MULTI-TASK LEARNING FOR FACE IDENTIFICATION AND ATTRIBUTE ESTIMATION

Hui-Lan Hsieh, Winston Hsu

National Taiwan University
Taipei, Taiwan

Yan-Ying Chen

Palo Alto Laboratory

ABSTRACT

Convolution neural network (CNN) has been shown as one of state-of-the-art approaches for learning face representations. However, previous works only utilized identity information instead of leveraging human attributes (e.g., gender and age) which contain high-level semantic meaning. In this work, we aim to incorporate identity and human attributes in learning discriminative face representations through multi-task learning. In our experiments, we learn face representation by using the largest publicly face dataset CASIA-WebFace with gender and age labels, and then evaluate learned features on widely-used LFW benchmark for face verification and identification. We also compare the effectiveness of different attributes for improving face identification. The results show that the proposed model outperforms the baseline CNN method without using multi-task learning and hand-crafted features such as high-dimensional LBP. We also do experiments on gender and age estimation on Adience benchmark to demonstrate that human attribute prediction can also benefit from the proposed multi-task representation learning.

***Index Terms*—** Face representation, multi-task learning, human attribute

1. INTRODUCTION

Face recognition has seen significant improvements recently using deep learning. Most works focus on face verification tasks, that usually only utilize identity information to learn face representations. However, with the advanced technology of human attribute detection, we also noticed that face recognition and human attributes are correlated and could complement each other. In Figure 1, we provide a multi-task learning framework by incorporating high-level human attributes with CNN method to learn more semantic face representations. A similar idea is proposed in [1] using fisher vectors with attributes for large-scale image retrieval, but they use early fusion to combine attribute scores and did not address face representations. There are other methods that leverage human attributes to achieve promising results in the fields including face verification [2], face identification [3], large scale face retrieval [4], but these work mainly use hand-crafted features while our work can take advantage of the powerful representations in the CNN framework.

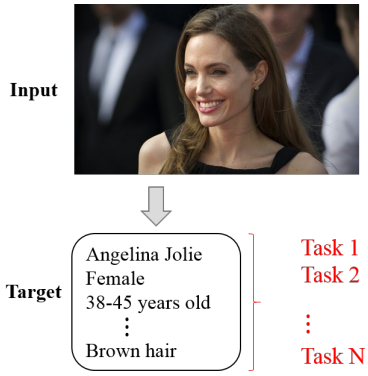


Fig. 1. We aim to learn discriminative features by incorporating high-level human attributes to improve face identification in the CNN framework. In a multi-tasking setup, we also exploit identity information to improve human attribute detection.

Face recognition could be referred to two types of tasks, face verification and face identification. The protocol of face verification is to decide whether two images are the same person or not. Current state-of-the-art methods achieve nearly or even outperform human recognition ability. However, face identification is still an unsolved problem. In [5], face identification is proved to be a more difficult problem than face verification. The goal of face identification is to identify a query image and retrieve the subject rank list from a gallery database. There are very large gallery database with labeled identities, and the scale of classes may reach up to millions or more. To make the retrieval results more precise, discriminating the subtle difference across the subjects in gallery database is important and critical. In this work, we argue that learning more discriminative features by leveraging attribute information makes face identification more precise and efficient in CNN framework.

Our main contributions are to propose a multi-task learning model that (1) first utilizes the human attributes to improve face identification and in return (2) improve attribute

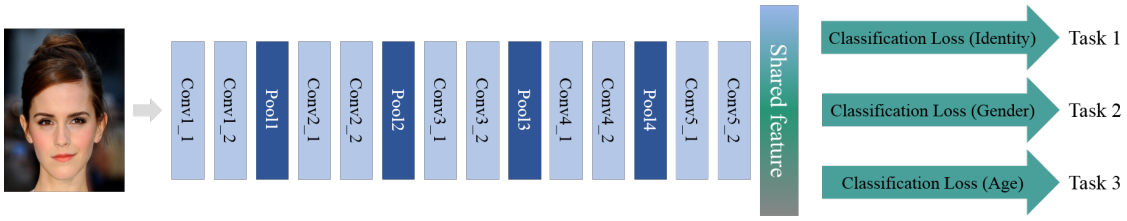


Fig. 2. Illustration of our CNN architecture. The shared features would be the input of multiple loss layers to classify human attributes and human identity. By utilizing the regularized multiple loss function, we can learn better representations for face identification and attribute detection.

detection by taking the advantages of rich identities incorporated in the same model.

2. RELATED WORK

Face verification using CNN features [6] has been shown outperforming most hand-crafted features. The current results in LFW benchmark are even approaching to human recognition ability. However, these models did not utilize attributes which have strongly correlated with face identification. Traditional methods have the similar idea for image retrieval [1], but do not specifically address human attributes that are special and can provide high-level semantic information (e.g., gender, race, and age) to understand an identity. Unlike the traditional methods, our features are learned from CNN-based model by incorporating face identity and attributes instead of hand-crafted face features such as high-dimensional LBP.

Predicting facial attributes from faces in the wild is also challenging due to the lighting and pose variations. Traditional methods usually extract hand-crafted features at predefined landmarks for the specific attribute recognition [7, 8, 9]. Recently, many researchers have used deep learning to achieve great success in this field. For gender esitmation, the authors conducted experiments on LFW benchmark by means of LBP features with an AdaBoost classifier [10]. In [11], Gil *et al.* did experiments with CNN on Adience benchmark which contains face images in an unconstrained environment collected by [12]. For age estimation, early methods are based on localizing facial landmarks and then extracting features to analyze ratios and measure wrinkles [13]. Some researchers represent the age process as a subspace [14] or a manifold [15] to estimate accurate age. The aforementioned methods require input images to be well-aligned, thus they conducted experiments on images that were taken under constrained environments. Our main goal is designing a CNN framework exploiting both identities and attributes to improve the performance of each task. We also investigate whether abundant identity information also helps attribute estimation.

Multi-task learning has proven effective in many computer vision problems. The multi-task strategy avoids the complex architecture and can leverage large amount of cross-task data. It is also expected to improve the performance because of the additional information. Facial landmark detection can also utilize some specific face expression related attributes. In [16], it uses more complicated multi-task learning with early stopping, but also learns a common features from CNN. That also motivates us to incorporate human attributes to multi-task learning and make use of rich attributes to learn face representations.

We will introduce the details of our method in Section 3. The experimental results are discussed in Section 4 and followed by the conclusion in Section 5.

3. PROPOSED METHOD

Our main concept of the proposed method is to embed the human attribute information into face representations. This network combines several useful and important characteristics, including very deep architecture, low dimensional representation and small filters. In this section, we will first describe the network structure we use in this paper. Next, we brief multi-task learning, and formulation of loss functions using in CNN model.

3.1. CNN Structure

Figure 2 illustrates our framework. The basic component of our model exploits very deep architecture and do not have fully connected layers. The proposed network includes 10 convolutional layers, 5 pooling layers, and fully connected output layers that map to different tasks. Instead of using many fully connected layer, We replace them with small filters to lower the complexity and parameters of CNN model and still achieve high performance. The size of all convolutional filters are 3×3, and each of convolutional layer is followed by ReLU and Normalization. Except pool5, all the pooling layers are max pooling. In the training stage, we use multiple loss functions to regularize pool5, and make the face representation embedded with semantic meaning in human attributes. We extract pool5 as the features of face images for face identification and verification.

3.2. Multi-task Learning

Instead of training on identity then fine-tuning by attributes, we make use of multiple loss function by multi-task learning. These ideas also decrease the needed size of dataset and reduce the computation efforts. In general, multi-task learning seek to improve the performance of multiple related tasks by learning them jointly. In this paper, we have one main task for identities, and two human attribute detection tasks. The weights of different tasks are different, so we set each task with distinct λ by cross validation. As a result, our loss function can be written as

$$\sum_{w^t} \lambda^t L(y^t, (x_i^t; w^t)) + R(w^t) \tag{1}$$

3.3. Training and Testing

Initialization. The network is trained from scratch using CASIA-WebFace, and all filters are initialized with random values from a zero mean Gaussian with standard deviation of 0.01. The learning rate is set to 0.01 at first 10,000 iteration and decreases gradually. For every images, we perform alignment with the height of eyes, then resize to 100×100. We train images with RGB channel because the color is highly related to human attributes. The corresponding output label of an input image is binary for gender and spans three classes for age separately because of limited by the lack of dataset with enough amount of images with identity and accurate age label.

Training. We only use CASIA-WebFace dataset to train from scratch without fine-tuning by any other external datasets. Input images are first aligned with eyes and scaled to 100×100. We also mirror images to augment the dataset and improve robustness of our model. All the color channels are processed directly by the network as we need to learn human attributes which are related to RGB. To avoid overfitting, we apply dropout learning after pool5, and the dropout ratio is 0.5 (50% chance of setting a neuron's output value to zero). Finally, the learning rate decreases gradually from 0.01 to avoid falling to local minima.

Testing. In the testing stage, we introduce two benchmarks to evaluate our proposed method. We focus on *Closed Set Identification* protocol proposed in [5] based on LFW benchmark to evaluate face identification. In gender and age estimation tasks, we evaluate gender and age estimation on Adience Benchmark with single-task learning and multi-task learning for comparison.

4. EXPERIMENTS

We use CASIA-WebFace dataset as the training and testing data. To get the groundtruth gender and age attributes of the dataset, we first crawled the information on IMDb celebrity profiles to get the gender information of images in CASIA-WebFace. We also used multi-class CNN classifiers to detect the age information of face images in CASIA-WebFace and corrected some obviously wrong labels by human efforts. Second, we train a multi-task CNN model on CASIA-WebFace with identity and attribute labels to learn the face representation.

4.1. Dataset

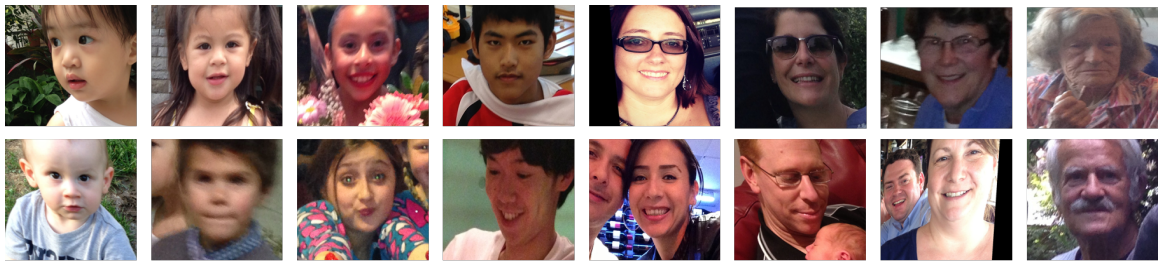
CASIA-WebFace. CASIA-WebFace [17], the largest public face dataset, has 10575 subjects and 494,414 face images crawled from IMDb photo gallery. Actually, this dataset contains only identity labels, thus we make the use of identities to collect gender information from internet and IMDb. We use about 40k images for training CNN model and the remaining images for validation.

LFW. LFW [18, 19] is commonly used to evaluate and report the performance of face verification. This benchmark consists of 13,323 web photos of 5,749 celebrities. We conducted experiments on *Closed Set Identification* proposed in [5]. In closed set identification, the 596 subjects who have at least two images in the LFW are used as gallery set. Only one frontal face image per subject is placed in the gallery, and the remaining images are used as probes (same as query database). The gallery set is further extend to 4,249 identities by adding 3,653 subjects with only a single image in LFW. As a result, all probe faces have a true match in the gallery.

Adience Benchmark. Adience benchmark [12] consists of roughly 26K images of 2,284 subjects, and most images have correspondent age and gender information labeled by human efforts. We follow the training and testing protocol in [12, 11] to evaluate the performance of our proposed method.

4.2. Results on Face Recognition

Table 1 presents our results for face verification and face identification on closed set, and we compare with high-dimensional LBP features and CNN baseline model which is trained without any attribute. Although our CNN baseline model and high dimensional LBP have achieved 93% in face verification, the results in face identification are not as good as those in face verification. For multi-task learning (denoted as MTL in the following section) with gender, we found that the rank-1 performance increased by 5% than baseline CNN, and the overall results are better than LBP and the baseline model. The performance of MTL with age are better than baseline model and high dimensional LBP too. In summary, face verification and identification can benefit from incorporating human attributes into training. The results show that we can achieve better performance in face verification and identification by utilizing multi-task learning with human attributes.



(a) 0-2 (b) 4-6 (c) 8-13 (d) 15-20 (e) 25-32 (f) 38-43 (g) 48-53 (h) 60-100

Fig. 3. The correct results of age classification for each age group. Although some samples are blurry or wearing glasses, we can recognize it correctly.

Table 1. The proposed method increases by 3% with the help by gender attribute, and 1% with the help by age attribute. Performance is measured in terms of the verification accuracy (%) under restricted protocol, and Rank-1 accuracy (%) on Closed Set. Please see details in Section 4.2.

Method	Rank 1	Rank 10	Rank 100	Verification
LBP	26.9	42.4	63.3	93.18
Baseline (w/o attribute)	47.9	76.8	91.7	93.06
MTL (with gender)	**50.8**	**76.9**	**92.7**	93.66
MTL (with age)	48.6	76.3	92.2	93.48

Table 2. Gender estimation results on Adience benchmark. Our proposed method is superior to the prior works.

Method	Accuracy
LBP in [12]	76.1
LBP in [20]	79.3
CNN in [11]	86.8
STL on gender (w/ finetuning)	87.5
MTL (w/o finetuning)	83.6
MTL (w/ finetuning)	**89.1**

Table 3. Age estimation results on Adience benchmark. Our proposed method is the best comparing to the prior works.

Method	Accuracy	1-off Age Accuracy
LBP in [20]	45.1	79.3
CNN in [11]	50.7	84.8
STL on age (w/ finetuning)	52.1	85.1
MTL (w/ finetuning)	52.6	**86.7**

4.3. Results on Gender and Age Estimation

In this section, we evaluate whether identity information also improves the gender and age classification. In Table 2 and Table 3, we show the results of our proposed model comparing with hand-crafted feature [20, 12] and a baseline CNN based method [11]. We also compare the difference between single-task learning (denoted as STL in the following section) model and MTL model.

For gender estimation, our MTL model without finetuning on Adience Benchmark achieves comparably results compare to the baseline CNN model [11], an AlexNet-like CNN. After finetuning on Adience benchmark, our performance increase roughly 2% than STL model and 3% than the baseline CNN model. The result of MTL model with finetuning achieves the best performance on Adience benchmark.

For age estimation, we finetune our MTL model on Adience bechmark because the number of age groups is different. In this dataset, there are about 20,000 images categorized into 8 age groups. We also provide the result of STL model compare to MTL model. The results in Fig. 3 show that even blurry photos or samples with glasses, we can estimate them correctly. Our proposed method also achieves better accuracy than hand-crafted features and the baseline CNN based model [11]. The MTL model also increase 1% accuracy compared to the STL model.

5. CONCLUSIONS

In this work, we propose a multi-task learning strategy to embed attribute information into face representations. We demonstrate that the face identification and verification can benefit from the incorporated attributes. Besides, we also demonstrate the improvement of attribute estimation with the help of identity information. In the future, we may collect more data to reduce the bias of CASIA-WebFace dataset. To discuss the relation between face recognition and attribute estimation, we want to introduce more attributes such as the style of hair and its color, the characteristic of skin color, eyebrow, lips, and so on into our experiments.

6. REFERENCES

[1] Matthijs Douze, Arnau Ramisa, and Cordelia Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 745–752.

[2] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Describable visual attributes for face verification and image search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1962–1977, 2011.

[3] Walter J Scheirer, Neeraj Kumar, Karl Ricanek, Peter N Belhumeur, and Terrance E Boult, "Fusing with context: a bayesian approach to combining descriptive attributes," in *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 2011, pp. 1–8.

[4] Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H Hsu, "Scalable face image retrieval using attribute-enhanced sparse codewords," *Multimedia, IEEE Transactions on*, vol. 15, no. 5, pp. 1163–1173, 2013.

[5] Lacey Best-Rowden, Hu Han, Christina Otto, Brendan F Klare, and Anubhav K Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *Information Forensics and Security, IEEE Transactions on*, vol. 9, no. 12, pp. 2144–2157, 2014.

[6] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[7] Thomas Berg and Peter Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 955–962.

[8] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1543–1550.

[9] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Attribute and simile classifiers for face verification," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 365–372.

[10] Caifeng Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431–437, 2012.

[11] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[12] Eran Eidinger, Roee Enbar, and Tal Hassner, "Age and gender estimation of unfiltered faces," *Information Forensics and Security, IEEE Transactions on*, vol. 9, no. 12, pp. 2170–2179, 2014.

[13] Young H Kwon and Niels da Vitoria Lobo, "Age classification from facial images," *Computer Vision and Image Understanding*, vol. 74, no. 1, pp. 1–21, 1999.

[14] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 12, pp. 2234–2240, 2007.

[15] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *Image Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1178–1188, 2008.

[16] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014*, pp. 94–108. Springer, 2014.

[17] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[18] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[19] Gary B. Huang Erik Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[20] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.