive experiments show that the network has a better understanding of face and achieves state-of-the-art result for most of these tasks.**

I. INTRODUCTION

Face analysis is a challenging and actively researched problem with applications to face recognition, emotion analysis, biometrics security, etc. Though the performance of few challenging face analysis tasks such as unconstrained face detection and face verification have greatly improved when CNN-based methods are used, other tasks such as face alignment, head-pose estimation, gender and smile recognition are still challenging due to lack of large publicly available training data. Furthermore, all these tasks have been approached as separate problems, which makes their integration into end-to-end systems inefficient. For example, a typical face recognition system needs to detect and align a face from the given image before determining the identity. This results in error accumulation across different modules. Even though the above mentioned tasks are correlated, existing methods do not leverage the synergy among them. It has been shown recently that jointly learning correlated tasks can boost the performance of individual tasks [59], [36], [5].

In this paper, we present a novel CNN model that simultaneously solves the tasks of face detection, landmark localization, pose estimation, gender recognition, smile detection, age estimation and face verification and recognition (see Fig. 1). We choose this set of tasks since they span a wide range of applications. We train a CNN jointly in a multi-task learning (MTL) framework (Caruana [3]), such that parameters from lower layers of CNN are shared among all the tasks. In this way, the lower layers learn general representation common to all the tasks, whereas upper layers are more specific to the given task, which reduces over-fitting in the shared layers. Thus, our model is able to learn robust features for distinct tasks. Employing multiple tasks enables the network to learn the correlations between data from different distributions in an effective way. This approach saves both time and memory in an end-to-end system, since it can simultaneously solve the tasks and requires the storage of a single CNN model instead of separate CNN for each task. To the best of our knowledge, this is the first work which simultaneously solves a diverse set of face analysis tasks using a single CNN in an end-to-end manner.

Fig. 1: The proposed method can simultaneously detect faces, predict their landmarks locations, pose angles, smile expression, gender, age as well as the identity from any unconstrained face image.

We initialize our network with the CNN model trained for face recognition task by Sankaranarayanan et al. [41]. We argue that a network pre-trained on face recognition task possesses fine-grained information of a face which can be used to train other face-related tasks efficiently. Task-specific sub-networks are branched out from different layers of this network depending on whether they rely on local or global information of the face. The complete network, when trained end-to-end, significantly improves the face recognition performance as well as other face analysis tasks. This paper makes the following contributions.

1) We propose a novel CNN architecture that simultaneously performs face detection, landmarks localization, pose estimation, gender recognition, smile detection, age estimation and face identification and verification.
2) We design a MTL framework for training, which regularizes the parameters of the network.
3) We achieve state-of-the-art performances on challenging unconstrained datasets for most of these tasks.

Ranjan et al. [36] recently proposed HyperFace that simultaneously performs the tasks of face detection, landmarks localization, pose estimation and gender recognition. The approach in this paper is different from HyperFace in the following aspects. Firstly, we additionally solve for the tasks of smile detection, age estimation and face recognition. Secondly, our MTL framework utilizes domain-based regularization by training on multiple datasets whereas HyperFace trains only on AFLW [24]. Finally, we initialize our network with weights from face recognition task [41] which

provides a more robust and domain specific initialization while HyperFace network is initialized using the weights from AlexNet [25].

This paper is organized as follows. Section II reviews closely related works. Section III describes the proposed algorithm in detail. Section IV provides the results of our method on challenging datasets for all the tasks. Finally, Section V concludes the paper with a brief discussion and future works.

II. Related Work

Multitask learning was first analyzed in detail by Caruana [3]. Since then, several approaches have used MTL for solving many problems in Computer Vision. One of the earlier methods for jointly learning face detection, landmarks localization and pose estimation was proposed in [59] which was later extended to [60]. It used a mixture of trees model with shared pool of parts, where a part represents a landmark location. Recently, several methods have incorporated the MTL framework with deep CNNs to train face-related tasks. Levi et al. [27] proposed a CNN for simultaneous age and gender estimation. HyperFace [36] trained a MTL network for face detection, landmarks localization, pose and gender estimation by fusing the intermediate layers of CNN for improved feature extraction. Ehrlich et al. [10] proposed a multi-task restricted Boltzmann machine to learn facial attributes, while Zhang et al. [56] improved landmarks localization by training it jointly with head-pose estimation and facial attribute inference. Although these methods perform MTL on small set of tasks, they do not allow training a large set of correlated tasks as proposed in this paper.

Significant research has been undertaken for improving individual face analysis tasks. Recent methods for face detection based on deep CNNs such as DP2MFD [35], Faceness [50], Hyperface [36], Faster-RCNN [20], etc., have significantly outperformed traditional approaches like TSM [59] and NDPFace [30].

Only a handful of methods have used deep CNNs for face alignment tasks [56], [26], [58], [36], due to lack of sufficient training data. Existing methods for landmark localization have focused mainly on near-frontal faces [2], [38], [22] where all the essential keypoints are visible. Recent methods such as PIFA [21], 3DDFA [58], HyperFace [36] and CCL [57] have explored face alignment over varying pose angles.

The task of pose estimation is to infer orientation of a person's head relative to the camera. Not much research has been carried out to solve this task for unconstrained images other than TSM [59], FaceDPL [60] and HyperFace [36].

The tasks of gender and smile classification from unconstrained images have been considered as a part of facial attribute inference. Recently, Liu et al. [31] released CelebA dataset containing about $200,000$ near-frontal images with 40 attributes including gender and smile, which accelerated the research in this field [46], [36], [55], [10]. Faces of the world [13] challenge dataset further advanced the research on these tasks for faces with varying scale, illumination and pose [28], [44], [53].

Age Estimation is the task of finding the real or apparent age of a person based on their face image. Few methods have already surpassed human error for the apparent age estimation challenge [12] using deep CNNs [40], [6].

Face Verification is the task of predicting whether a pair of faces belong to the same person. Recent methods such as DeepFace [43], Facenet [42], VGG-Face [34] have significantly improved the verification accuracy on the LFW [17] dataset by training deep CNN models on millions of annotated data. However, it is still a challenging problem for unconstrained faces with large variations in viewpoint and illumination (IJB-A [23] dataset). We address this issue by regularizing the CNN parameters using the MTL framework, with only half-a-million samples (CASIA [51]) for training.

III. Proposed Method

We propose a multi-purpose CNN which can simultaneously detect faces, extract key-points and pose angles, determine smile expression, age and gender from any unconstrained image of a face. Additionally, it assigns an identity descriptor to each face which can be used for face recognition and verification. The proposed algorithm is trained in a MTL framework which builds a synergy among different face related tasks improving the performance for each of them. In this section we discuss the advantages of MTL in the context of face analysis and provide the details of the network design, training and testing procedures.

A. Multi-task Learning

Typically, a face analysis task requires a cropped face region as the input. The deep CNN processes the face to obtain a representation and extract meaningful information related to the task. According to [52], lower layers of CNN learn features common to a general set of face analysis tasks whereas upper layers are more specific to individual tasks. Therefore, we share the parameters of lower layers of CNN among different tasks to produce a generic face representation which is subsequently processed by the task-specific layers to generate the required outputs (Fig. 2). Goodfellow et al. [14] interprets MTL as a regularization methodology for deep CNNs. The MTL approach used in our framework can be explained by following two types of regularization.

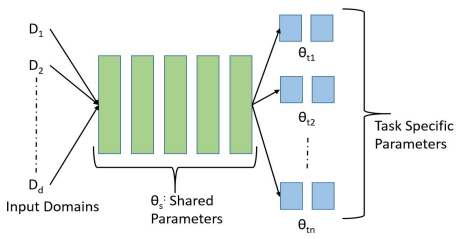


Fig. 2: A general multitask learning framework for deep CNN architecture. The lower layers are shared among all the tasks and input domains.

TABLE I: Datasets used for training

Dataset	Face Analysis Tasks	# training samples
CASIA [51]	Identification, Gender	490,356
MORPH [39]	Age, Gender	55,608
IMDB+WIKI [40]	Age, Gender	224,840
Adience [27]	Age	19,370
CelebA [31]	Smile, Gender	182,637
AFLW [24]	Detection, Pose, Fiducials	20,342
Total		**993,153**

1) Task-based Regularization: Let the cost function for a given task t_i with shared parameters θ_s and task-specific parameters θ_{t_i} be $J_i(\theta_s, \theta_{t_i}; D)$, where D is the input data. For isolated learning, the optimum network parameters $(\theta_s^*, \theta_{t_i}^*)$ can be computed using (1)

$$(\theta_s^*, \theta_{t_i}^*) = \underset{(\theta_s, \theta_{t_i})}{\arg\min}\, J_i(\theta_s, \theta_{t_i}; D) \tag{1}$$

For MTL, the optimal parameters for the task t_i can be obtained by minimizing the weighted sum of loss functions for each task, as shown in (2). The loss weight for task t_i is denoted by α_i.

$$(\theta_s^*, \theta_{t_i}^*) = \underset{(\theta_s, \theta_{t_i})}{\arg\min}\, \alpha_i J_i(\theta_s, \theta_{t_i}; D) + \sum_{j \neq i}^{n} \alpha_j J_j(\theta_s, \theta_{t_j}; D) \tag{2}$$

Since other tasks contribute only to the learning of shared parameters, they can be interpreted as a regularizer R_i on θ_s with respect to the given task t_i, as shown in (3). Thus, MTL shrinks the solution space of θ_s such that the learned parameter vector is in consensus with all the tasks, thus reducing over-fitting and enabling the optimization procedure to find a more robust solution.

$$(\theta_s^*, \theta_{t_i}^*) = \underset{(\theta_s, \theta_{t_i})}{\arg\min}\, J_i(\theta_s, \theta_{t_i}; D) + \lambda R_i(\theta_s; D) \tag{3}$$

2) Domain-based Regularization: For face analysis tasks, we do not have a large dataset with annotations for face bounding box, fiducial points, pose, gender, age, smile and identity information available. Hence, we adopt the approach of training multiple CNNs with respect to task-related datasets D_i, and share the parameters among them. In this way, the shared parameter θ_s adapts to the complete set of domains $(D_1, D_2, ...D_d)$ instead of fitting to a task-specific domain. Additionally, the total number of training samples increases to roughly one-million, which is advantageous for training deep CNNs. Table I lists the different datasets used for training our all-in-one CNN, along with their respective tasks and sample sizes.

B. Network Architecture

The all-in-one CNN architecture is shown in Fig. 3. We start with the pre-trained face identification network from Sankaranarayanan et al. [41]. The network consists of seven convolutional layers followed by three fully connected layers. We use it as a backbone network for training the face identification task and sharing the parameters from its first six convolution layers with other face-related tasks. Parametric Rectifier Linear units (PReLUs) are used as the activation function. We argue that a CNN pre-trained on face identification task provides a better initialization for a generic face analysis task, since the filters retain discriminative face information.

We divide the tasks into two groups: 1) subject-independent tasks which include face detection, keypoints localization and visibility, pose estimation and smile prediction, and 2) subject-dependent tasks which include age estimation, gender prediction and face recognition. Similar to HyperFace [36] we fuse the first, third and fifth convolutional layers for training the subject-independent tasks, as they rely more on local information available from the lower layers of the network. We attach two convolution layers and a pooling layer respectively to these layers, to obtain a consistent feature map size of 6×6. A dimensionality reduction layer is added to reduce the number of feature maps to 256. It is followed by a fully connected layer of dimension 2048, which forms a generic representation for subject-independent tasks. At this point, the specific tasks are branched into fully connected layers of dimension 512 each, which are followed by the output layers respectively.

The subject-dependent tasks of age estimation and gender classification are branched out from the sixth convolutional layer of the backbone network after performing the max pooling operation. The global features thus obtained are fed to a 3-layered fully connected network for each of these tasks. We keep the seventh convolutional layer unshared to adapt it specifically to the face recognition task. Task-specific loss functions are used to train the complete network end-to-end.

C. Training

The training CNN model contains five sub-networks with parameters shared among them. The tasks of face detection, key-points localization and visibility, and pose estimation are trained in a single sub-network, since all of them use a common dataset (AFLW [24]) for training. The remaining tasks of smile detection, gender recognition, age estimation and face recognition are trained using separate sub-networks. At test time, these sub-networks are fused together into a single all-in-one CNN (Fig. 3). All tasks are trained end-to-end simultaneously using Caffe [19]. Here, we discuss the loss functions and training dataset for each of them.

1) Face Detection, Key-points Localization and Pose Estimation: These tasks are trained in a similar manner as HyperFace [36], using AFLW [24] dataset. We randomly select 1000 images from the dataset for testing, and use the remaining images for training. We use the Selective Search [45] algorithm to generate region proposals for faces from an image. Regions with Intersection-Over-Union (IOU) overlap of more than 0.5 with the ground truth bounding-box are considered positive examples whereas regions with IOU<0.35 are chosen as negative examples for training the detection task using a softmax loss function. Facial landmarks, key-points visibility and pose estimation tasks

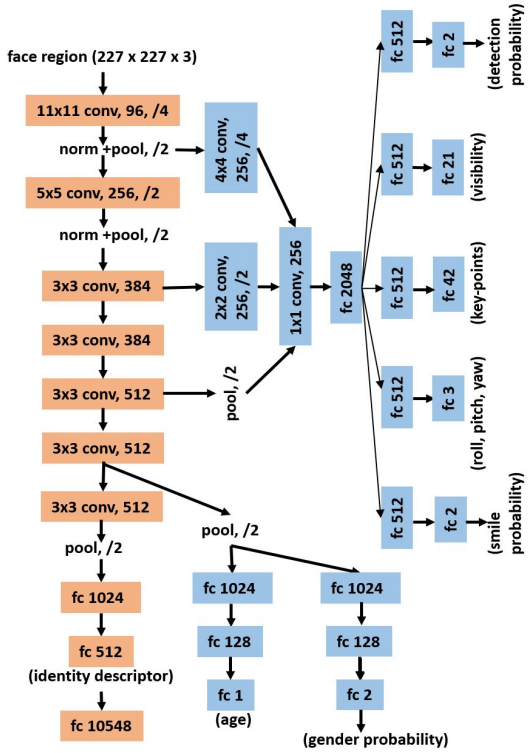


Fig. 3: CNN Architecture for the proposed method. Each layer is represented by filter kernel size, type of layer, number of feature maps and the filter stride. Orange represents the pre-trained network from Sankaranarayanan et al. [41], while blue represents added layers for MTL.

are treated as regression problems and trained with the Euclidean loss. Only regions with IOU>0.35 contribute to back-propagation during their training.

2) Gender Recognition: It is a binary classification problem similar to face detection. The datasets used for training gender are listed in Table I. The training images are first aligned using facial key-points which are either provided by the dataset or computed using HyperFace [36]. A cross-entropy loss L_G is used for training as shown in (4)

$$L_G = -(1-g)\cdot\log(1-p_g) - g\cdot\log(p_g), \qquad (4)$$

where $g = 0$ for male and 1 for female. p_g is the predicted probability that the input face is a female.

3) Smile Detection: The smile attribute is trained to make the network robust to expression variations for face recognition. We use CelebA [31] dataset for training. Similar to the gender classification task, the the images are aligned before passing them through the network. The loss function L_S is given by (5)

$$L_S = -(1-s)\cdot\log(1-p_s) - s\cdot\log(p_s), \qquad (5)$$

where $s = 1$ for a smiling face and 0 otherwise. p_s is the predicted probability that the input face is a smiling.

4) Age Estimation: We formulate the age estimation task as a regression problem in which the network learns to predict the age from a face image. We use IMDB+WIKI [40], Adience [27] and MORPH [39] datasets for training. It has been shown by Ranjan et. al. [37] that Gaussian loss works better than Euclidean loss for apparent age estimation when the standard deviation of age is given. However, the gradient of Gaussian loss is close to zero when the predicted age is far from the true age (Fig. 4), which slows the training process. Hence, we use a linear combination of these two loss functions weighted by λ as shown in (6)

$$L_A = (1-\lambda)\frac{1}{2}(y-a)^2 + \lambda\left(1 - exp(-\frac{(y-a)^2}{2\sigma^2})\right), \qquad (6)$$

where L_A is the age loss, y is the predicted age, a is the ground-truth age and σ is the standard deviation of the annotated age value. λ is initialized with 0 at the start of the training, and increased to 1 subsequently. In our implementation, we keep $\lambda = 0$ initially and switch it to 1 after $20k$ iterations. σ is fixed at 3 if not provided by the training set.

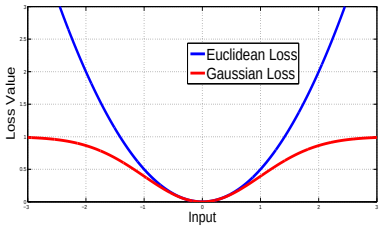


Fig. 4: Euclidean and Gaussian loss functions.

5) Face Recognition: We use $10,548$ subjects from CASIA [51] dataset to train the face identification task. The images are aligned using HyperFace [36] before passing them through the network. We deploy a multi-class cross-entropy loss function L_R for training as shown in (7)

$$L_R = \sum_{c=0}^{10547} -y_c\cdot\log(p_c), \qquad (7)$$

where $y_c = 1$ if the sample belongs to class c, otherwise 0. The predicted probability that a sample belongs to class c is given by p_c.

The final overall loss L is the weighted sum of individual loss functions, given in (8)

$$L = \sum_{t=1}^{t=8} \lambda_t L_t, \qquad (8)$$

where L_t is the loss and λ_t is the loss-weight corresponding to task t. The loss-weights are chosen empirically. We assign a higher weight to regression tasks as they tend to have lower loss magnitude than classification tasks.

D. Testing

We deploy a two-stage process during test time as shown in Fig. 5. In the first stage, we use the Selective Search [45] to generate region proposals from a test image, which are passed through our all-in-one network to obtain the detection scores, pose estimates, fiducial points and their visibility. We use Iterative Region Proposals and Landmarks-based NMS [36] to filter out non-faces and improve fiducials and pose estimates.

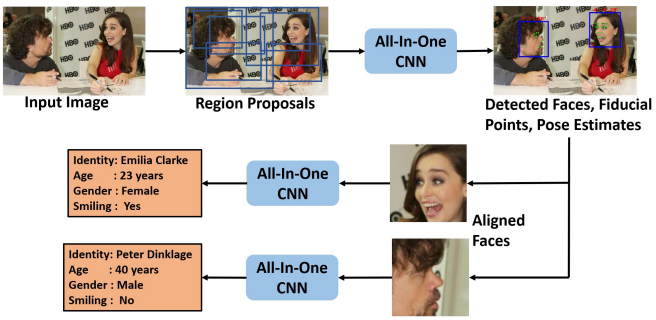


Fig. 5: The end-to-end pipeline for the proposed method during test time.

For the second stage, we use the obtained fiducial points to align each detected face to a canonical view using similarity transform. The aligned faces, along with their flipped versions are passed again through the network to get the smile, gender, age and identity information. We use the 512-dimensional feature from the penultimate fully connected layer of the identification network as the identity descriptor.

IV. Experiments

The proposed method is evaluated for all the tasks on which it was trained except the key-points visibility, due to the lack of a proper evaluation protocol. We select HyperFace [36] as a comparison baseline for the tasks of face detection, pose estimation, landmarks localization and gender recognition. For face recognition task, the method from Sankaranarayanan et. al. [41], which is used as the initialization, is used as the baseline.

A. Face Detection

We evaluate the face detection task on Annotated Face in-the-Wild (AFW) [59], PASCAL faces [48] and Face Detection Dataset and Benchmark (FDDB) [18] datasets. All these datasets contain faces with wide variations in appearance, scale, viewpoint and illumination. To evaluate on AFW and PASCAL datasets, we finetune the face detection branch of the network on FDDB. To evaluate on the FDDB dataset, we finetune according to the 10-fold cross validation experiments [18].

The precision-recall curves for AFW and PASCAL dataset, and Receiver Operating Characteristic (ROC) curve for FDDB dataset are shown in Fig. 6. It can be seen from the figures that our method achieves state-of-the-art performance on AFW and PASCAL dataset with mean average precision (mAP) of 98.5% and 95.01% respectively. On FDDB dataset, our method performs better than most of the reported algorithms. It gets lower recall than Faster-RCNN [20] and Zhang et al. [54], since small faces of FDDB fail to get captured in any of the region proposals. Other recently published methods compared in our detection evaluations include DP2MFD [35], Faceness [50], Headhunter [33], Joint Cascade [5], Structured Models [48], Cascade CNN [29], NDPFace [30], TSM [59], as well as three commercial systems Face++, Picasa and Face.com.

B. Landmarks Localization

We evaluate our performance on AFW [59] and AFLW [24] datasets as they contain large variations in viewpoints of faces. The landmarks location is computed as the mean of the predicted landmarks corresponding to region proposals having IOU>0.5 with the test face. For AFLW [24] evaluation, we follow the protocol given in [58]. We randomly create a subset of 450 samples from our test set such that the absolute yaw angles within $[0°, 30°]$, $[30°, 60°]$ and $[60°, 90°]$ are $1/3$ each. Table II compares the Normalized Mean Error (NME) for our method with recent face alignment method adapted to face profoling [58], for each of the yaw bins. Our method significantly outperforms the previous best HyperFace [36], reducing the error by more than 30%. A low standard deviation of 0.13 suggests that landmarks prediction is consistent as pose angles vary.

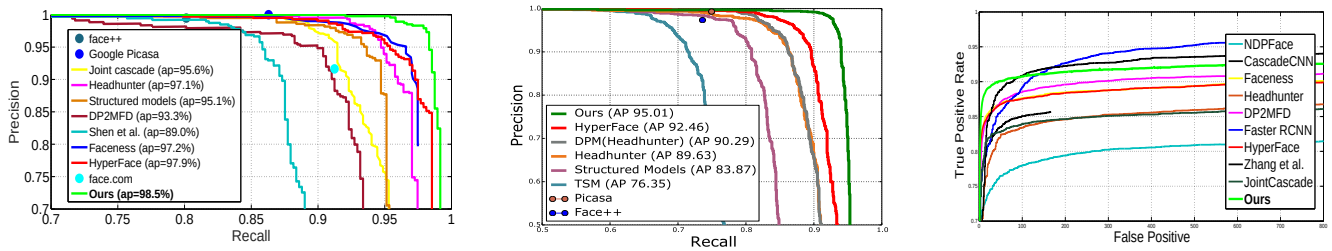
TABLE II: The NME(%) of face alignment results on AFLW test set.

	AFLW Dataset (21 pts)				
Method	[0, 30]	[30, 60]	[60, 90]	mean	std
RCPR [1]	5.43	6.58	11.53	7.85	3.24
ESR [2]	5.66	7.12	11.94	8.24	3.29
SDM [47]	4.75	5.55	9.34	6.55	2.45
3DDFA [58]	5.00	5.06	6.74	5.60	0.99
3DDFA+SDM	4.75	4.83	6.38	5.32	0.92
HyperFace [36]	3.93	4.14	4.71	4.26	0.41
Ours	**2.84**	**2.94**	**3.09**	**2.96**	**0.13**

For AFW [59] evaluation, we follow the protocol described in [57]. Fig. 7(a) shows comparisons with recently published methods such as CCL [57], HyperFace [36], LBF [38], SDM [47], ERT [22] and RCPR [1]. It is evident that the proposed algorithm performs better than existing methods on unconstrained and profile faces since it predicts landmarks with less than 5% NME on more than 95.5% of test faces. However, it lacks in pixel-accurate precise localization of key-points for easy faces, which can be inferred from the lower end of the curve. Most of these algorithms use cascade stage-wise regression to improve the localization, which is slower compared to a single forward pass of the network.

C. Pose Estimation

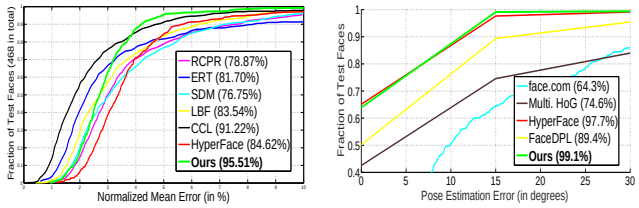
We evaluate our metod on AFW [59] dataset for the pose estimation task. According to the protocol defined in [59], we compute the absolute error only for the yaw angles. Since, the ground-truth yaw angles are provided in multiples of $15°$, we round-off our predicted yaw to the nearest $15°$ for



(a) (b) (c)

Fig. 6: Performance evaluation on (a) the AFW dataset, (b) the PASCAL faces dataset and (c) FDDB dataset. The numbers in the legend are the mean average precision for the corresponding datasets.

evaluation. Fig. 7(b) shows the comparison of our method with HyperFace [36], Face DPL [60], Multiview HoG [59] and face.com. It is clear that the proposed algorithm performs better than competing methods and is able to predict the yaw in the range of $\pm 15^\circ$ for more than 99% of the faces.



(a) (b)

Fig. 7: Performance evaluation on AFW dataset for (a) landmarks localization task, (b) pose estimation task. The numbers in the legend are the percentage of test faces with (a) NME less than 5%, (b) absolute yaw error less than or equal to 15°

D. Gender and Smile Recognition

We evaluate the smile and gender recognition performance on Large-scale CelebFaces Attributes (CelebA) [31] and ChaLearn Faces of the World [13] datasets. While CelebA [31] has wide variety of subjects, they mostly contain frontal faces. Faces of the World [13] has wide variations in scale and viewpoints of faces. We take the mean of the predicted scores obtained from region proposals having IOU>0.5 with the given face, as our final score for smile and gender attributes. Table III compares the gender and smile accuracy with recently published methods. On CelebA [31], we outperform all the methods for gender accuracy. Our smile accuracy is lower only to Walk and Learn [46] which uses other contextual information to improve the prediction. The gender and smile branches of the network were finetuned on the training set of Faces of the World [13] before its evaluation. We achieve state-of-the-art performance for both gender and smile classification on their validation set.

E. Age Estimation

We use Chalearn LAP2015 [12] apparent age estimation challenge dataset and FG-NET [15] Aging Database for evaluating our age estimation task. We fine-tune the age-task branch of the network on the training set of the challenge dataset, and show the results on the validation set. The error is computed according to the protocol described in [12]. For FG-Net [15], we follow the standard Leave-One-Out-Protocol (LOPO). Table IV lists the evaluation error for both these datasets. We surpass human error of 0.34 and perform comparable to state-of-the-art methods, obtaining an error of 0.293 on Chalearn LAP2015 [12] dataset. On FG-Net [15], we significantly outperform other methods, achieving an average error of 2 years.

TABLE III: Accuray (%) for Gender and Smile classification on CelebA [31] (left) and Faces of the World [13] (right)

Method	Gender	Smile
PANDA-1 [55]	97	92
MT-RBM [10]	90	88
LNets+ANet [31]	98	92
HyperFace [36]	97	-
Walk & Learn [46]	96	**98**
Ours	**99**	93

Method	Gender	Smile
MT-RBM [10]	71.7	80.8
CMP+ETH [44]	89.15	79.03
DeepBE [28]	90.44	88.43
SIAT_MMLAB [53]	91.66	89.34
Ours	**93.12**	**90.83**

TABLE IV: Age Estimation error on LAP2015(left) and FG-NET(right)

Method	Error
UMD [37]	0.359
Human	0.34
CascadeAge [6]	0.297
CVL_ETHZ	0.295
ICT-VIPL	**0.292**
Ours	0.293

Method	Error
Han2013 [15]	4.6
Chao2013 [4]	4.38
Hong2013 [16]	4.18
El Dib2010 [11]	3.17
CascadeAge [6]	3.49
Ours	**2.00**

F. Face Identification/Verification

We evaluate the tasks of face recognition and verification on the IARPA Janus Benchmark-A (IJB-A) [23] dataset. The dataset contains 500 subjects with a total of $25,813$ images including $5,399$ still images and $20,414$ video frames. It contains faces with extreme viewpoints, resolution and illumination which makes it more challenging than the commonly used LFW [17] dataset.

For IJB-A dataset, given a template containing multiple faces, we generate a common vector representation by media pooling the individual face descriptors, as explained in [41]. A naive way to measure the similarity of a template pair, is by taking cosine distance between their descriptors. A better way is to learn an embedding space where features

TABLE V: Face Identification and Verification Evaluation on IJB-A dataset

	IJB-A Verification (TAR@FAR)			IJB-A Identification			
Method	0.001	0.01	0.1	FPIR=0.01	FPIR=0.1	Rank=1	Rank=10
GOTS [23]	0.2(0.008)	0.41(0.014)	0.63(0.023)	0.047(0.02)	0.235(0.03)	0.443(0.02)	-
VGG-Face [34]	0.604(0.06)	0.805(0.03)	0.937(0.01)	0.46(0.07)	0.67(0.03)	0.913(0.01)	0.981(0.005)
Chen et al. [7]	-	0.838(0.042)	0.967(0.009)	-	-	0.903(0.012)	0.977(0.007)
Masi et al. [32]	0.725	0.886	-	-	-	0.906	0.977
NAN [49]	0.785(0.03)	0.897(0.01)	0.959(0.005)	-	-	-	-
Sankaranarayanan et al. [41] w/o TPE	0.766(0.02)	0.871(0.01)	0.952(0.005)	0.67(0.05)	0.82(0.013)	0.925(0.01)	0.978(0.005)
Sankaranarayanan et al. [41]	0.813(0.02)	0.90(0.01)	0.964(0.005)	0.753(0.03)	0.863(0.014)	0.932(0.01)	0.977(0.005)
Crosswhite et al. [9]	-	**0.939(0.013)**	-	0.774(0.049)	0.882(0.016)	0.928(0.01)	0.986(0.003)
Ours	0.787(0.04)	0.893(0.01)	0.968(0.006)	0.704(0.04)	0.836(0.014)	0.941(0.008)	0.988(0.003)
Ours + TPE [41]	**0.823(0.02)**	0.922(0.01)	**0.976(0.004)**	**0.792(0.02)**	**0.887(0.014)**	**0.947(0.008)**	**0.988(0.003)**

TABLE VI: Comparison of End-to-End face recognition systems on IJB-A

Face Detection	Face Alignment	Identity Descriptor	Metric Learning	Verif @FAR=0.01	Ident Rank=1
DP2MFD [35]	LDDR [26]	Chen et al. [8]	Joint Bayesian [8]	0.776(0.033)	0.834(0.017)
HyperFace		Sankaranarayanan et al [41]	cosine	0.871(0.01)	0.925(0.01)
HyperFace		Sankaranarayanan et al [41]	TPE [41]	0.90(0.01)	0.932(0.01)
HyperFace		**Ours**	cosine	0.889(0.01)	0.939(0.01)
Ours			cosine	0.893(0.01)	0.941(0.008)
Ours			TPE [41]	**0.922(0.01)**	**0.947(0.008)**

corresponding to similar pairs are close to each other while dissimilar pairs are far away. We train a Triplet Probabilistic Embedding (TPE) [41] using the training splits provided by the dataset. Table V compares with recently published methods on IJB-A. We achieve state-of-the-results for the face identification task. Although we perform comparable to template-adaptaion learning (Crosswhite et al. [9]) on verification task, we achieve a significantly faster query time ($0.1s$ after face detection per image pair). We get a consistent improvement of 2% to 3% over the baseline network [41] for all metrics.

We also compare our performance with end-to-end face recognition methods in Table VI. Our method outperforms existing end-to-end systems which shows that training all the tasks in the pipeline simultaneously, reduces the error. We see a two-fold improvement, i.e., about 80% performance gain is a result of improved identity descriptor and 20% gain is due to improved face alignment.

G. Runtime

We implemented our all-in-one network on a machine with 8 CPU cores and GTX TITAN-X GPU. It takes an average of 3.5s to process an image. The major bottleneck for speed is the process of generating region proposals and passing each of them through the CNN. The second stage of our method takes merely 0.1s of computation time.

V. CONCLUSIONS AND FUTURE WORKS

In this paper we presented a multi-task CNN-based method for simultaneous face detection, face alignment, pose estimation, gender and smile classification, age estimation and face verification and recognition. Extensive experiments on available unconstrained datasets show that we achieve state-of-the-art results for majority of these tasks. Our method performs significantly better than HyperFace, even though both of them use the MTL framework. This work demonstrates that subject-independent tasks benefit from domain-based regularization and network initialization from face recognition task. Also, the improvement in face verification and recognition performance compared to [41] clearly suggests that MTL helps in learning robust feature descriptors. In future, we plan to extend this method for other face-related tasks and make the algorithm real time. Several qualitative results of our method are shown in Figure 8.

VI. ACKNOWLEDGMENTS

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

REFERENCES

[1] X. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[3] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[4] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[5] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *European Conference on Computer Vision*, volume 8694, pages 109–122. 2014.

Fig. 8: Qualitative results of our method. The blue boxes denote detected faces. The green dots provide the landmark locations. Pose estimates for each face are shown on top of the boxes in the order of roll, pitch and yaw. The predicted identity, age, gender and smile attributes are shown below the face-boxes. Although the algorithm generates these attributes for all the faces, we show them only for subjects that are present in the IJB-A dataset for better image clarity.

[6] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa. A cascaded convolutional neural network for age estimation of unconstrained faces.

[7] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[8] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 118–126, 2015.

[9] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[10] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.

[11] M. Y. El Dib and M. El-Saban. Human age estimation using enhanced bio-inspired features (ebif). In *2010 IEEE International Conference on Image Processing*, pages 1589–1592. IEEE, 2010.

[12] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[13] S. Escalera, M. Torres, B. Martinez, X. Baró, H. J. Escalante, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of IEEE conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[14] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[15] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *2013 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2013.

[16] L. Hong, D. Wen, C. Fang, and X. Ding. A new biologically inspired active appearance model for face age estimation by using local ordinal ranking. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 327–330. ACM, 2013.

[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, Oct. 2007.

[18] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[20] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. *arXiv preprint arXiv:1606.03473*, 2016.

[21] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[22] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014.

[23] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[24] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, Nov 2011.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[26] A. Kumar, R. Ranjan, V. M. Patel, and R. Chellappa. Face alignment by local deep descriptor regression. *CoRR*, abs/1601.07950, 2016.

[27] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, June 2015.

[28] C. Li, Q. Kang, G. Ge, Q. Song, H. Lu, and J. Cheng. Deepbe: Learning deep binary encoding for multi-label classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–46, 2016.

[29] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, June 2015.

[30] S. Liao, A. Jain, and S. Li. A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, Dec. 2015.

[32] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? *arXiv preprint arXiv:1603.07057*, 2016.

[33] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, volume 8692, pages 720–735. 2014.

[34] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[35] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In *International Conference on Biometrics Theory, Applications and Systems*, 2015.

[36] 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. *CoRR*, abs/1603.01249, 2016.

[37] R. Ranjan, S. Zhou, J. C. Chen, A. Kumar, A. Alavi, V. M. Patel, and R. Chellappa. Unconstrained age estimation with deep convolutional neural networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, ICCVW '15, pages 351–359, Washington, DC, USA, 2015. IEEE Computer Society.

[38] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[39] K. Ricanek Jr. and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, FGR '06, pages 341–345, Washington, DC, USA, 2006. IEEE Computer Society.

[40] R. Rothe, R. Timofte, and L. V. Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[41] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. *CoRR*, abs/1604.05417, 2016.

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[44] M. Uricár, C. FEE, R. Timofte, E. CVL, R. Rothe, J. Matas, and L. Van Gool. Structured output svm prediction of apparent age, gender and smile from deep features.

[45] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1879–1886, Washington, DC, USA, 2011. IEEE Computer Society.

[46] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. *arXiv preprint arXiv:1604.06433*, 2016.

[47] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[48] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790 – 799, 2014.

[49] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *CoRR*, abs/1603.05474, 2016.

[50] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *IEEE International Conference on Computer Vision*, 2015.

[51] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[52] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[53] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[54] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *arXiv preprint arXiv:1604.02878*, 2016.

[55] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[56] Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.

[57] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning.

[58] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. *arXiv preprint arXiv:1511.07212*, 2015.

[59] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, June 2012.

[60] X. Zhu and D. Ramanan. FaceDPL: Detection, pose estimation, and landmark localization in the wild. preprint 2015.