A Providing a hybrid method for face detection, gender recognition, facial landmarks localization and pose estimation using deep learning to improve accuracy

Peyman jabraelzadeh[1], Asghar charmin [2]*, Mohsen Ebadpour[3]
[1]*Department of Electrical Engineering Ahar branch, Islamic Azad university, Ahar, Iran*
[2]*Department of Electrical Engineering Ahar branch, Islamic Azad university, Ahar, Iran*
[3]*Department of Electrical Engineering Ahar branch, Islamic Azad university, Ahar, Iran*

Abstract

In general, identifying and locating faces in images or videos is considered as the first step in face recognition. It is quite clear that an accurate detection algorithm can significantly benefit system performance and vice versa. Therefore, face recognition is one of the key steps in the application of face recognition systems. In deep learning algorithms are able to learn high-level features, which have been highly regarded by researchers for use in the field of machine vision, as well as in a variety of fields such as image classification and human gesture estimation, which are the key activities for image perception. In this paper, we present a hybrid method called Hyper-Yolo-face to identify faces, facial landmarks localization, pose estimation and recognize the gender of a given image using deep convolutional neural networks, the Yolo algorithm, and local binary patterns. The proposed network architecture is based on the AlexNet model and the integration of the binary pattern operator and Yolov3, which results in increasing performance and accuracy. Yolo changes the architecture of face recognition systems and looks at the problem of recognition as a regression problem which goes directly from the pixels of the image to the coordinates of the box and the probability of the classes. Experiments on the AFLW and FDDB datasets indicated that the proposed model performs significantly better than other algorithms and methods and improves detection accuracy.

Keywords: Convolutional network, Face detection, Yolo, Pose estimation, Gender recognition, Facial landmarks localization, local binary pattern.

1. INTRODUCTION

Face recognition and analysis is a challenging issue in machine vision and is researched for some applications including face recognition, face tracking, face recognition, and more. Although methods based on deep convolutional neural networks achieved significant results in face recognition [1, 2 and 3], facial landmark localization, head pose estimation and gender recognition are difficult for facial images with extreme gestures, light, and resolution changes. This paper presents a new CNN-based method for face recognition, pose estimation, facial landmark localization, and gender recognition simultaneously in a given image. A CNN architecture was designed to learn the common features of these tasks and use the synergy between them. Since the binary pattern uses both statistical and structural features of the texture, it is a powerful tool for texture analysis. Therefore, it was used when feeding

*Corresponding Author Institutional Email: a_charmin@sut.ac.ir (A. charmin)

features to the grid, and input dimensions were upgraded to 6 dimensions, which improved the accuracy of diagnosis. In the local binary pattern operator, local texture patterns are extracted by comparing the value of adjacent pixels with the value of the central pixel and are represented by binary codes. The local binary model was first proposed by Ocala et al. in 1996. It is one of the most common descriptors due to its resistance to brightness changes, low computational complexity, and ability to encode details. Researchers use this method in most image processing researches to improve accuracy. As in [4], this fact is applied for distributing the information in the features across the network hierarchically. The lower layers are used to assess the edges and corners. Thus, they have better location properties and are more appropriate for identifying the landmarks of the image and estimating poses. However, the deeper layers are category-dependent and appropriate for learning complicated tasks such as identifying face and gender. In fact, different tasks could be taught by using all the middle layers of a deep CNN. Since a CNN architecture includes several layers with hundreds of feature mappings in each layer, it cannot be efficiently used for learning multiple tasks due to the total size of the super-features. In addition, it should be interrlated to encode common features across many tasks efficiently. Recent research on deep learning showed that CNNs can be employed to estimate the desired complex function. Therefore, a separate hybrid CNN was produced to mix the super-features. Several loss functions were simultaneously used to teach them how to do the tasks, i.e. recognizing the faces, which led to an increase in the performance of each individual task. The present study aims to propose a novel architecture of CNN to identify face and gender by integrtaing the middle layers of the network. This method focuses on the architecture of AlexNet model [5] and utilizes a Yolo algorithm named Yolo_face, instead of the selective search algorithm [6], on R-CNN [7] to provide area and face crop suggestions. This method can help quickly recognize and cut the faces of the images by using the introduced Yolo algorithm and entering them into the proposed network along with some other information added by Local Binary Patterns (LBP). The YOLO v3 architecture was applied for face recognition network and could successfully improve it by suggesting a definition for the loss function of the new regression including MSE losses and GIOU losses as well as more suitable bounding boxes to identify faces with k-means clustering. This paper is organized as follows. The review of literature is covered in Section 2. Details of the proposed method are described in Section 3. Implementations of the proposed deep CNN Hyper-Yolo-face approaches, as well as the results of the proposed approach on the data set are presented in Section 4. The conclusions are summarized and discussed in Section 5.

2. Review of the literature

Ramaya et al. [8] proposed a convolutional neural network to solve the problem of changes in brightness and head position in face recognition. In the proposed method, people are distinguished from each other using local patterns on their faces. The accuracy of this method was improved by 96.4% compared to the previous methods and by evaluating this method on the Yale

dataset, the detection accuracy was 95.99%. Gao et al. [9] proposed a new type of building block for deep architecture called an automatic encoder with an observer to identify the face by a training sample from each individual. In this method, all of the different faces are first modeled to map the normal face of each person. Then, the features related to the same person are extracted. Finally, they are extracted, which make face recognition easier using a self-encoder with the feature supervisor resistant to light scattering and opacity. By evaluating this method, the detection accuracy in AR datasets, Extended Yale-B, CMU-PIE, and Multi-PIE was 21.85¸22.82, 79.82, and 97.93%, respectively. Zhang et al. [10] proposed Sparse coding neural networks and Softmax classification for solving the problem of changes in brightness, state, and low image quality in face recognition. Face image preprocessing is used for hierarchical building and training of a deep network. The deep neural network is trained by a recursive algorithm and optimized using two different schemes. The evaluation of ORL datasets showed an identification accuracy of 97. 5%, Yale datasets gave an identification accuracy of 94.67%, Yale-B datasets had an identification accuracy of 82% and PERET datasets showed 92.78% of identification accuracy. Viola-Jones detector [11] is a traditional method which uses Haar-like feature classifications to identify faces. This method provides instant face recognition and works well for full faces with enough light. In addition, face recognition methods based on the Deformable Parts Model (DPM) [12] were presented in which a face is defined as a set of parts [13, 14]. It was shown that in face recognition without limitation, features such as HOG or Haar wavelets do not receive distinct face information in different poses or brightness changes. Various CNN-based face recognition methods were proposed to overcome these limitations [15, 16, 17, 18, and 19]. These methods generated new results on many of the challenging existing face recognition datasets. Other face recognition methods include NPDF Faces [20], Adapt [21] and [22]. One of the first approaches to consider the combined tasks of face recognition, pose estimation, and locating symbols was suggested in [23] and then developed in [24]. This model is based on a combination of a tree with shared reservoirs of parts, in which each face symbol is modeled as a part and uses total combinations to capture topological changes due to point-of-view changes. Recently, a cascade method was proposed in [25] to detect faces and landmarks localization simultaneously in a given image. This method improves detection by having a face alignment step in this cascade structure. Further, multi-task learning using CNNs was studied recently. Eigen and Fergus [26] proposed a multi-scale CNN to predict depth, surface normalities, and semantic tags simultaneously from an image. They used CNNs at different scales, in which the output of the smallest network scale is considered as the input of the larger scale. Yang and Ramanan [27] proposed the DAG-CNN method, which extracts features from multiple layers to have top, middle, and bottom features for image categorization. Sermanet et al. [28] combined the output of first class of CNN with the classifier input after sub-classification for pedestrian detection. Previous works on gender recognition

focused on finding good distinguishing features for classification. Various methods used a combination of one or more features such as LBP, SURF, HOG or SIFT. During recent years, feature-based methods have attracted a great deal of attention in face recognition. Binary classifiers were used in [29] for any characteristic including masculinity, long hair, white skin, etc. Separate features were calculated for different items and used to train each of the SVMs in each feature. CNN-based methods were also proposed for feature-based learning demonstrations [30, 31].

3. Proposed approach and architecture

This study proposes a hybrid CNN model for face recognition, gender recognition, landmarks localization, and pose estimation simultaneously. This proposed algorithm consists of two modules. The first module uses the proposed Yolo algorithm called Yolo_face, which crops faces from images and scales them to 227×227 pixels. The second module is a CNN which receives these resized cropped areas and classifies them as faces or non-faces, and then uses the LBP feature to increase the size of the input images to six color channels. The suggested network, including five convolutional layers with three fully connected layers (Fig. 1), is used in two ways. First, features on CNN are hierarchically distributed on the network, which are at a lower level for identifying the faces and estimating the pose. Instead, the properties of the higher layers are appropriate for more complex tasks such as detecting or classifying [34]. In addition, the simultaneous learning of several interrelated tasks can result in synergy and improve the performance of each task as reported in previous studies [35, 36]. All intermediate layers were not used for combination since the nearby layers were highly interrelated. The max_1, $conv_3$, $pool_5$ layers from Alexnet were mixed by a separate network. The direct combination of these features is known as a simple method. Since the mappings of this feature for these layers have different dimensions ($6 \times 6 \times 256, 13 \times 13 \times 384,$ and $27 \times 27 \times 96,$ respectively), they cannot be easily combined with each other. Therefore, $conv_{1a}$ and $conv_{3a}$ convolutional layers were added to the $pool_1$ and $conv_3$ layers in the output to get feature mappings compatible with the $6 \times 6 \times 256$ dimensions. Next, the outputs of these layers were mixed with $pool_5$ to create a feature mapping with $6 \times 6 \times 768$ dimensions, which are too large to train a multi-task framework. Therefore, a core convolutional layer ($conv_{all}$) $1 \times 1$was added to decrease these dimensions to $6 \times 6 \times 192$. The fully connected (fc_{all}) layer was added to $conv_{all}$ with feature vector outputs of 3072 dimension. Next, the network was seperated into five separate branches for each task.

The fully connected $fc_{detection}$, $fc_{landmarks}$, $fc_{visibility}$, fc_{pose}, fc_{gender} layers with 512 dimensions were added to fc_{all}. Finally, a fully connected layer was added to each branch to use the labels for each task. An activation function (ReLU) was used after each convolution or a fully connected layer. No integration function was conducted in this hybrid network and resulted in local immutability, which was not appropriate for localizing the facial landmarks. Then, task-specific loss functions were employed to learn the weights related to the network.

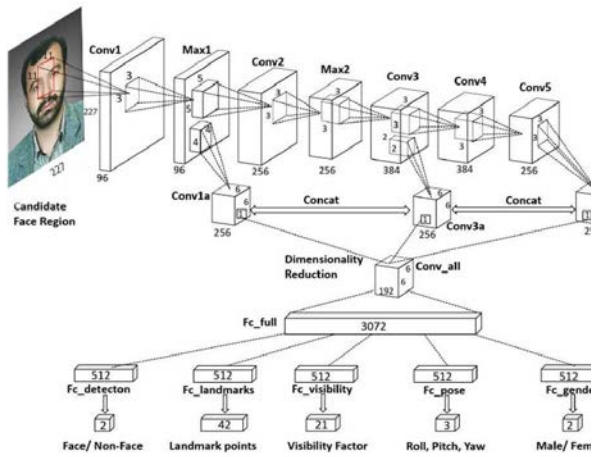

Fig.1. The structure of the proposed network

The original LBP operator is introduced as a powerful descriptor for image texture [32]. This operator generates a binary number for each pixel according to the adjacent 3×3 pixel labels. Labels are obtained by thresholding the value of neighboring pixels with the center pixel value. Thus, the label is 1 for pixels with a value greater than or equal to the value of the center pixel, while the label is 0 for pixels with values less than the value of the center pixel. These labels are then rotated side by side to form an 8-bit number. The performance of this operator is shown in Figure 2.

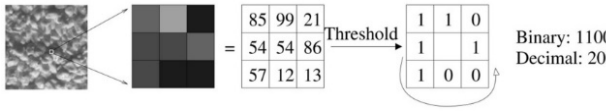

Fig. 2. Operator of local binary patterns [32]

During the recent years, the LBP method underwent changes in order to improve performance in various applications such as improving separation strength, increasing tolerance toward changes, selecting a neighbor, and combining it with other methods. In this method, for each RGB color channel, the LBP operator was applied separately on each of the color channels of the input images to the network to increase the channel size of each image to 6 dimensions. The output of this operator on the sample image is given in Figure 3.

Fig.3. LBP output on three RGB channels of image

If an area is categorized as a face, it can provide the facial landmarks and estimate the head pose with respect to the camera and gender information. The cropped face from the images by this hypothetical Yolo was a candidate to be used in the main network for estimating other parametersincluding gender detection and pose estimation. The backbone of the face recognition network included Darknet-53. It should be noted that the feature extraction network was based on Darknet. In order to obtain a suitable bounding box for face detection, the first stage was to adapt the number of the searches (seed) k experimentally for the clustered bounding boxes. Then, randomly k is the bounding box as the initial clustering centers, and then calculating the IoU of the bounding boxes of k and all other bounding boxes. All face labels were divided into k classes by employing the IoU as the intersection criterion for bounding

boxes. Next, the mean values of the size of bounding boxes of k class were considered as the centers of the new cluster, which were repeated several times to reach convergence. The centers of the initial k cluster were divided into 9 classes in the experiments. YOLO optimizes a multi-part loss function while training this model , including reliability, objective function of regression, , classification, and responsibility for the absence of any object. However, face recognition is consisdred as a binary classification problem. The weights experimentally changed to 2: 1: 0.5: 0.5 in order to make the total objective function more appropriate for recognizing face . The final objective function is obtained as Eq. 1:

$$L = 2 \cdot \sum L_{reg} + \sum L_{objconf} + 0/5 \cdot \sum L_{noobjconf} + 0/5 \cdot \sum L_{cld} \quad (1)$$

Where L_{reg} is the coordinate regression loss, $L_{narconf}$ demonstrates the confidence loss of the bounding box with objects, $L_{noobjconf}$ denotes the loss of trust to bounding boxes with no objects, and L_{cls} shows the classification loss. The predicted location of IoU and the related labels of the corresponding monitored data are usually used to evaluate optimization, and the MSE function is applied as the regression loss. However, a gap was available between maximizing IoU and optimizing MSE.

Particularly, it is impossible to optimize non-intersecting boxes. A generalization to the IoU was suggested as a new metric called GIoU to address this weakness. There is a strong relationship between optimizing the MSE function and the metric itself in the new metric. Based on [33], the regression loss function was enhanced by integrating the main soft error l_n with the weight loss of GIoU generalization. The new regression loss could be computed as Eqs. 2, 3, and 4.

$$GIoU = IoU - \frac{A_{c-U}}{A_C} \quad (2)$$

$$L_{GIoU} = 1 - GIoU \quad (3)$$

$$L_{reg} = \sum_{c=x.y.w.h} \sum \left(|\Delta c_{pred} - \Delta c_{truth}| + \alpha \cdot L_{GIoU}\right)^2 = \sum \left(|\Delta x_{pred} - \Delta x_{truth}| + \alpha \cdot L_{GIoU}\right)^2 \sum \left(|\Delta y_{pred} - \Delta y_{truth}| + \alpha \cdot L_{GIoU}\right)^2 + \sum \left(|\Delta w_{pred} - \Delta w_{truth}| + \alpha \cdot L_{GIoU}\right)^2 + \sum \left(|\Delta h_{pred} - \Delta h_{truth}| + \alpha \cdot L_{GIoU}\right)^2 \quad (4)$$

where A_c is the smallest convex set enclosingthe predicted location and the correct labels, α shows a real value factor, and x, y, w, and *h* represent the locations and size of the bounding boxes, respectively. The α factor was set as 0.1 in this model.

3.1 Training and testing process of network

The AFLW dataset [36] was used to train and test the proposed network. This dataset contains 25,993 faces with 21,997 full real-time images, face shapes, race, age and gender. Further, there are 21 landmarks for each face with a box bounded to the face, face pose (left and right bending, up and down bending and rotation) and gender information. From this collection, 2400 images were randomly selected for testing and the rest for training. Various loss

functions were used to train facial recognition tasks, facial landmarks localization, pose estimation, and gender recognition.

Face recognition in images: The proposed Yolo algorithm was applied for face recognition and face cropping. The architecture and loss functions of this algorithm are presented and discussed in Section 3.

Facial landmark localization: A total of 21 markup language points were utilized to localize facial landmarks as the AFLW dataset. Some of these landmarks become non-visible due to the fact that these faces have different perfect poses. The candidate box areas with intersection (IOU) arger than 0.35 of the target label map were used to learn this task while others were neglected. An area could be determined by {x, y, w, h}, where the (x, y) is the coordinates of the center of the area and w and h denote the width and height, respectively. Each point of the visible sign was replaced based on the center of the area (x, y) and normalized by (w, h) as Eq. (5).

$$(a_i . b_i) = \left(\frac{x_i - x}{w} . \frac{y_i - y}{h}\right) \quad (5)$$

Where the (xi, yi) reliability coordinates are the basis of known correctness, and (ai, bi) are used as labels for training the localization task using weighted Euclidean loss with a coefficient of visibility. This loss was applied to predict the location of the landmarks calculated from Eq. (6).

$$loss_L = \frac{1}{2N}\sum_{I=1}^{N} v_i\left((\hat{x}_i - a_i)\right)^2 + ((\hat{y}_i - b_i)^2) \quad (6)$$

Where the $(\hat{x}_i, \hat{y}_i)$ is the ith location of the predicted landmarks by this network with respect toa known area and N denotes the total number of landmark points (21 for AFLW [37]). The coefficient of visibility (vi) is 1 when the ith landmark is obserevd in the candidate area. Otherwise, it is zero, i.e., there is no loss related to the non-visible points. Therefore, it will not considered in the post-release phase.

Visibility learning: The coefficient of visibility can test the presence of a facial landmark on the predicted face. Regarding an area with an intersection greaterthan 0.35, the Euclidean loss was utilized to train visibility as in Eq. (7).

$$loss_v = \frac{1}{N}\sum_{I=1}^{N}\left((\hat{v}_i - v_i)\right)^2 \quad (7)$$

where $\hat{v}_i$ indicates the predicted visibility of the ith landmark. The true visibility (v_i) is one when the ith landmark is visible in the candidate area. Otherwise, it is equal to zero.

Pose estimation: Euclidean loss was used to train the estimation of the head rotation (P1), bending towards up and down (P2), and bending towards left and right (P3). The loss of a candidate area with more than 0.5 intersection with the target map is calculated by Eq. (8).

$$loss_P = \frac{(\hat{p}_1 - p_1)^2 + (\hat{p}_2 - p_2)^2 + (\hat{p}_3 - p_3)^3}{3} \quad (8)$$

Where $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ represent the estimated poses.

Gender recognition: Gender is a two-part issue similar to face recognition. For a candidate area with 0.5 intersection with the target map, the Softmax loss was calculated based on Eq. (9).

$$loss_G = -(1 - g) \cdot log(1 - p_g) - g \cdot log(p_g) \quad (9)$$

where g=0 when the gender is male; otherwise, g is 1. (p0, p1) is the two-dimensional probability vector measure from this network. The total loss is calculated as the weighted sum of each of the four losses as in Eq. (10).

$$loss_{full} = \sum_{i=1}^{i=5} \lambda_{t_i} loss_{t_i} \qquad (10)$$

where ti indicates the ith entry of the T = {D, L, V, P, G} set of tasks. The weight parameter (λ_{t_i}) is decided based on the importance of this task in the overall loss. The $(\lambda_D = 1, \lambda_L = 5, \lambda_V = 0.5, \lambda_P = 5, \lambda_G = 2)$ values were chosen for the experiments. More weights were assigned to the tasks of locating the facial landmarks and estimating the pose due to the need for spatial accuracy.

4. Analysis of results and experiments

Face recognition results for the AFW and FDDB datasets are presented in this section. The AFW [38] dataset is collected by Flicker, and the images in this dataset contain many changes in appearance and point of view. There are a total of 468 faces in this dataset. The FDDB database [39] consists of 2,845 images including 5,171 images collected from news articles on the Yahoo web site. Some recently published methods which were compared in this evaluation include DP2MFD [40], CascadeCNN [41] and Hyper face [42]. FDDB dataset is very challenging for the proposed method and other R-CNN-based face recognition methods, which could be due to its multiple blurry and small faces. Some of these faces are not in the candidate search areas. Additionally, resizing small faces to the 227×227 input size further distorts the face, which can lead to a low detection score. However, the performance of the proposed model could be compaed with some recently published deep learning face recognition methods such as DP2MFD [40] and Faceness [43] on the FDDB dataset with 91.1% mAP. Figs. 4 and 5 display the precision-recall curves of the various detectors related to the AFW and PASCAL face datasets, respectively. Furthemore, Fig. 6 shows the comparison of the performances of various detectors by Receiver Operating Characteristic (ROCs) curves on the FDDB dataset. As can be seen, the suggested method outperformed all the commercial and academic detectors reported on the AFW and PASCAL datasets. Hyper-Yolo-face has an average accuracy of 99.3% (mAP) and 98.20% for the AFW and PASCAL datasets, respectively.

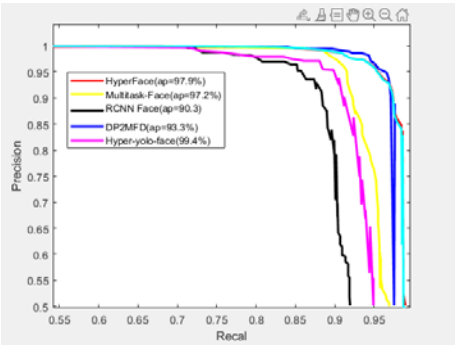

Fig. 4. Evaluation of face recognition performance on the AFW dataset (The numbers in the guide represent the mean accuracy (mAP) for the relevant dataset)

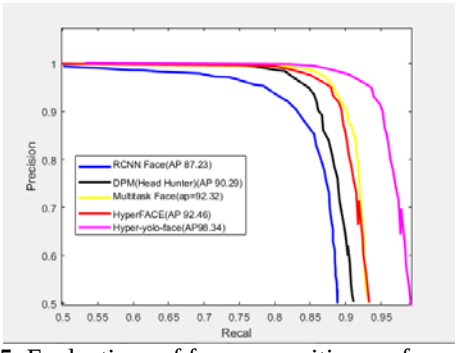

Fig.5. Evaluations of face recognition performance on the PASCAL face dataset (The numbers in the guide represent the mean accuracy (mAP) for the relevant data set)

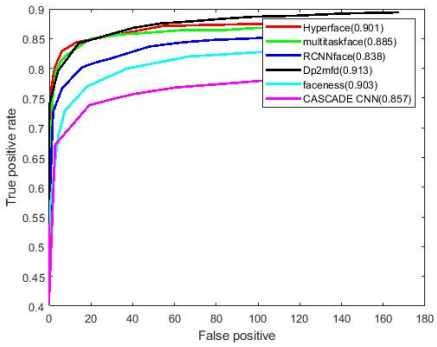

Fig.6. Evaluation of face recognition performance on the FDDB dataset (The numbers in the guide represent the mean accuracy)

Figures 5 and 6 illustrate that multi-task CNNs (Multitask_Face and Hyper-Yolo-face) operate with a wider margin than R-CNN-Face. This increase in performance is mainly related to the fact that the multi-task learning approach helps to the network to learn improved features for face recognition, as evidenced by their mAP values in the AFLW dataset.

The performance of various facial landmarks localization algorithms was evaluated on the AFW [38] and AFLW [36] datasets. Both of these datasets contain faces with complete pose changes. Some of the compared methods in terms of facial landmarks localization include FaceDPL [44], JointCascade [45], SDM [46] and 3DDFA [47]. The performance of different methods of facial landmark localization on the AFW dataset using the defined protocol in [44] is shown in Figure 7. As shown, (*) represents the models evaluated on full close-up faces or using manual initial values [38]. This dataset contains six main points for each face includingleft_eye_center, right_eye_center, nose_tip, mouth_left, mouth_center, and mouth_right. The error was calculated as the average distance between these predicted main points and the correct normalized labels to the size of the face. Diagrams for this comparison are obtained from [44].

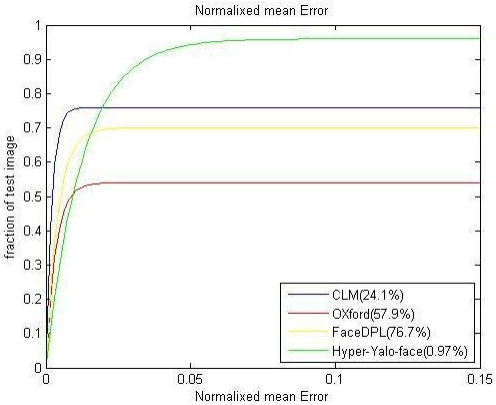

Fig.7. Cumulative error distribution curves for facial landmark localization on the AFW dataset (The numbers in the guide are part of the test faces which have an average error lower than (5%) of the size of the face)

This error was calculated for the AFLW dataset using all visible points. The same protocol defined in [47] was used for AFW. Here, the only difference was that the AFLW test set contained only 1000 images with 1132 face samples and the rest of the images were used for training. It randomly generated a subset of 450 samples from this test set with left and right bending angles of 1/3,[60°, 90°], [30°,60°], [0°,30°] to comply with this protocol. The performance of different methods of facial landmark localization is compared in Figure 8. The diagrams of this comparison are obtained from [47] in which evaluations for RCPR, ESR and SDM are performed by using these algorithms for face file processing. The normalized mean error (NME) for the AFLW dataset for each pose group is given in Table 1.

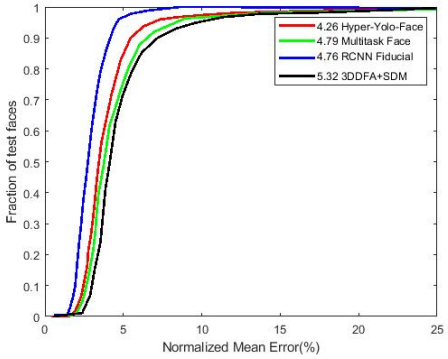

Fig.8. Cumulative error distribution curves for landmarks localization on the AFLW dataset (The numbers in the guide represent the average NME for the test images.) The test samples were selected so that they are 1/3 of the absolute left and right bending angle ,[60°, 90°], [30°,60°], [0°,30°])

As shown, the proposed Hyper-Yolo-face performs better than many new methods of facial landmarks localization including FaceDPL [44], 3DDFA [47] and SDM [46]. Table 1 shows that the proposed method worked accurately on all pose angles.

Table 1. NME (in percentage) of the alignment results of face on the AFLW test set with the best results

	AFLW Dataset 21(pts)				
Method	[0,30]	[30, 60]	[60, 90]	Mean	std
CDM[62]	8.15	13.03	16.17	12.44	4.04
ESR[4]	5.66	7.12	11.94	8.24	3.29
SDM[56]	4.75	5.55	9.34	6.55	2.45
3DDFA [69]	5.00	5.06	6.74	5.60	2.45
Hyper face[42]	3.93	4.14	4.71	4.26	0.41
Hyper-Yolo-face[ours method]	3.00	2.89	3.56	3.90	0.38

The proposed method on the AFW dataset [37] was evaluated for the gesture estimation task. Recognition boxes are used to evaluate the facial landmarks localization and initialization. For the AFW dataset, the proposed approach was compared with Multi.AAM [38], HoG with multiple view angles [38], FaceDPL [44] and face.com [] and Hyperface [42]. Cumulative error distribution curves on the AFW dataset are given in Figure 9. This curve shows the part of the faces where the estimated pose is in the fluctuation range of error. As displayed , the proposed method operates with a much better margin than the existing methods.

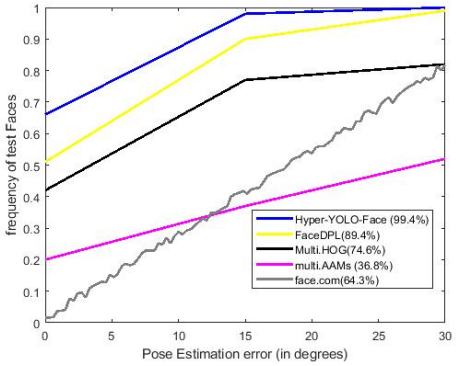

Figure 9. Cumulative error distribution curves of pose estimation on the AFW dataset (The numbers in the guide represent the percentage of faces labeled in $\pm 15°$ of error fluctuations)

The performance of gender recognition was evaluated on the CelebA [48] and LFWA [49] datasets since these datasets contain gender information. The CelebA dataset contains 10,000 attributes and 200,000 images. The LFWA dataset contains 13,233 images with 5,749 attributes. The proposed approach was compared with FaceTracer [50], PANDA-w [51], PANDA-1 [51] and Hyper face [42]. The performance of different methods of gender recognition is reported in Table 2. The proposed method performs best on the LFWA dataset compared to all of the methods listed in the table.

Table 2. Comparison of performance (by percentage) of gender detection on the CelebA and LFWA datasets

Method	CelebA	LFWA
FaceTracer[31]	91	84
PANDA-W[64]	93	86
PANDA-1[64]	97	92
Hyperface	97	94
Hyper-Yolo-face[oure method]	99	100

5. Discussion and conclusion

This paper presented a Hyper-Yolo-Face multi-task deep learning method for face recognition, facial landmark localization, head pose estimation, and gender recognition simultaneously. Numerous experiments have demonstrated the effectiveness of this method in all four tasks using multiple publicly available datasets. In the future, the performance of this method will be evaluated in other applications including human recognition and human pose estimation, object identification and pedestrian recognition simultaneously. Based on the results, some observations are presented. First, all facial functions benefit from using a multi-task learning framework. This benefit is mainly due to the network's ability to learn more distinguishing features and post-processing techniques which can be improved by facial landmark localization and recognizing points for an area. Second, the composition of the middle layers improves this function for tasks related to the structure of pose estimation and locating the prominent points of the face because these properties do not change in the deeper layers of CNN relative to geometry. The Hyper-face uses these observations to improve the performance of all four tasks. Several qualitative results of the proposed method on the AFW, AFLW and FDDB datasets are shown in Figure 10. As observed, the proposed method can simultaneously perform all four tasks on images including poses, brightness, and sharp resolution changes with a cluttered background.

Figure 10. Qualitative results of the proposed method (The blue boxes represent the recognized faces of the women and the pink boxes represent the recognized faces of the man. Pose estimation on each face is shown at the top of these boxes for rotation, up and down bending, and left and right bending, respectively)

6. References

[1]. S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In International Conferenc on Multimedia Retrieval, 2015.

[2]. 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.

[3]. 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.

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

[5]. 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.

[6]. 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.

[7]. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation.In Computer Vision and Pattern Recognition, 2014.

[8]. Ramaiah, N. P., Ijjina, E. P., & Mohan, C. K. (2015, February). Illumination invariant face recognition using convolutional neural networks. In Signal Processing, Informatics, Communication and Energy Systems (SPICES), 2015 IEEE International Conference on (pp. 1-4). IEEE.

[9]. Gao, S., Zhang, Y., Jia, K., Lu, J., & Zhang, Y. (2015). Single sample face recognition via learning deep supervised autoencoders. IEEE Transactions on Information Forensics and Security, 10(10), 2108-2118.

[10]. Zhang, Z., Li, J., & Zhu, R. (2015, October). Deep neural network for face recognition based on sparse autoencoder. In Image and Signal Processing (CISP), 2015 8th International Congress on (pp. 594-598). IEEE.

[11]. P. A. Viola and M. J. Jones. Robust real-time face detection.International Journal of Computer Vision, 57(2):137–154, 2004.

[12]. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan.Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence,32(9):1627–1645, Sept 2010.

[13]. 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.

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

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

[16]. [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.

[17]. 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.

[18]. S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In International Conference on Multimedia Retrieval, 2015.

[19]. B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features.In IEEE International Conference on Computer Vision, 2015.

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

[21]. H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In IEEE International Conference on Computer Vision, pages 793–800, Dec 2013.

[22]. 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.

[23]. 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.

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

[25]. 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.

[26]. D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture.In Proceedings of the IEEE International Conference on Computer Vision, pages 2650–2658, 2015.

[27]. S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns.In The IEEE International Conference on Computer Vision (ICCV), December 2015.

[28]. P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun. Pedestrian detection with unsupervised multi-stage feature learning.In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’13, pages 3626–3633, Washington, DC,USA, 2013. IEEE Computer Society.

[29]. N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In European Conference on Computer Vision (ECCV), pages 340–353, Oct 2008.

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

[31]. 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.

[32]. Ojala, T., Pietikäinen, M., and Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 7 (2002), 971–987.

[33]. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

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

[35]. Z. Zhang, P. Luo, C. Loy, and X. Tang. Facial landmark detectionby deep multi-task learning. In European Conference on ComputerVision, pages 94–108, 2014.

[36]. 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.

[37]. 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.

[38]. 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.

[39]. 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.

[40]. 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.

[41]. 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.

[42]. 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 TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2016.

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

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

[45]. 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.

[46]. 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.

[47]. X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. CoRR, abs/1511.07212, 2015.

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

[49]. 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.

[50]. N. Kumar, P. N. Belhumeur, and S. K. Nayar. aceTracer: A Search Engine for Large Collections of Images with Faces. In European Conference on computer Vision (ECCV), pages 340–353, Oct 2008.

[51]. 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.