FACIAL POINT DETECTION USING CONVOLUTIONAL NEURAL NETWORK TRANSFERRED FROM A HETEROGENEOUS TASK

*Takayoshi Yamashita** *Taro Watasue*** *Yuji Yamauchi** *Hironobu Fujiyoshi**

*Chubu University, **Tome R&D
1200, Matsumoto-cho, Kasugai, AICHI

ABSTRACT

We present a novel training approach that uses convolutional neural network for facial part detection. In the proposed training procedure, we use the parameters of a network obtained for a heterogeneous task as the initial parameters of the network for a target task. We employ a convolutional neural network for facial part labeling in the heterogeneous task, and then transfer the trained network so as to provide initial parameters of the network for facial point detection. This transfer of network is advantageous in the training for a target task in that 1) the network obtains representation kernels for extraction of facial part regions and 2) the network reduces detection errors at distant positions. The performance of the proposed method applied to BioID and Labeled Face Parts in the Wild datasets is comparable to that of state-of-the art methods. In addition, since our network structure is simple, processing takes approximately 3ms for one face on a standard CPU.

Index Terms— convolutional neural network, heterogeneous task, facial point, face labeling

1. INTRODUCTION

The detection of facial points is an important area of active research in computer vision, and is essential preprocessing for applications such as facial identification, facial expression estimation and facial analysis. These applications require the accurate detection of facial feature points even if facial images are taken with various poses, lighting conditions, expressions and occlusions.

Researchers have tackled the facial parts detection under these difficult conditions employing several approaches that can be divided two categories: classification methods[1][11][21] and direct prediction methods[2][3][5][13][17][19]. Classification methods extract candidate regions using local sliding windows. Both true facial points and similar regions of the face or background are detected as candidates. The optimal points are estimated from these candidates with shape constraints [1][13][21]. The prediction methods employ a regressor to detect the facial points from the whole face region without scanning. In this approach, the positions of facial points are iteratively updated until convergence is achieved. Recently, classification and prediction methods have been combined in a coarse-to-fine framework to improve accuracy. In this framework, the initial positions of facial points are first predicted and fine positions of facial points are then estimated [17][20][15]. Although the initial positions are critical in detecting facial points, a mean shape or a shape taken from training images, which can greatly differ from the test facial image, is used to provide the initial positions. In addition, many approaches face the issue of deciding what kind of feature representation to employ. Together with shape information, appearance information is important in detecting facial parts and yet it is underspecified.

In this work, we present a method of simultaneously resolving the two above problems in the detection of facial points. Our method is based on transferred convolutional neural network that is pre-trained for a heterogeneous task. Conventional pre-training is layer-wise unsupervised learning that obtains initial parameters that are efficient for subsequent learning. The initial parameters are then updated iteratively by supervised learning. The same dataset is used in the two training stages for the homogeneous task. In our proposed method, the network is trained twice with different datasets for the heterogeneous task. First, the network is pre-trained for the labeling of facial parts to extract an important facial region. This means that efficient feature extraction is automatically performed by this pre-training from a raw image without any knowledge. The pre-trained network is then transferred as initial network and updated using the facial point dataset.

2. RELATED WORKS

Facial point detection is important preprocessing for face recognition and facial analysis. Active shape models and the active appearance model, which model the holistic appearance or shape, are representative methods applied in early works[3][4]. These methods are limited in finding facial part regions when there are large variations in appearance.

The shape constraint is one solution to reducing the effect of facial variations[6][16] . These constraints provide a robust method to variation of facial expression. Belhumerur et al.

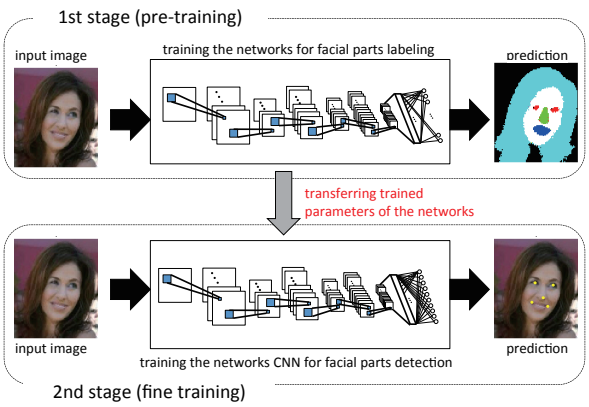

Fig. 1. The training frameworks. In the first stage, network is trained for facial feature labeling. The trained network is transferred as the initial parameters of the network for facial point detection.

proposed a Bayesian model combining the outputs of local detectors with a consensus of a non-parametric global model for part locations[1]. The model provides high accuracy but requires high-resolution images.

Regression-based methods are at the forefront of research. Dantone et al. detected facial points using regression forests of each face pose[5]. Valstar et al. employed support vector regression and a conditional Markov random field to obtain global consistency[19]. Cao et al. proposed a method based on the regression of random ferns that receive the whole face region as input[2]. However, these methods either lack flexibility in pose variation or incorrectly detect the frames of eyeglasses.

Deep convolutional neural network (CNN) [12] have high performance similar to that of human experts in object recognition[10] and object localization[9]. Krizhevsky applied CNN to an object recognition benchmark to classify 1,000 different classes and achieved good performance[10]. The advantage of CNN is that it is able to extract complex and suitable features for the task. It can reduce the burden of designing features since the entire system is trained from raw pixels. Sun et al. proposed a method based on CNN that cascade from the whole facial region to local regions[18]. Although they achieved state-of-the art performance, there method has a complex structure and is prone to incorrect detection in the presence of accessories such as eyeglasses.

3. TRANSFERRED CNN

Our proposed method has two training stages as shown in Fig. 1. In the first stage, we train the network to predict a facial part label for each pixel. The trained network is transferred to the second training stage as initial parameters. The last layer of the transferred network replaces the layer that predicts facial points. The parameters of network are updated iteratively with a facial point dataset. In the prediction phase, we use only the network of facial point detection that are trained in the second stage. Since the pre-trained network is transferred, the network structure without the last layer is the same.

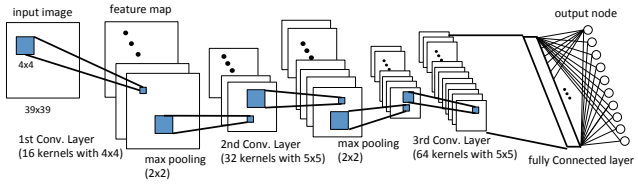

Fig. 2. Structure of CNN consisting of seven layers: three convolutional layers, three pooling layers and one fully connected layer. The fully connected layer has 1521 and 10 nodes as output for facial part labeling and facial part detection, respectively.

3.1. pre-training for facial parts labeling

To extract efficient initial parameters from a relative region of facial points explicitly, we train the network for facial part labeling in the first stage. The network have three convolutional layers with one max-pooling layer, and one fully connected layer. Maxout[7] and sigmoid functions are employed for activation in the convolutional layers and fully connection layer respectively. Maxout selects the maximum convolution value from certain neighboring kernels and has powerful representation ability. The output nodes correspond to pixels of the input image. We define six facial labels: eye, nose, mouth, skin, hair, eyeglasses and background. The network outputs a facial part label for each pixel. Random values are given as initial parameters of network, and these values are then updated by back propagation[14]. All parameters $\boldsymbol{W}$ are updated by back propagation iteratively with mini-batch. First, the random values are given as initial parameters $\boldsymbol{W_0}$. The updated parameters $\boldsymbol{W_{t+1}}$ in iteration $t+1$ are defined by Equation (1).

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \epsilon_{t+1}\frac{\partial L}{\partial \boldsymbol{W}} \tag{1}$$

The update efficiency decreases as the update time t. Binary cross entropy is employed for the cost function $L(w)$ as shown by Equation (2).

$$L(\boldsymbol{W}) = -\sum_{n=1}^{N}\{y_n \ln y' + (1-y_n)\ln(1-y'_n)\} \tag{2}$$

N is the number of training samples, y_n denotes label data and y'_n is the corresponding value on the output node for the n th training sample.

3.2. fine training for facial points detection

The parameters of network for facial part labeling are transferred as initial parameters of the network for facial point de-

tection. While the convolutional layers have the same structure as the network of facial part labeling, the fully connected layer is changed for regression of facial part detection. The x is the number of output nodes and y is coordinates of facial points. In the case of our facial point detection, the network have 10 output nodes since we define five facial points: the center of the left eye, center of the right eye, nose, left corner of the mouth and right corner of the mouth. The parameters of network are updated iteratively in the same manner as pre-training for facial parts labeling by back propagation. We applied the L2 norm to the loss function for the network of facial point detection.

By transferring the network of facial part labeling, the network of facial point detection are trained to prevent the occurrence of a local minimum like that corresponding to the frames of eyeglasses or shadow over the eye or mouth.

4. IMPLEMENTATION DETAIL

Figure 2 illustrate the network architecture of each stage. In the designed network architecture, the input is a gray image of 39×39 pixels. Even given a gray image, the network is able to label each pixel with a facial part in the first training stage. The first convolutional layer has 16 kernels of 4×4, and the second and third convolutional layers have 32 and 64 kernels of 5×5, respectively. The maxout function is employed as an activation function and 2×2 max pooling follows the first and second convolutional layers. The maxout function selects a maximum value from the neighboring two feature maps, and as a result, the number of feature maps is half the number of kernels. The output node of the third layer is treated as the input node of the fully connected layer. The fully connected layer has 39×39 nodes for output nodes. There are a total of approximately 7 million parameters of the network of facial labeling.

All parameters except those in the fully connected layer are transferred as initial parameters of the network of facial point detection. The fully connected layer is replenished as a new layer. This fully connected layer has 10 output nodes that correspond to the coordinates of each facial point. There are a total of approximately 60,000 parameters of the network of facial point detection.

Input images are augmented by scaling, translation and small rotation to obtain the robustness of the proposed method. We generate approximately 10 augmented images from a original image. These augmented images are divided into subsets and presented as a mini-batch during the training phase. We set the size of the mini-batch to 10 and the updating parameters to 20,000 and 0.006 respectively. We train the network on a NVIDIA GTX780 4-GB GPU.

	structure	BioID	LFPW
without Transferring	# of Conv. : 3 (8-16-32) # of Full : 1 (10)	0.051	0.065
	# of Conv. :3 (16-32-64) # of Full : 1 (10)	0.049	0.053
	# of Conv. :3(32-64-128) # of Full : 1 (10)	0.059	0.072
	# of Conv. :3 (16-32-64) # of Full : 2 (200-10)	0.124	0.142
with Transferring	# of Conv. : 3 (8-16-32) # of Full : 1 (10)	0.038	0.050
	# of Conv. :3 (16-32-64) # of Full : 1 (10)	**0.025**	**0.032**
	# of Conv. :3(32-64-128) # of Full : 1 (10)	0.037	0.047
	# of Conv. :3 (16-32-64) # of Full : 2 (200-10)	0.044	0.053

Table 1. Normalized error rate of eight network structures applied to two datasets. The error rate is the mean value for five facial points.

5. EXPERIMENTS

In experiments, we evaluated the CNN architectures to demonstrate the efficiency of the proposed method and compared the efficiency with that of several state-of-the art methods. In both experiments, we used the same training dataset as [18], which is the Labeled Face in the Wild (LFW) dataset available on the Internet[8]. We prepared 1,000,000 images from 10,000 images for training via augmentation with scaling, translation and rotation. In the training of network for facial part labeling, we first prepared the annotation of the facial labels for each pixel. We manually annotated facial labels to 4,000 images of LFW dataset. We assigned one of eight classes – eye, mouth, nose, skin, hair, eyeglasses, accessory and background — to each pixel. These annotation data will be publicly available from our website. We augmented 3,000 images of the images to produce 30,000 images for training of facial part labeling.

We evaluated detection methods applied to two major public datasets, BioID and the Labeled Face Parts in the Wild (LFPW), as test datasets. BioID contains facial images collected in the laboratory while LFPW contains facial images from the Internet. BioID includes 1,521 frontal facial images of 23 subjects with variations in illumination and expression. LFPW contains 1,432 facial images that were taken with large variations in illumination, facial pose, expression, and partial occlusion. The LFPW website allows the download of 1,132 training images and 300 test images. However, since some links on the website are no longer valid, we evaluated 249 images that we could download from original links.

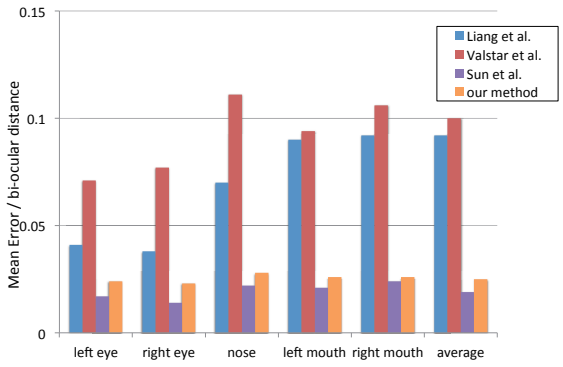

Fig. 3. Comparison of our method with state-of-the art methods applied to BioID.

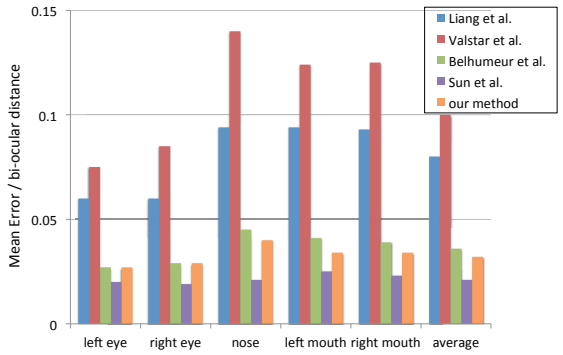

Fig. 4. Comparison of our method with state-of-the-art methods applied to FLPW.

We compared the proposed method with state-of-the art methods representing different approaches: a component-based discriminative search, boosted regression with Markov network and the use of cascaded deep convolutional network. We evaluated the normalized detection error by the biocular distance[1].

5.1. Comparison of network structures

Table 1 illustrates the performance of each network structure applied to the BioID and LFW datasets. We trained eight different structures that differ according to the number of layers and kernels. The first four structures are based on the conventional approach of training directly using a facial point dataset. The remaining four structures are trained for facial part labeling initially, and the trained network are then transferred to the training of network for facial part detection as initial parameters. The third structure with transferred network achieves average error rates of 0.059 and 0.072 for the two datasets. The transferred CNN that has three convolutional layers (with 16, 32, and 64 kernels in each layer) and one fully connected layer achieved the best performance for both datasets. The proposed method can reduce the error rate through transferring CNN.

Fig. 5. Facial point detection result for the LFPW dataset.

5.2. comparison

Figures 3 and 4 compare the performance of the proposed method with that of recent state-of-the art methods. The evaluation dataset is the same as in the above experiment. The results of comparison methods are cited from original papers. Since the condition in BioID is moderation, the error rate is lower than that for the LFPW dataset. Our method performs better than conventional classification and regression methods. The method proposed by Sun is also based on conventional network and employs a cascaded structure with several networks. Even though our method has single convolutional network, the results obtained using our method are similar to those of Sun's method. The average error rates of Sun's method are 0.019 and 0.021 for BioID and LFPW, respectively, while our results are 0.025 and 0.032. This difference reaches sub-pixel order when the size of the face is 39×39 pixels. We show resultant images of each dataset in Figure 5. Our method detects the facial points well for various conditions of the facial pose, expression and illumination. The processing speed of our method on C implementation takes 3 ms to one image on 3.4GHz CPU. This speed is 40 times faster than Sun's method, since the network structure is quite simple.

6. CONCLUSION

We described a novel training method of facial point detection based on convolutional neural network. Our primary contribution was to transfer the network from a heterogeneous task so as to provide the initial parameters in the target task. The transferred convolutional network can train representative kernels to extract facial part regions. As a result, the network do not fail in facial point detection. We demonstrated that the performance of the proposed method is comparable to that of state-of-the art methods applied to BioID and LFPW datasets.

7. REFERENCES

[1] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using an consensus of exemplars. In CVPR, 2011.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In CVPR, 2012.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In ECCV, 1998.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models -their training and application. Computer vision and image understanding, 1995.

[5] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. In CVPR, 2012.

[6] M. Everingham, J. Sivic, and A. Zisserman. Hello!, may name is ... buffy - automatic naming of characters in tv video. In BMVC, 2006.

[7] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. arXiv preprint arXiv:1302.4389, 2013.

[8] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces int the wild: A database for studying face recognition in unconstrained environments. Technical Report07-49, University of Massachusetts, 2007.

[9] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. Wahat is the best multi-stage architecture for object recognition?. In ICCV, 2009.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. magenet classification with deep convolutional neural networks. In NIPS, 2012.

[11] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component based discriminative search. In ECCVV, 2008.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P.Haffner. Gradient-based learning applied to document recognition. In IEEE, 1998.

[13] X. Liu. Generic face alignment using boosted appearance model. In CVPR, 2007.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In Proc. Explorations in the Microstructures of Cognition, 1986.

[15] S, Ren, X. Cao, Y. Wei, J. Sun. Face Alignment at 30000 FPS via Regressing Local Binary Features. In CVPR, 2014.

[16] G.Roig, X. Boix, F. De la Torre, J. Serrat, and C. Vilella. Hierarchical crf with product label spaces for parts-based models. In FG, pp.657–271, 2011.

[17] P. Sauer, T. Cootes, and C. Taylor. Accurate regression procedures for active appearance models. In BMVC, 2011.

[18] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In CVPR, 2013.

[19] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using regression and graph models. In CVPR, 2010.

[20] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free Facial Landmark Fitting via Optimized Parts Mixture and Cascaded Deformable Shape Model. In ICCV, 2013.

[21] X. Zhu, and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012.