M.S. THESIS

Deep Multi-Dataset Multi-Domain Multi-Task Frame Works for Facial Expression Recognition, Age and Gender Estimation

얼굴 표정 인식, 나이 및 성별 추정을 위한 다중 데이터셋 다중 도메인 다중작업 네트워크

BY

Sepidehsadat Hosseini
AUGUST 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Deep Multi-Dataset Multi-Domain Multi-Task Frame Works for Facial Expression Recognition, Age and Gender Estimation

얼굴 표정 인식, 나이 및 성별 추정을 위한 다중 데이터셋 다중 도메인 다중작업 네트워크

BY

Sepidehsadat Hosseini
AUGUST 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Deep Multi-Dataset Multi-Domain Multi-Task Frame Works for Facial Expression Recognition, Age and Gender Estimation

얼굴 표정 인식, 나이 및 성별 추정을 위한 다중 데이터셋 다중 도메인 다중작업 네트워크

지도교수 Prof. Nam Ik Cho

이 논문을 공학석사 학위논문으로 제출함

2019년 8월

서울대학교 대학원

전기 컴퓨터 공학부

세피데사닷

세피데사닷의 공학석사 학위 논문을 인준함

2019년 8월

위 원 장: _____
부위원장: _____
위 원: _____

Abstract

The convolutional neural network (CNN) works very well in many computer vision tasks including the face-related problems. However, in the case of age estimation and facial expression recognition (FER), the accuracy provided by the CNN is still not good enough to be used for the real-world problems. It seems that the CNN does not well find the subtle differences in thickness and amount of wrinkles on the face, which are the essential features for the age estimation and FER. Also, the face images in the real world have many variations due to the face rotation and illumination, where the CNN is not robust in finding the rotated objects when not every possible variation is in the training data. Moreover, The Multi Task Learning (MTL) Based based methods can be much helpful to achieve the real-time visual understanding of a dynamic scene, as they are able to perform several different perceptual tasks simultaneously and efficiently. In the exemplary MTL methods, we need to consider constructing a dataset that contains all the labels for different tasks together. However, as the target task becomes multi-faceted and more complicated, sometimes unduly large dataset with stronger labels is required. Hence, the cost of generating desired labeled data for complicated learning tasks is often an obstacle, especially for multi-task learning. Therefore, first to alleviate these problems, we first propose few methods in order to improve single task baseline performance using gabor filters and Capsule Based Networks , Then We propose a new semi-supervised learning method on face-related tasks based on Multi-Task Learning (MTL) and data distillation.

keywords: Face-related Tasks, Capsule Net, Data Distillation, Multi Task Learning, Domain Adaptation

student number: 2017-26727

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Motivation

Researchers have applied convolutional neural networks (CNNs) to many image processing and computer vision tasks, including the face-related problems that we focus on in this paper. For example, the CNNs in [1, 2] are shown to provide better face detection performance than the conventional methods that use hand-crafted features [3, 4]. Recent researches on age estimation indicate that the CNN-based techniques [5, 6] also yield more accurate results than the methods based on the hand-crafted features, specifically the bio-inspired feature (BIF) [7] which is one of the best non-CNN approaches. In most CNN-based computer vision applications, we usually feed the CNN with raw images (not the features) as the input. This is based on the belief that the CNNs learn and extract the right features through the training with the image input. However, in the face-related problems, we need to tell the subtle differences of facial features such as the wrinkle, and also the differences in the positional relationship of facial features that the plain CNNs cannot well detect and define. Hence we need more efforts other than using plain CNNs with raw image input.

To be precise, the most important features in estimating the age are the amount and thickness of wrinkles, and the sizes and relative distances of facial landmarks (eyes,

eyebrow, ears, nose, mouth, etc), where it seems that the plain CNNs cannot well find the subtle features. The other problem with the CNN is the use of max pooling in the network. It was originally intended to reduce the data size and positional invariance, but the spatial relationships between higher level features are lost due to the pooling. Also, the CNNs do not well deal with different viewpoints, or they need a large amount of data augmentation for the view-invariance.

In this thesis, we attempt to alleviate the above-stated problems in conducting the face-related tasks. First, we show that feeding useful hand-crafted features to the CNN, along with the input image, can enhance the performance of CNN for the age/gender estimation and FER. In other words, we stimulate the CNN with the relevant hand-crafted features, which helps the CNN to find the right features at the earlier layers and thus increases the performance. Moreover, based on the Capsule Network (CapsNet) [8] which is intended to alleviate the problems of the CNN-based architectures (weakness in view-point change and loss of spatial relationship of features), we further increase the accuracy of age/gender estimation and FER. Then, we use hand-crafted features along with the CapsNet-based architecture, which is shown to outperform the baseline CapsNets.

Then as we mentioned, in order to achieve more generalized and realistic information, we can use multi-task networks, however the cost of generating desired labeled data for complicated multi tasks learning network is too high. Therefore, studies on semi/self/omni-supervised learning are getting attention recently because they can obviate such strong labeling. In the most semi-supervised learning methods, they exploits part of annotated data and considers the rest as unlabeled [70, 71]. Recently a new regime of semi-supervised learning has been proposed called as omni-supervised learning [64]. In the omni-supervised learning, the learner uses as much labeled data as possible and also uses an unlimited amount of unannotated data from other sources.

In this thesis, we propose a data distillation framework on weakly labeled datasets to help to improve the multi-task learning on facial expression recognition. Previous

works on distillation adopted omni-supervised learning methods [64] which used unlabeled auxiliary datasets. However, we argue that instead of feeding the network with unlabeled images for providing a new target labeled dataset, we can use datasets from other related tasks as weakly labeled images. By doing so, we can train the network in the manner of multi-task learning (MTL) and then use the trained network to produce the target labels for the related tasks' datasets. Then, similar to [52, 64], we retrain the network in a single task manner with the union of the original and the newly labeled datasets. By doing so, we can benefit from making the network familiar with the features of the new datasets and having a more powerful teacher for data distillation.

Moreover, In the exemplary MTL methods, we need to consider constructing a dataset that contains all the labels for different tasks together. Without such a dataset, training the multi-task network in a common approach will result in a negative effect due to the cross-dataset distribution shift. To the best of our knowledge, the first work which mentioned this problem is StarGAN [49] proposed by Choi *et al.*. Their model can simultaneously be trained on different datasets by alternating between different datasets. However, the alternating scheme still has the cross-dataset distribution shift problem, and the network cannot be applied to datasets with different domains. Recently, Guosheng Hu *et al.* [55] addressed this issue by proposing the trace norm-based knowledge sharing. In their method, multiple networks, one for each task, are stacked horizontally together to form a one-order higher tensor. Then, by using a tensor trace norm regularizer, they share knowledge between these networks. In comparison with [55], our method is simpler, easier to implement, and more efficient in both aspects of memory and computation.

1.2 Background

1.2.1 Age and Gender Estimation

Aging depends on several factors such as living habit, race, genetics, etc. Hence, predicting a person's age from a single image is one of the hardest tasks both for human and machines. Researches on age estimation are mainly following two paths: designing age-related features [7, 9] or using the CNN. Researches without using the CNN are well summarized in Zafeiriou et al.'s survey [10]. Recent works are mostly based on the CNN, for examples, Levi and Hassner's work [5] was the first to adopt the CNN for age/gender estimation and Xing et al. [6] considered the influence of race and gender by proposing a multi-task network.

1.2.2 Facial Expression Recognition (FER)

The FER is a relatively complicated task among many face-related works. Since the FER plays an important role in human-machine interaction, many researches have also been conducted on this subject. Li and Deng [61] published a survey on the deep facial expression recognition methods. Recently. For some examples of conventional methods, Georgescu et al. used the support vector machine (SVM) to improve the Bag of Visual words (BOW) approach [11], and Hassani et al. used the advantage of facial landmarks along with CNNs [12]. More recent studies are focused on using the CNNs for the FER [13, 14, 15, 16]. acial Expression Recognition (FER) has also attained increasing attention recently. Yang *et al.* [78] proposed to recognize facial expressions by extracting information of the expressive component through a de-expression learning procedure, called De-expression Residue Learning (DeRL). Zhang *et al.* [82] proposed joint pose and expression modeling by disentangling the expression and pose from the facial images and produce images with arbitrary expressions and poses using a new discriminator and a content-similarity loss for generative adversarial networks. Zeng *et al.* [81] addressed the inconsistency between FER datasets for the first time by

proposing an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) framework to train a FER model from multiple inconsistently labeled datasets and large-scale unlabeled data. Our method can be considered a generalization of this work because we can use datasets with inconsistent labels instead of datasets with different-task labels.

1.2.3 Capsule networks (CapsNet)

Hinton et al. [18] proposed a new method for robust unsupervised learning called capsules. The capsules are the group of neurons to recognize the presence of a visual entity within a limited range of viewing condition and deformation. A group of capsules makes a capsule-layer, where the outputs of the capsule-layers are vectors instead of scalars. The length of the capsule expresses the probability of the entity being present, and the orientation of capsule represents the abstraction of parameters of entity. Afterward, Sabour et al. [19] made capsules feasible as CapsNet which uses a routing-by-agreement mechanism. In this mechanism, an active-capsule at each level (layer L) actives capsules by using a transformation matrix to predict the presence of parameters of capsules in the higher level (Layer $L + 1$), and the higher level capsules become active if several of those predictions agree. Later on, Hinton et al. [8] proposed CapsNet with expectation maximization (EM) routing structure that uses matrix capsules, which produces a logistic unit (activation) and a 4×4 matrix (pose matrix) to represent the presence of a visual entity and relationship between that entity and the pose respectively.

1.2.4 Semi-Supervised Learning.

Zhu*et al.* [84] and Sheikhpour *et al.* [71] have done comprehensive surveys on semi-supervised learning methods. The first trial on self semi-supervised learning was based on the soft self-training technique [70], which is to predict labels of unannotated data. Then those labels are used to train itself, which is known as one of the simplest and commonly used approaches in semi-supervised learning. Recently, many approaches

attempt designing deep learning based semi-supervised frameworks [58, 64, 65]. Rasmus *et al.* [65] proposed the Ladder network-based method, by exploiting unsupervised auxiliary tasks. Laine *et al.*[58] annotate unlabeled data using the outputs of the network-in-training under different conditions such as regularization input augmentation. In Omni-supervised method [64], they use knowledge distillation from larger data, in the other word their model generates annotations on unlabeled data using a model trained on large amounts of labeled data. Then, they retrain the model using the extra generated annotations.

1.2.5 Multi-Task Learning.

Multi-task learning has demonstrated performance improvement in several computer vision applications such as facial landmark detection [83] and human pose estimation [62]. The primary intuition behind Multi-Task Learning (MTL) is how humans apply their knowledge and skills obtained from other tasks on more complicated tasks. There are different methods to exploit MTL: joint learning, parallel multi-task learning with auxiliary tasks, and continual learning are a few examples of MTL based methods. The parallel multi-task based methods integrated different tasks contemporaneously, which has been widely deployed in face-related tasks [55, 75].

1.2.6 Knowledge and data distillation.

There are a large number of researches attempt to transfer knowledge from a teacher model to a student model. Romero *et al.* [68] proposed FitNets, a two-stage strategy to train networks by providing *hint* from the teacher middle layers. Knowledge Distillation (KD) proposed by Hinton *et al.* [54] leverage the predictions of a larger model as the *soft target* to better training of a smaller model. After that, Chen *et al.* [46] improved the efficiency and the accuracy of an object detector by transferring the knowledge from a powerful teacher in case of model architecture or the input data resolution to a weaker student. Zagoruyko *et al.* [80] proposed several ways to transfer

the attention from a teacher network to a student. Polino *et al.* [63] proposed quantized distillation to compress a network in terms of depth by using knowledge distillation. Furlanello *et al.* [52] used knowledge distillation on a student the same as the teacher to improve the performance of the networks by teaching selves.

Inspired by knowledge distillation, Radosavovic *et al.* [64] proposed data distillation to tackle omni-supervised learning. They generate annotations for unlabeled data by using a trained model on a labeled dataset and then retrain the model on the union of these two datasets to improve the accuracy. There are also other works trying to use unlabeled data to retrain the model [47, 58, 60, 79]. Gupta *et al.*[53] proposed a method to transfer supervision between different modalities which needs unlabeled paired images. Laine and Aila [58] proposed to use ensemble from different checkpoints with different regularizations and input augmentations.

1.2.7 Domain Adaptation.

Saenko *et al.* [69] was one of the first researchers who proposed a method to solve the domain shift problem. More recent works are based on deep neural network aiming to align features by minimizing domain gaps using some distance function [66, 74]. In these methods, domain discriminator trains to distinguish different domains while the generator tries to fool discriminator through the learning of more general representation and features.

1.3 Datasets

We consider 7 facial related datasets to evaluate our method.

CK+ [41] is one of the constrained datasets widely used for FER. It contains 593 video sequences from 123 persons. The sequences start from neutral faces and shift to one of anger, contempt, disgust, fear, happiness, sadness, and surprise expressions peak. Among these 593 sequences, only 327 sequences from 118 persons are labeled to those seven expressions.

Oulu Casia [44] contains 2,880 sequences of 180 subjects, in six different expressions (anger, disgust, fear, happiness, sadness, and surprise) per subject. Similar to CK+ each sequence starts from a neutral face and gradually shows the expression. Following other researches, we also use only images under visible light and strong illumination condition.

FER2013 [40] is annotated with seven basic facial expressions (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, and 6=Neutral), which contains about 32K images, 28.5K for training and 3.5K for the test. All pictures in this dataset are collected automatically by the Google image search API which is one of the frequently used unconstrained datasets.

MORPHII [32] is one of the most popular large-scale age estimation datasets created by the Face Aging Group at the University of North Carolina. It contains 55,134 images of 13,000 subjects with about three images per subject, age ranging from 16 to 77 year. The images in this dataset are mainly frontal.

Adience [29] compared to MORPHII which contains frontal and constrained images, has been captured from Flicker.com albums. Hence, they are totally unconstrained and no manual filtering has been applied, which makes them a good representation of the real world. It consists of 26K facial images of 2,284 identities.

Gallagher [30] consists of images from flickr.com, including pictures with large variations in pose, appearance, lighting condition, unusual facial expressions, etc. It has 5K images with 28K labeled faces, divided into 7 classes (0-2, 3-7,8-12, 13-19, 20-36,

37-65, 66+).

FG-Net [33] which contains 1002 images of 82 subjects (age-range from 0 to 69 and has more frontal pictures, and there are several pictures of the same person in different years, which makes the dataset a suitable benchmark for age regression.

Chapter 2

GF-CapsNet: Using Gabor Jet and Capsule Networks for Face-Related Tasks

2.1 Feeding CNN with Hand-Crafted Features

2.1.1 Preparation of Input

Nobel prize winners Hubel and Wiesel discovered that there are simple cells in the primary visual cortex, where its receptive field is divided into subregions which are the layers covering the whole field [20]. Petkov [21] proposed the Gabor filter, as a suitable approximation of mammal's visual cortex receptive field. The 2D Gabor filter is a Gaussian kernel function adjusted by a sinusoidal wave, consisting of both imaginary and real parts, where the real part can be described as:

$$g_{\lambda,\theta,\sigma,\gamma}(x,y) = \exp\left(-\frac{x' + \gamma y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \phi\right) \tag{2.1}$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, and λ, θ, ϕ, γ and σ are the wavelength of the real part of Gabor filter kernel, the orientation of the normal to the stripes of function, phase offset, spatial ratio and standard deviation of the Gaussian envelope representatives respectively. Fig. 2.1 is an example of Gabor filter responses to a face image, which shows that they find the textures that correspond to the given θ

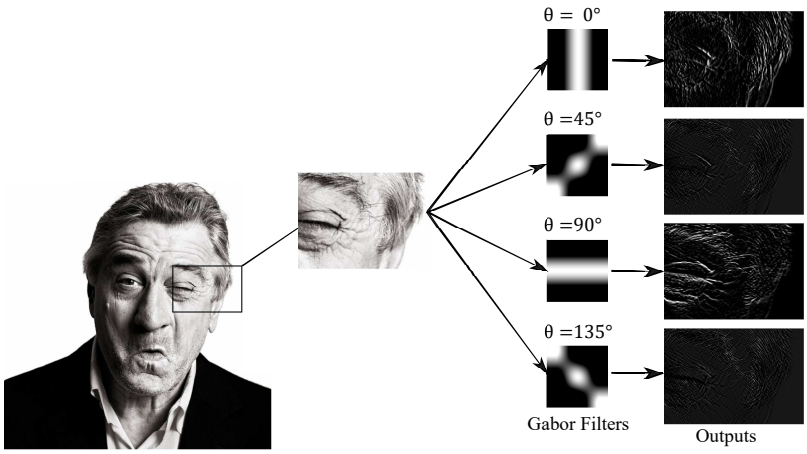


Figure 2.1: Demonstration of Gabor filter bank and their responses with kernel size = 3 applied to an image. Responses for four orientations ($\theta = 0, \pi/4, \pi/2, 3\pi/4$) are shown.

very well. Hence, the Gabor filter responses have been used in the applications where the orientational textures play an important role such as fingerprint recognition [22], face detection [23], facial expression recognition [24], and age/gender estimation [7]. A recent research [25] also showed that using Gabor responses as the input can increase the performance of CNN.

However, only a single λ was used in [25], which means that we cannot fully observe the different depths of wrinkles. Hence, in this paper, we use the Gabor jet proposed in [7], which is a set of the multi-scale version of Gabor filters with different spatial scales and orientations. In the other words, we use 32 Gabor filters with $\lambda = \{2.3, 2.5, 3, 3.8\}$, $\theta = \{0, \pi/4, \pi/2, 3\pi/4\}$ and $\pi = \{0, \pi/2\}$. From the extensive experiments, we found that the optimal σ in different cases are highly dependent on λ, specifically $\sigma = \lambda/2$. Also, we fix $\gamma = 0.1$ in all of our experiments.

For feeding the Gabor responses to the network, we extract several Gabor filter responses and concatenate them with the input image, which forms a tensor input like a multi-channel image. Let N_f be the number of Gabor filters, and let F_g^k be the

response of the k-th Gabor filter. Normally, we may just concatenate the input image (a gray input image of size $W \times H$) and N_f responses as $W \times H \times (N_f + 1)$ tensor input to the CNN as illustrated in Fig. 2.2(a). On the other hand, we may consider fusing the

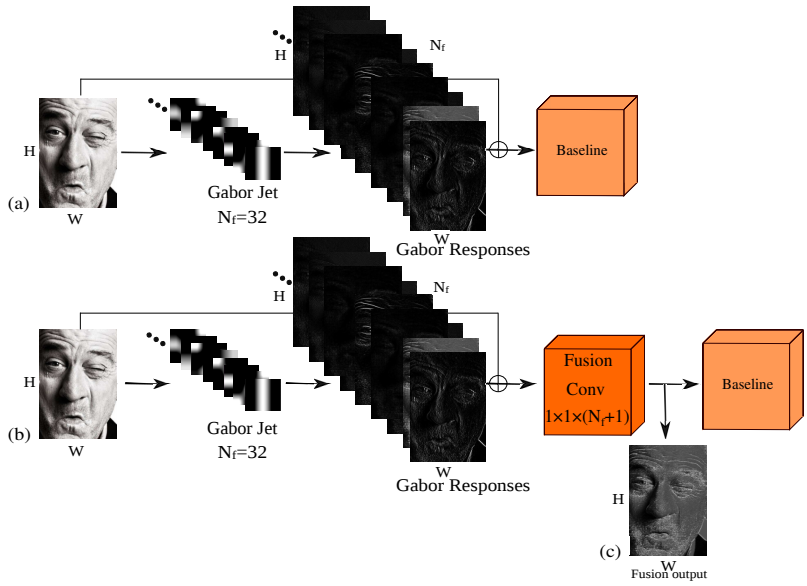


Figure 2.2: Illustration of two input feeding methods. (a) The tensor input is fed to the CNN. (b) The tensor input is fused to be an image and fed to the CNN. (c) An example of a fusion image which is the weighted sum of image and Gabor responses.

input and Gabor responses as a single input (matrix), and feed the matrix to the CNN as shown in Fig. 2.2(b). The Figure also shows that fusing the input image and Gabor responses can be interpreted as convolving the $W \times H \times (N_f + 1)$ tensor input with $1 \times 1 \times (N_f + 1)$ filter. If we denote the coefficients of this filter as $[w_1,\ w_2, \cdots, w_{N_f}]$ and w_{image} (w_k is multiplied to the k-th Gabor response and w_{image} is multiplied to the input image), then the fused input is represented as

$$F^{in} = w_{image} I + \sum_{k=1}^{N_f} w_k F_g^k \tag{2.2}$$

which is similar to the weighted fusion method in [6]. These weights are trained along with the rest of network parameters in the end-to-end manner. Fig. 2.2(c) is an example of fused input, which can be considered a "wrinkle-enhanced" image.

Both concatenation and fusion approaches inject the Gabor responses as the input to the CNN. From the extensive experiments, while the concatenation approach shows slight improvement compared to the baseline, the fusion approach in Fig. 2.2(b) shows much better performance than the baseline (about 8 %p increase in the case of age estimation using the network purposed in Levi [5] as a baseline, and also similar improvements when using the other networks as baselines). Also, it requires less number of parameters than the concatenation and almost the same amount of parameters as the baseline.

Analysis of feature maps from the network (shown in Fig. 4.1 which will be discussed later) shows that the wrinkle features and face shapes are more enhanced in our CNN than the conventional one that uses only the pixel values as the input. As a result, the accuracy of age/gender estimation is much improved compared to the state-of-the-art image-domain CNNs [5, 6]. Moreover, we test our approach on facial expression recognition and also obtain some gains over the existing CNN-based methods [13, 14, 15, 16]. In other tasks where some of the hand-crafted features are effective, we hope that feeding such features along with the image may bring better results.

2.1.2 Age and Gender Estimation using the Gabor Responses

The gender estimation is just a binary classification, while the age estimation is implemented as a classification or regression problem. In the case of age estimation as a classification problem (segmenting the age into several ranges), we apply our input fusion scheme in Fig. 2.2(b) to three different baselines. One of them is the most simple age estimation network similar to Levi [5] (Fig. 2.3), and the two others are VGG16 [27] and ResNet-101 [28]. For the gender estimation, in addition to using Levi (Fig. 2.3) and ResNet [28] as baseline, we also examine our method on VGG16-Hybrid network (Fig. 2.4) which estimates the gender, and use the gender-result for more accurate gender-specific age estimation. For training the VGG16-Hybrid net-

work, we first pre-train the gender network and each of the gender-specific networks separately on their specific data. Then, the network is finely tuned using the whole dataset.

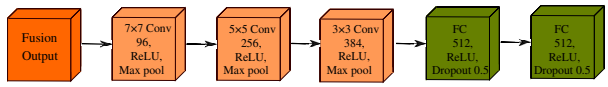


Figure 2.3: Baseline age classification network (Levi's network).

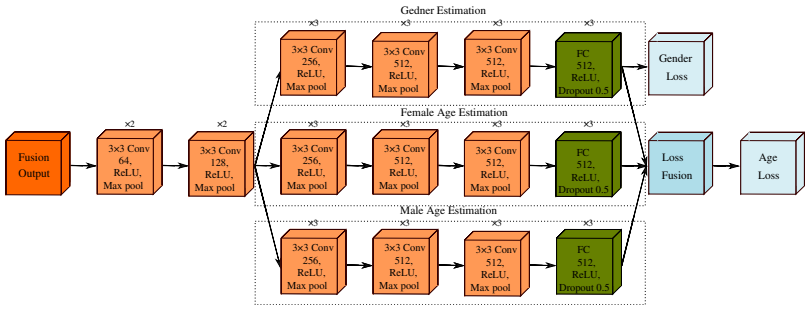


Figure 2.4: Baseline age classification network (VGG16-Hybrid).

Age estimation can also be implemented as a regression problem when we wish to tell a person's exact age, rather than as a classification problem which tells the range (class) of ages. We use two networks: one is the VGG16-Hybrid network in Fig. 2.4 and the other is the Resnet [28]. One of the main differences between the classification and regression problem is that they need different loss functions. For the classification problem, we use the softmax loss defined as:

$$L(x) = -\frac{1}{N}\sum_{i=1}^{N} Y_{iy_i} \log p_{iy_i} \tag{2.3}$$

where N is the number of classes, Y_{iy_i} is one-hot encoding of the sample's age label and p_{iy_i} is the y_i-th element of predicted probability vector for x_i. For the regression, we use Mean Absolute Error (MAE) as the loss function. To be precise, the MAE is defined as

$$L(x) = -\frac{1}{M}\sum_{i=1}^{M} |\hat{y}_i - y_i| \tag{2.4}$$

where M is the maximum age that we set, and $\hat{y}_i$ is the estimate of true age y_i.

2.2 GF-CapsNet

2.2.1 Modification of CapsNet

In the previous section, we showed that feeding Gabor features to the CNNs can increase their performances in face-related problems. However, the best performances shown in the tables do not still seem good enough to solve the real-world problems. Hence, we attempt to further increase the performance by using the recently developed CapsNet in this section.

As we mentioned before, there are some problems in using the CNN for face-related tasks such as age estimation and FER. First, CNNs are not good at finding the spatial relations of facial landmarks, and secondly, they are invariant to changes in viewpoints. On the other hand, the CapsNet can capture the parameters of the specific feature along with its likeliness. Hence, it can not only detect features but also learn and detect their variants. To construct a CapsNet-based age estimation architecture, we adopt the EM routing mechanism in [8]. This method employs the EM clustering technique to cluster the lower layer capsules in Gaussian distribution and create a part-whole relationship. We use the matrix capsule which detects the likeliness and 4×4 pose matrices which define the change of viewpoint of features. Also, in the CapsNet, there are 4×4 transformation matrices W between the capsules in the L-th layer and their parent capsules in the $(L+1)$-th layer. Then, the votes matrix is defined as the multiplication of the pose matrix with the transformation matrix as:

$$v_{ij} = M_i W_{ij} \tag{2.5}$$

where v_{ij} is a vote for a capsule j to be the parent of capsule i, M_i is a pose matrix for the capsule i, and W_{ij} is the transformation matrix between the capsules i and j. Then, by using the EM routing on these votes, the parent-children relation can be made.

According to the method in [8], the capsule j will be activated depending on the

activation function

$$\begin{aligned} a_j &= \text{sigmoid}(\lambda(b_j - \sum_h \text{cost}_j^h)) \\ &= \text{sigmoid}(\lambda(b_j - \sum_h \sum_i r_{ij}\text{cost}_{ij}^h)) \end{aligned} \tag{2.6}$$

where r_{ij} is the runtime assignment probability which shows the amount data of capsule i assigned to the capsule j, and h refers to h^{th} component of pose matrix. The b_j is related with the capsule j's mean and variance, which can be approximated through the optimization of cost function cost_{ij}, which is the cost for the capsule i in the L-th layer to activate the parent capsule j in the layer $L+1$. The pose matrix is generally modeled as Gaussian, and then the cost_{ij}^h is defined as

$$\text{cost}_{ij}^h = -ln\left(\frac{1}{\sqrt{2\pi(\sigma_j^h)^2}}\exp\left(-\frac{(v_{ij}^h - \mu_j^h)^2}{2(\sigma_j^h)2}\right)\right) \tag{2.7}$$

where μ_j and σ_j are capsule j's mean and variance respectively. Note that r_{ij}, μ_j^h, σ_j^h and a_j are computed using the EM routing whose main objective is to fit the data points to a Gaussian model. Further details of EM routing algorithm can be found in [8].

In summary, the baseline network that we use is the one proposed in [8], which has a simple convolutional layer at the head to extract the features, followed by three capsule layers. In our proposed CapsNet, we inject Gabor features along with the image to the network. Also, considering the complexity of face-related tasks, we add one more convolution layer at the head to extract more features as shown Fig. 2.5.

About the loss function for the training, the "spread loss" is defined in [8] which is expressed as

$$L_{spread} = \sum_{i \neq t}(max(0, m - (a_t - a_i)))^2 \tag{2.8}$$

where m is the margin which is initially 0.05 and linearly increased to 0.95, and a_t and a_i correspond to activation target and wrong class respectively. In our implementation for the classification, we use this spread loss function. However, in the case of age

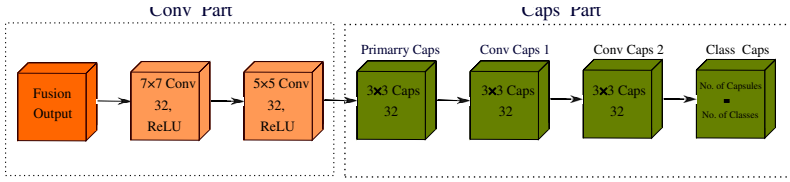


Figure 2.5: Our modified Capsule network.

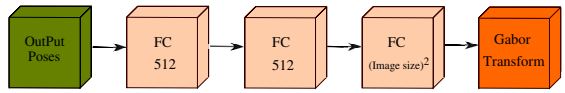


Figure 2.6: Reconstruction of input image using the decoder network.

regression, the spread loss does not differentiate between the wrong estimates and different values of errors. Hence, we add one more capsule after the final layer (Class Caps), and use Mean Absolute Error (MAE) instead of the spread loss.

As Sabur et al. suggested in [19], adding the reconstruction error to the total loss can improve the performance and acts as a regularization method. However, unlike the MNIST dataset used in [19], the face-related datasets are more complicated and reconstructing the whole image is hard and unnecessary. Hence, in the proposed reconstruction loss, we extract the Gabor features of the original and the reconstructed image which is obtained from the last layer of CapsNet (output poses). That is, we modify the loss function to

$$L = L_{spread\ or\ MAE} + r||G(I_{org}) - G(I_{rec})|| \tag{2.9}$$

where γ is the regularization scale and $G(I_{org})$ and $G(I_{rec})$ are Gabor features extracted from the original and reconstructed image respectively. The procedure of image reconstruction is illustrated in Fig. 2.6.

subsectionFacial Expression Recognition (FER)

In the FER experiments, we choose some state-of-the-art networks as the baselines and show they yield improved results when fed with fusion input. The first baseline is VGG-19 [27] which shows the best results on FER2013. We add one more drop out

after the last fully connected layer to decrease the overlapping as shown in Fig. 2.7. We also choose the zero-bias CNN+AD [15] shown in Fig. 2.8, which uses three convolutional layers followed by one fully connected layer.

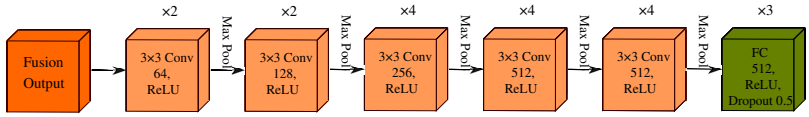


Figure 2.7: Illustration of GF-VGG network for facial expression recognition on FER2013.

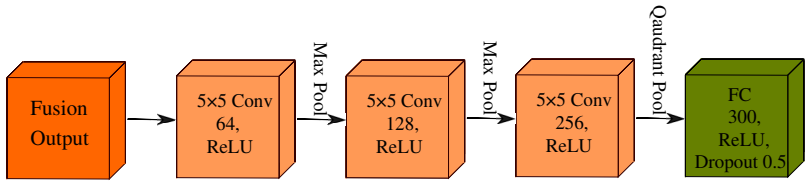


Figure 2.8: Illustration of GF-Zero-bias CNN+AD network for facial expression recognition on CK+.

Chapter 3

Distill-2MD-MTL: Data Distillation based on Multi-Dataset Multi-Domain Multi-Task Frame Work to Solve Face Related Tasks

3.1 MTL learning

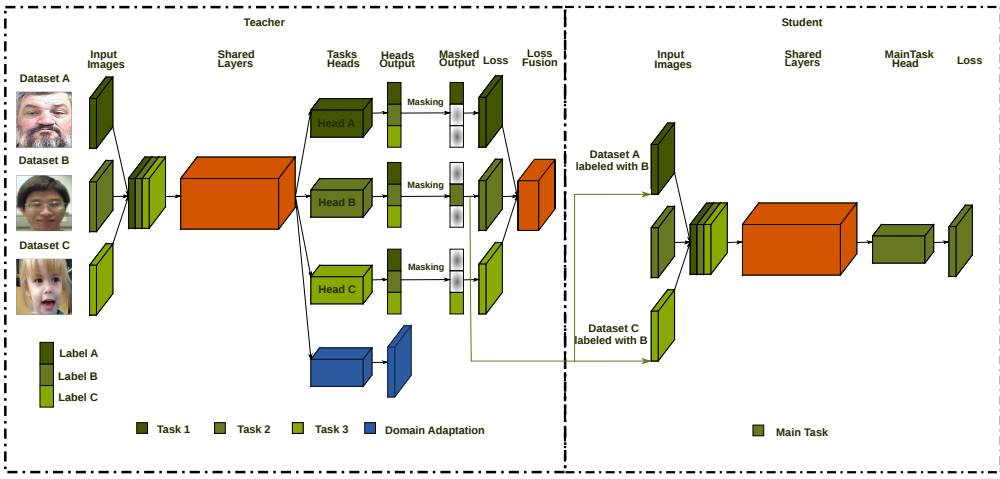


Figure 3.1: The proposed method. Right: the first step of training using the proposed 2MD-MTL network (teacher). Left: the second step of training using a simple single task network (student) with labels produced by the 2MD-MTL network (teacher).

Suppose we have t tasks T_i for $1 \leq i \leq t$, and d datasets D_j for $1 \leq j \leq d$ in which each dataset contains labels for a subset of the t tasks. Without loss of generality, suppose the target task is T_1 and at least one of the datasets contains the related labels for the target task. By defining a multi-task network N_m, we train N_m with the datasets D_j in a multi-dataset multi-domain multi-task (2MD-MTL) manner.

To be more clear, instead of training the network with alternating inputs from each dataset, we construct an input batch of size $b = t \times \hat{b}$ as a combination of $\hat{b}$ images from each task T_i. Therefore, by evaluation of the network N_m on the input batch, we will have a matrix L of size $b \times t$ related to the loss functions of the different tasks, in which cell $l_{i,j}$ means the loss value for the i-th image and the j-th task. Now, we construct a mask matrix M of the same size by putting $m_{i,j} = \alpha_j$, where α_j is equal to the coefficient of the loss due to the task T_j, if the i-th image contains the label for task T_j and 0 otherwise. Then, the final loss will be equal to the dot product of these two matrices. In other words, we use all the tasks parallelly in the network by considering only the valid loss values at the end.

The loss function in multi-task learning is generally defined as $L = \Sigma_i \omega_i L_i$, where L_i is a Loss function and ω_i is a scalar coefficient for the i-th task respectively. In most of the cases, it is challenging to find the best value for each ω_i which not only need huge efforts and extensive experiments but also decrease network generalization. We use the gradient normalization [48] to solve the loss balancing problem, which obviates the expensive time-consuming grid search for tuning the ω_is.

Figure 3.1 shows the proposed framework, where we use VGG-16 [72] network as the baseline, all tasks are sharing convolutional layers (5 convolutional blocks), and each task has its own Fully connected layers and also their own loss as their head.

While the features learned above on multiple tasks will be more general-purpose ones than those learned on a single task, there may be still a problem if the dataset domains are so different. For example, Figure 3.2 shows some images used in our experiments, where the images from MORPHII and Casia datasets are frontal images

Figure 3.2: Diverse-domains - diverse-tasks datasets

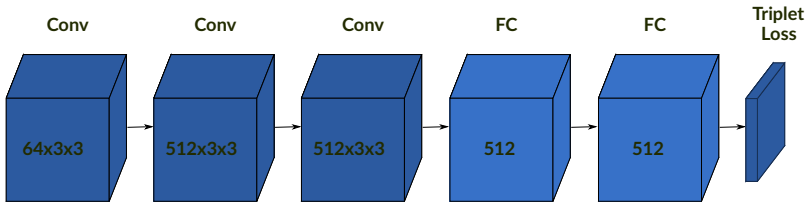


Figure 3.3: Discriminator Head.

while the photos from the Adience dataset are mainly wild. On the other hand, MORPHII subjects are photos of prisoners' who don't show that many emotions while Casia is emotion dataset. In order to minimize the domain gap between different datasets and also extract more generalized features, we use metric learning based discriminator.

In our proposed method, we add the discriminator head shown in Figure 3.3 after the shared layers. Then, we apply a triplet loss which aims to pull samples belonging to the same dataset into nearby points on a manifold surface and push samples from different datasets apart from each other. The labels of the training images for the discriminator's head can be easily provided by the dataset to which they belong (for example 0: age, 1: gender, and 2: emotion). Then, they are selected and formed into triplets as $T_i = (x^a, y^p, y^n)$, where x^a and y^p are the anchor and positive samples respectively which belong to one dataset and y^n is the negative sample which belongs to another. Then, we train the discriminator head to decrease the triplet loss (Eq. 3.1) and the rest of the network to minimize the total loss (Eq. 3.2) where N is the number of tasks and L_i is the loss function per task. In the other word, we train network in

the way that the discriminator's head is not able to distinguish between datasets, while still the other heads being able to extract informative feature for all the task, therefore shared layer features will be generalized on all tasks and domains.

$$L_{DA} = \Sigma_{t=1}^{T}[||x_t^a - y_t^b||_2^2 - ||x_t^a - y_t^a||_2^2 + 0.2] \tag{3.1}$$

$$L_{Total} = \Sigma_{i=1}^{N}\omega_i L_i - L_{DA} \tag{3.2}$$

Using the triplet loss for the discriminator's head can help us to overcome the class imbalance problem due to the different sizes of datasets. For example, the number of MORPHII images is one order of magnitude greater than the images in CK+. Therefore, without considering a solution for the class imbalance, the discriminator will be biased to MORPHII based on a large number of images in that class.

3.2 Data Distillation

Hinton *et al.* [54] proposed knowledge distillation (KD) in order to transfer the knowledge from a cumbersome teacher model to a smaller student model. They use the class probabilities predicted by the teacher model as a *soft target* to guide the student model. Furlanello *et al.*in born-again neural networks [52] show that we can also adopt student network architecture as the teacher in order to improve the model by guiding itself. Radosavovic *et al.* [64] apply this idea to omni-supervised learning. They showed that by using a trained model with a labeled dataset, we can generate labels for an unlabeled dataset by applying the model on multiple transformations of the input images and aggregate the results as the *hard labels* similar to the ground truth labels. It has been shown that the aggregation will improve the results in [51, 55, 76]. Comparing to the previous methods, we believe that using weakly labeled datasets in a multi-task learning manner instead of an unlabeled one has advantages especially when the distributions of the labeled datasets and the unlabeled one is highly different.

For example, in the case of face-related tasks, if we have a dataset consists of images in domain "A" in a specific task "X" (such as facial expression recognition), and we have a datasets of images in domain "B" which they have different features with images in domain "A" and they are labeled by the other task "Y" (such as age estimation). Then if we want to use a model trained on domain "A" to estimate the facial expression of domain "B", the model which is trained only on a specific domain probably will suffer from the differences of features between the domain and won't show a good performance. Therefore, the proposed method in [64] cannot produce good labels without adopting the new domain. Therefore, in our method, we used our proposed trained MTL framework, which can learn more general features, to generate Unknown labels for all datasets (Figure 3.1). For examples, if dataset "A" has been annotated for task "X" but not task "Y" and "Z" we use our MTL network to generate "Y & Z" labels for task "A", then by doing so, we can generate more accurate predictions and can train our network in the classic MTL manner.

Chapter 4

Experiments and Results

4.1 Experiments on GF-CNN and GF-CapsNet

4.2 GF-CNN Result

We perform age classification based on the standard five-fold, subject-exclusive cross-validation protocol for fair comparison. Tables 4.1 and 4.2 show the results for age estimation, with comparisons to baselines and state-of-the-art methods. The results show that adding the Gabor responses along with images improves the performance compared to the baselines.

For the visual analysis of the effects of Gabor response feeding, we compare some feature maps in Fig. 4.1. The feature maps from our GF-Levi are shown Fig. 4.1(b), and those from the original Levi are shown in Fig. 4.1(c). As can be observed, the feature maps from the GF-Levi contain stronger facial features and wrinkle textures than the original network, which is believed to be the cause of better performance.

Table 4.1: The accuracy (%) of age estimation (classification) on Adience & Gallagher datasets, compared with the baseline. The network with prefix "GF-" uses the Gabor response input to the baseline network.

Network	Adience	Gallagher
Levi [5]	50.7±5.1	-
GF-Levi	**58.3±2.1**	71.0±0.9
VGG16 [27]	53.2±1.0	68.1±0.6
GF-VGG16	**59.2±1.3**	**72.0±0.3**
ResNet-101 [28]	54.6±2.3	69.1±0.8
GF-ResNet-101	**59.8±1.2**	**72.6±0.7**

Table 4.2: The accuracy (%) of age estimation (classification) on Adience & Gallagher datasets, compared with sate of the art techniques.

Method	Adience	Gallagher
LBP [7]	41.1	58.0
LBP+FPLBP+Dropout 0.8 [34]	45.1	66.6
Eidinger [29]	45.1	-
Levi [5]	50.7	-
PTP [35]	53.27	68.6
DAPP [35]	54.9	69.9
GF-ResNet-101 (ours)	**59.8±1.2**	**72.6±0.7**

In the case of age regression, we use four-fold cross-validation protocol for Webface dataset and the Leave-One-Person-Out (LOPO) test strategy when working on FG-Net because the number of pictures in FG-Net is small. Table 4.3 shows the result of age regression, which also indicates that our network yields better performance than the state of the art method.

The results on gender estimation is summarized in Table 4.4, which also shows that our method outperforms the other techniques on Adience. The table also shows that

h!

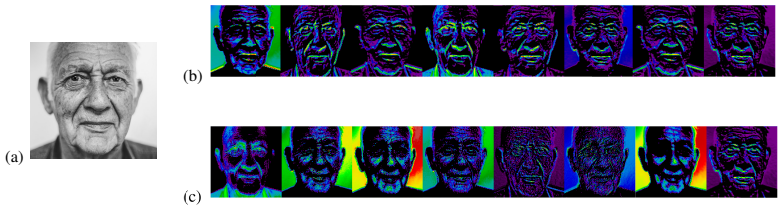


Figure 4.1: Comparison of feature maps after the first convolution layer in two networks: (a) input image, (b) feature map of GF-Levi network, (c) feature map of the original Levi Network.

Table 4.3: Mean absolute error of age regression methods on Webface, MorphII and FG-Net datasets. The last two methods with prefix GF are the networks that take the Gabor responses as the input.

Method	Casia Webface	MorphII	FG-Net
BIF[7]	10.65	5.09	4.77
EBIF[36]	-	4.11	3.17
ResNet [28]	5.80	3.13	3.10
OR-CNN[37]	5.93	3.27	-
VGG16-Hybrid [6]	5.75	2.96	-
Ranking-CNN[38]	5.71	2.96	-
GF-ResNet	5.61±0.04	2.95±0.06	**3.04±0.01**
GF-VGG16-Hybrid	**5.53±0.02**	**2.93±0.05**	3.06±0.02

the proposed method can increase the performance of baseline and also outperforms the other techniques on Webface.

Table 4.4: The accuracy (%) of gender estimation on Adience & Webface datasets.

Method	Adience	Casia Webface
BIF [7]	-	79.3
Eidinger [29]	77.8	-
Levi [5]	86.8	-
ResNet [28]	88.5	89.2
VGG16-Hybrid [6]	-	92.3
GF-Levi	90.1±1.3	92.1±1.5
GF-ResNet	**90.7±1.1**	92.4±0.5
GF-VGG16-Hybrid	90.6±0.5	**93.1±0.4**

Kernel Size

For determining the appropriate Gabor filter size, we conduct the experiments using several kernel sizes and summarize the result in Table 4.5. We can see that the smaller size works better, and we use only 3×3 filter instead of 5×5 or 7×7 like [39].

Table 4.5: The accuracy (%) of age classification on Adience and Gallagher datasets depending on the kernel size of Gabor filter banks.

GF-Levi	Age (Adience)	Age (Gallagher)
7×7	56.7	70.1
5×5	57.6	70.4
3×3	58.3	71.0

Gabor Jet vs. Single-Scale Gabor

In [25], they used the Gabor filter in eq. 2.1 with five hyper-parameters (λ, θ, ϕ, γ and σ) which are tuned depending on the given problem. Specifically, the grid search was conducted for each of the problems to find the appropriate parameters. However, since this manual optimization is time-consuming, we use Gabor jet instead of the single-scale Gabor filter used in the previous work. The Gabor jet is a set of Gabor filters with different scales and orientations, and thus using the Gabor jet is to add the multi-scale filters to the previous Gabor filters. As stated previously, we combine several scales and orientations, resulting in 32 filter banks in total. The results with Gabor jet are compared with those using the optimized single-scale Gabor in Table 4.6, which shows that using a larger number of multi-scale filters may bring better results than using a fewer number of single-scale filters with manual optimization.

Table 4.6: Comparison of using Gabor jet and single-scale Gabor filter bank on age and gender estimation. The estimation accuracy ($\%$) is measured with Adience dataset and the mean absolute error is obtained with Morph II.

Method	Age (Adience)	Gender (Adience)	Age (Morph II)
Single-Scale (Levi)	57.8±1.8	89.6±1.0	3.34±0.04
Gabor Jet (Levi)	58.3±2.1	90.1±1.3	3.30±0.05
Single-Scale (VGG16-Hybrid)	58.0±0.9	89.8±0.2	2.95±0.03
Gabor Jet (VGG16-Hybrid)	59.2±1.3	90.6±0.5	2.93±0.05

4.2.1 GF-CapsNet Results

We perform age classification on Adience and Gallagher datasets with the baseline network and our modifications, i.e., with the modified loss function and/or Gabor response input. Table 4.7 shows the effect of our modifications, and Table 4.8 shows the comparison with other methods.

Table 4.7: The effect of our modifications on CapsNet for the age classification.

Method	Accuracy on Adience (%)	Accuracy on Gallagher (%)
Caspnet EM routing (Baseline) [8]	54.9±0.9	68.1±0.4
Caspnet-2 EM routing (with an additional convolution layer)	58.7±0.8	70.4±0.5
GF-Capsnet-2 with Gabor features and raw image	56.9±0.6	71.6±0.3
GF-Capsnet-2 with reconstruction loss [19]	58.3±0.9	71.5±0.6
Capsnet-2 with modified loss function in eq.(2.9)	62.3±1.1	72.1±0.5
GF-Capsnet-2 with modified loss function in eq.(2.9)	**64.8±0.9**	**73.2±0.8**

Table 4.8: Age classification accuracy (%) on Adience and Ghallagher datasets, and Gender classification accuracy on Adience and Webface datasets.

Network	No. of parameters	Age (Adience)	Age (Gallagher)	Gender (Adience)	Gender (WebFace)
Levi [5]	22.6M	50.7	-	86.8	-
PTP [35]	-	53.27	68.6	-	-
ResBet-101 [28]	46.0	59.2	72.0	88.5	89.2
DAPP [35]	-	54.9	69.9	-	-
GF-Levi	22.7 M	58.3±1.4	71.0±0.5	90.1±1.3	92.1±1.5
GF-ResNet-101	46.3 M	59.8±0.9	72.6±0.3	90.7±1.1	92.4±0.5
GF-Capsnet (our best)	19.1 M	**64.8±0.9**	**73.2±0.8**	**92.0±0.8**	**94.0±1.0**

Also Table 4.9 shows the MAE of age regression on Webface, Morph II, and FG-Net. Regarding the network complexity, the CapsNet generally requires less number of parameters than the CNN for the same problem (see the Table 4.8). The total number of our network is 19.10M, which is even less than the number of parameters of the simplest CNN in this paper (Levi's network in Fig. 2.3).

Table 4.9: Mean absolute error of age regression on Webface, Morph II, and FG-Net datasets.

Method	Casia Webface	Morph II	FG-Net
BIF [7]	10.65	5.09	4.77
OR-CNN [39]	5.93	3.27	-
Ranking-CNN [40]	5.71	2.96	-
ODFL [41]	-	3.12	3.89
Mean Variance Loss [42]	-	2.41	2.68
GF-Capsnet (our best)	**5.32±0.46**	**2.40±0.03**	**2.61±0.08**

We conduct experiments on FER using CK+, FER2013, and Oulu-CASIA [44] datasets. All the experiment settings are the same as the previous section. Table 4.10, Table 4.11, and Table 4.12 show the results on CK+, FER2013 and Oulu-CASIA respectively. It can be seen that our network yields better performance than others on CK+ and FER2013, and comparable results on Oulu-CASIA.

Table 4.10: FER results on CK+ dataset.

Method	Accuracy on CK+ (%)
Zero-bias CNN+AD [15]	95.1±0.5
FN2EN [14]	96.8
DeRL [45]	97.30
CapsNet	97.12±0.2
GF-CapsNet (our best)	**98.13±0.3**

Table 4.11: FER results on FER2013 dataset.

Method	Accuracy on FER 2013 (%)
Maxim Milakov [40]	68.82
VGGNet [13] (Baseline)	72.18±1.1
GF-VGGNet (ours)	74.93±0.9
CapsNet	74.87±1.1
GF-CapsNet (our best)	**76.46±1.3**

Table 4.12: FER results on Oulu-CASIA dataset.

Method	Accuracy on Oulu-CASIA (%)
AUDN [43]	92.5
IACNN [17]	**95.37**
PPDN [16]	84.59
FN2EN [14]	87.71
DeRL [45]	88.0
CapsNet	84.8 ± 0.5
GF-Capsnet (our best)	88.12 ± 0.4

4.3 Experiment on Distill-2MD-MTL

We conduct experiments mainly on Facial Expression recognition, and we divide our analysis into two main parts. In the first part, we compare our network with a single-task baseline when both training data and test data are from the same domain and the same dataset (Sec. 4.3.1). In Sec. 4.3.1, we also evaluate our network on the auxiliary tasks (age and Gender estimation) to prove that not only our network shows the better result on the main task (FER), but also simultaneously improves the performance of those auxiliary ones. Following the previous works, we use 10-fold cross validation protocol for all the experiments on both CK+ and Casia datasets. We repeat each experiment 10 times and report the average result. At the last part, we compare our result when test data are from a different dataset from other domain in order to evaluate the generalization of our proposed network (Sec. 4.3.2). In this part we evaluate our network on FER2013 which has been already divided to train part and private test set by publishers [40], we follow their protocol and evaluate our network on private test part of FER2013.

In the experiments, all the images are resized to 48×48, and batch size to 128 which is divided into three parts 32, 32, and 64 for age, gender, and emotion datasets respectively, and we train our network for 200 epochs per each experiment. We utilize conventional data augmentation in the form of random sampling and horizontal flipping. To adapt VGG-16 network to our 48×48 input, we omit the last pooling layer right after VGG-16 5th block.

For optimization, we used Momentum optimizer and fix the momentum to be 0.9. We use two different methods in order to adjust the learning rate, the first one is the classic method where the learning rate starts from $10e-2$ and dropped exponentially, in the second method as the recent researches demonstrated that instead of monotonically decreasing the learning rate, vary learning rate cyclically will cause improve in performance without a need to tune and often in fewer iterations [73], we proposed a dynamic learning rate, in our method network gets feedback from loss difference

between iteration and if it seems not decreasing enough it will decrease learning rate ($lr_{T+1} = lr_T * 0.1^{(current-training-step/10^k)}$) and in each k_th times (lets call it "cycle") that this situation happens instead of decreasing learning rate, we will increase it to the initial learning rate call it as lr_{MUX}, as a result, the minimum learning rates in each cycle, step $k-1$ in each cycle, will be equal to the same value as if the learning rate has been exponentially decreased, in our experiments we set "k" to 5.

4.3.1 Semi-Supervised MTL

In this section we use three tasks; age estimation on MORPHII, gender estimation on Audience and FER on CK+, Oulu Casia. For facilitating the age estimation task, we divide it into two classes, those who are younger than 38 years old and those who are older than 43 years old and we ignore the rest. Then we evaluate our network on CK+ and Oulu Casia respectively. To have a fair compression with state of the arts as they mostly pre-trained their network [14, 55, 59, 61, 78], we also follow their method and pre-trained our network on LSEMSW same as [55] and then fine tune our network on CK+ and Casia While pre-training we didn't change the age and gender datasets.

The result has been shown in Table 4.13, Confusion matrices are also has been shown in Figure 4.3 and Figure 4.2. Moreover, we evaluate our network on MORPHII (age Estimation) and Adience (gender Estimation), while we use all images of Oulu Casia for training the FER. We divide both age and gender to two parts of train and test with a portion of 4 to 1. The result has been shown in Table 4.14. "DA" prefix indicates networks with Domain Adaptation, "Distill" for the network using knowledge distillation and "DR" for the network being trained using proposed dynamic learning rate. As the results show, the proposed method not only gets a great improvement over the baseline by exploiting the information of the other datasets from other tasks but also it works better than other multi-task approaches [49, 55] and other states of the art techniques.

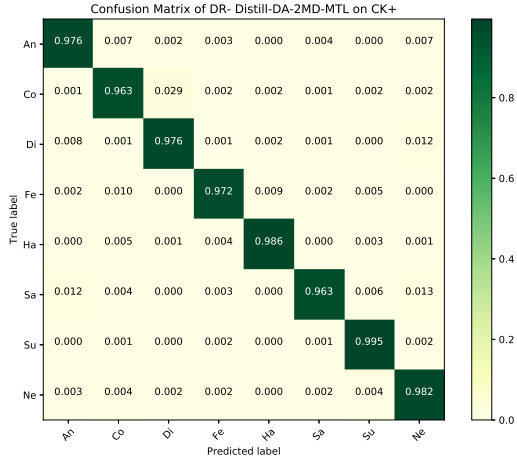
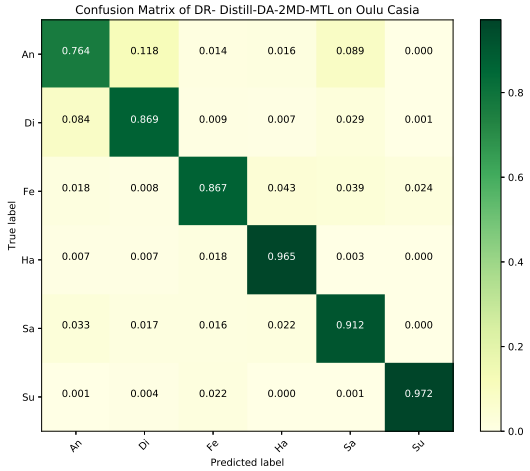


Figure 4.2: Confusion Matrix from DR- Distill-DA-2MD-MTL on CK+. The darker the color, the higher the accuracy.



5

Figure 4.3: Confusion Matrix from DR- Distill-DA-2MD-MTL on Casia. The darker the color, the higher the accuracy.

Table 4.13: Facial expression recognition result on Oulu-Casia and CK+ dataset.

Method	FER on Oulu-Casia	FER on CK+
HOG 3D [56]	70.63	91.44
IPA2LT [81]	61.02	91.67
FN2EN [14]	87.71	96.8
DeRL [78]	88.0	97.30 (7classes)
RN+LAF+ADA [55]	87.1	96.4
Star-Gan [49]	84.3	91.6
Baseline	83.6	89.4
2MD-MTL	86.83	93.51
DA-2MD-MTL	87.1	93.4
Distill-Baseline	84.1	89.2
Distill-2MD-MTL	89.13	94.5
Distill-DA-2MD-MTL	89.3	96.73
DR-Distill-DA-2MD-MTL	**90.1**	**97.68**

4.3.2 Cross Datasets Cross-Domain Evaluation

Not only our method benefits from all of the datasets to improve the results of the target dataset, but it is also capable of predicting the target labels on the domain of the auxiliary datasets. For validating these properties, we train our network same as Sec. 4.3.1, except that we evaluate our network on FER2013. For training the network we use all CK+ dataset while training on CK+, and all Casia dataset while training on Casia also we use all Casia and CK+ dataset together as training set as the number of training image in each individual dataset were so low and the network could be easily get overfed. Results are provided in Table 4.15, which shows that our method achieves significant results without seeing any labeled image of the target task in the domain of FER2013.

Table 4.14: Age estimation performance on MORPHII and Gender Estimation on Adience.

Method	Age on MORPH	Gender On Adience
Baseline	84.2	87.6
2MD-MTL	89.3	90.8
DA-2MD-MTL	88.9	91.4
Distill-Baseline	84.9	87.9
Distill-2MD-MTL	89.8	91.2
Distill-DA-2MD-MTL	89.5	91.7
DR- Distill-DA-2MD-MTL	**90.3**	**92.9**

Table 4.15: Cross-Domain facial expression recognition result on FER 2013.

Method	Trained on CK+	Trained on Casia	Trained on Both
Baseline	33.1	35.2	39.8
2MD-MTL	35.6	38.3	47.0
DA-2MD-MTL	36.7	38.5	47.3
Distill-Baseline	34.7	35.9	41.3
Distill-2MD-MTL	38.4	38.7	54.0
Distill-DA-2MD-MTL	38.1	38.5	54.2
DR- Distill-DA-2MD-MTL	**38.9**	**39.7**	**55.4**

Chapter 5

Conclusion

We have proposed techniques to increase the performance of age/gender estimation and FER. It is believed that the most important features for these problems are the shape, amount, and depth of wrinkles on the face, and the algorithms need to be robust to the variation of face rotations. We have proposed to use Gabor filter responses as the input to the deep network, which enhances the wrinkles and hence helps the network to find the wrinkle-enhanced features at the earlier stage of the convolutional layers. We have also employed the capsule network and designed appropriate loss functions, which also adds the performance improvement. In summary, using the Gabor responses as the input to the deep networks (both in the case of CNN and CapsNet) increases their performance in face-related problems. Moreover, We have proposed an end-to-end multi-dataset, multi-domain, and multi-task deep learning framework for joint facial expression, age, and gender estimation. The proposed scheme is able to exploit multiple datasets which the labels for different domains or tasks in the manner of semi-supervised learning. Hence, unlike the supervised multi-task network that needs expensive multiple labeled datasets, the proposed method is more efficiently trained. Using domain adaptation and data distillation, we were able to enhance the network generalization and solve the cross-domain adaptivity problem.

Bibliography

[1] S. Yang, P. Luo, P. C.C. Loy,and X. Tang, “From facial parts responses to face detection: A deep learning approach,” IEEE Conf. on Computer Vision (2015) 3676–3684.

[2] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, “Joint face detection and alignment using multi-task cascaded convolutional networks,” IEEE Signal Processing Letters 23 (2016) 1499–1503.

[3] P. Viola,M. J. Jones “Robust real-time face detection,” Intl. journal of computer vision 57 (2004) 137–154.

[4] O. Bilaniuk, E. Fazl-Ersi, R. Laganire, C. Xu, D. Laroche, C. Moulder, “Fast lbp face detection on low-power simd architectures,” Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (2014) 616–622.

[5] G. Levi, T. Hassner, “Age and gender classification using convolutional neural network,” CVPR Workshops (2015) 34–42.

[6] J. Xing, K. Li, W. Hu, C. Yuan, H. Ling, “Diagnosing deep learning models for high accuracy age estimation from a single image,” Pattern Recognit 66 (2017) 106–116.

[7] G. Guo, G. Mu, Y. Fu, T.S. Huang, “Human age estimation using bio-inspired features,” CVPR (2009).

[8] G.E. Hinton, S. Sabour, N. Frosst, “Matrix capsules with EM routing,” Intl. Conf. on Learning Representations (2008) 1 – 4.

[9] A. Gnay Ylmaz, V. Nabiyev, “Automatic age classification with lbp,” Intl. Symposium on Computer and Information Sciences (2008) 1–4.

[10] S. Zafeiriou, C. Zhang, Z. Zhang, “Age synthesis and estimation via faces:A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1955–1976.

[11] M. Georgescu, R.T. Ionescu, M. Popescu, “Local learning with deep and hand-crafted features for facial expression recognition,” CoRR abs/1804.10892 (2018).

[12] B. Hassani, M.H. Mahoor, “Facial expression recognition using enhanced deep 3d convolutional neural networks,” CVPR Workshops (2017).

[13] C. Pramerdorfer, M. Kampel, “Facial expression recognition using convolutional neural networks: State of the art,” CoRR abs/1612.02903 (2016).

[14] H. Ding, S.K. Zhou, R. Chellappa, “Facenet2expnet: Regularizing a deep face recognition net for expression recognition,” IEEE Intl. Conf. on Automatic Face and Gesture Recognition (2017).

[15] P. Khorrami, T.L. Paine, T.S. Huang, “Do deep neural networks learn facial action units when doing expression recognition?” IEEE ICCV Workshop (2015).

[16] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, “Peak-piloted deep network for facial expression recognition,” ECCV (2016).

[17] Z. Meng, P. Liu, J. Cai, S. Han, Y.Y. Tong, “Identity-aware convolutional neural network for facial expression recognition,” FG, 12th IEEE International Conference on, pages 558–565. IEEE, 2017.

[18] G.E. Hinton, A. Krizhevsky, S.D. Wang, "Transforming auto-encoders," ICANN (2011).

[19] S. Sabour, N. Frosst, G.E. Hinton, "Dynamic routing between capsules," NIPS (2017).

[20] D.H. Hubel, T. Wiese, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," Journal of Physiology 160 (1962).

[21] N. Petkov, "Biologically motivated computationally intensive approaches to image pattern recognition. Future Generation Computer Systems 11 (1995).

[22] C. Gottschlich, "Curved-region-based ridge frequency estimation and curved gabor filters for fingerprint image enhancement," IEEE Transactions on Image Processing 21 (2011) 2220–2227.

[23] L.L. Huang, A. Shimizu, H. Kobatake, "Robust face detection using gabor filter features," Pattern Recognition Letters 26 (2005) 1641–1649.

[24] M.J. Lyons, S. Akamatsu, M.G. Kamachi, J. Gyoba, "Coding facial expressions with gabor wavelets," IEEE Conf. on Automatic Face and Gesture Recognition. Proceedings (1998).

[25] S. Hosseini, S. H. Lee, H. J. Kwon, H. I. Koo, and N. I. Cho, "Age and gender classification using wide convolutional neural network and Gabor filter," IWAIT , 2018.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, X. Wojna, "Rethinking the inception architecture for computer vision," CVPR (2016).

[27] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR abs/1409.1556 (2014).

[28] S. H. Lee, S. Hosseini, H. J. Kwon, J. W. Moon, H. I. Koo, and N. I. Cho, "Age and gender estimation using deep residual learning network," IWAIT, 2018.

[29] E. Eidinger, R. Enbar, T. Hassner, “Age and gender estimation of unfiltered faces,” IEEE Transactions on Information Forensics and Security 9 (2014) 2170–2179.

[30] A.C. Gallagher, T. Chen, “Understanding images of groups of people,” CVPR (2009) 256–263.

[31] D. Yi, Z. Lei, S. Liao, S.Z, Li, “Learning face representation from scratch,” CoRR abs/1411.7923 (2014).

[32] K. Ricanek, T. Tesafaye, “Morph: A longitudinal image database of normal adult age-progression,” in: Proceedings of the IEEE International Conference Face Gesture, 2006, pp. 341345.

[33] A. Lanitis, “Msfgnet-ist-2000-26434 face and gesture recognition working group,” Available at, “http://yanweifu.github.io/F G N ET data/index.html Accessed: 2017-12-24.

[34] T. Wu, R. Chellappa, “Age invariant face verification with relative craniofacial growth model,” ECVV 7577 (2012) 58–71.

[35] M.T.B Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, O. Chae, “Directional age-primitive pattern (dapp) for human age group recognition and age estimation,” IEEE Transactions on Information Forensics and Security 12 (2017)2505–2517.

[36] M.E Deeb, M. El-Saban, “Human age estimation using enhanced bio-inspired features (ebif),” ICIP (2010).

[37] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, “Ordinal regression with multiple output cnn for age estimation,” CVPR (2016).

[38] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, “Using ranking-cnn for age estimation,” CVPR (2017).

[39] T. Serre, L. Wolf, S.M. Bileschi, M. Riesenhuber, T.A. Poggio, “Robust object recognition with cortex-like mechanisms,” IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 411–426.

[40] I. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang,Y. Bengio, “Challenges in representation learning: A report on three machine learning contests,” (2013).

[41] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion specified expression,” 3rd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2010).

[42] P. Liu, S. Han, Z. Meng, Y. Tong, “Facial expression recognition via a boosted deep belief network,” CVPR (2014) 1805–1812.

[43] M. Liu, S. Li, S. Shan, X. Chen, “Au-inspired deep networks for facial expression feature learning,” Neurocomput. (159) 126–136.

[44] G. Zhao, X. Huang, M. Taini, S.Z. Li, and M. Pietikinen, “Facial expression recognition from near-infrared videos,” Image and Vision Computing, 29(9):607-619.

[45] H. Yang, U. Ciftci, L. Yin, “Facial, Expression Recognition by De-expression Residue Learning,” CVPR (2018).

[46] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan,

and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 742–751. Curran Associates, Inc., 2017.

[47] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In International Conference on Computer Vision (ICCV), December 2013.

[48] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multi-task networks. In ICML, 2018.

[49] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-toimage translation. In CVPR, 2017.

[50] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces.In IEEE Transactions on Information Forensics and Security, volume 9, pages 2170–2179, Dec 2014. doi: 10.1109/TIFS.2014.2359646.

[51] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. In IEEE transactions on pattern analysis and machine intelligence, volume 32, pages 1627–1645. IEEE, 2010.

[52] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In arXiv preprint arXiv:1805.04770, 2018.

[53] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In CoRR, volume abs/1507.00448, 2015.

[54] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015.

[55] Guosheng Hu, Li Liu, Yang Yuan, Zehao Yu, Yang Hua, Zhihong Zhang, Fumin Shen,Ling Shao, Timothy M. Hospedales, Neil Martin Robertson, and Yongxin Yang. Deep multi-task learning to recognise subtle facial expressions of mental states. In ECCV, 2018.

[56] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Mark Everingham, Chris Needham, and Roberto Fraile, editors, BMVC 2008 - 19th British Machine Vision Conference, pages 275:1–10, Leeds, United Kingdom, September 2008. British Machine Vision Association.

[57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[58] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In CoRR, volume abs/1610.02242, 2016.

[59] Jiaxing Li, Dexiang Zhang, Jingjing Zhang, Jun Zhang, Teng Li, Yi Xia, Qing Yan, and Lina Xun. Facial expression recognition with faster r-cnn. In Procedia Computer Science, volume 107, pages 135 – 140, 2017. ICICT2017.

[60] L. Li, G. Wang, and Li Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In CVPR, pages 1–8, June 2007. doi: 10.1109/CVPR.2007. 383048.

[61] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. In CoRR, volume abs/1804.08348, 2018.

[62] Sijin LI, Zhi-Qiang Liu, and Antoni B. Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In CVPR, June 2014.

[63] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In International Conference on Learning Representations, 2018.

[64] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data Distillation: Towards Omni-Supervised Learning. In CVPR, 2018.

[65] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. In CoRR, volume abs/1507.02672, 2015.

[66] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In CoRR, volume abs/1711.09082, 2017.

[67] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult ageprogression. In 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pages 341–345, April 2006. doi: 10.1109/FGR.2006.78.

[68] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[69] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, ECCV, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[70] H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. In IEEE Trans. Information Theory, volume 11, pages 363–371, 1965.

[71] Razieh Sheikhpour, Mehdi Agha Sarram, Sajjad Gharaghani, and Mohammad Ali Zare Chahooki. A survey on semi-supervised feature selection methods. In Pattern Recognition, volume 64, pages 141–158. Elsevier, 2017.

[72] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for largescale image recognition. In CoRR, volume abs/1409.1556, 2014.

[73] L. N. Smith. Cyclical learning rates for training neural networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 464–472, March 2017. doi: 10.1109/WACV.2017.58.

[74] Hao Su, Qixing Huang, Niloy J. Mitra, Yangyan Li, and Leonidas Guibas. Estimating image depth using shape collections. In Transactions on Graphics (Special issue of SIGGRAPH 2014), 2014.

[75] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In CVPR, pages 2892–2900, 2015.

[76] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe.Scalable, high-quality object detection. In arXiv preprint arXiv:1412.1441, 2014.

[77] M. Taini, G. Zhao, S. Z. Li, and M. Pietikainen. Facial expression recognition from near-infrared video sequences. In ICPR, pages 1–4, 2008.

[78] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by deexpression residue learning. In CVPR, 2018.

[79] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd annual meeting on Association for Compu-

tationalLinguistics, pages 189–196. Association for Computational Linguistics, 1995.

[80] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In ICLR,2017.

[81] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In ECCV, September 2018.

[82] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In CVPR, pages 3359–3368, 2018.

[83] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, ECCV, pages 94–108, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

[84] Xiaojin Zhu. Semi-supervised learning literature survey. In Computer Science, University of Wisconsin-Madison, volume 2, page 4, 2006.

초 록

컨볼 루션 뉴럴 네트워크 (CNN)는 얼굴과 관련된 문제를 포함하여 많은 컴퓨터 비전 작업에서 매우 잘 작동합니다. 그러나 연령 추정 및 얼굴 표정 인식 (FER)의 경우 CNN이 제공 한 정확도는 여전히 실제 문제에 대해 충분하지 않습니다. CNN은 얼굴의 주름의 두께와 양의 미묘한 차이를 발견하지 못했지만, 이것은 연령 추정과 FER에 필수적입니다. 또한 실제 세계에서의 얼굴 이미지는 CNN이 훈련 데이터에서 가능할 때 회전 된 물체를 찾는 데 강건하지 않은 회전 및 조명으로 인해 많은 차이가 있습니다. 또한 MTL (Multi Task Learning)은 여러 가지 지각 작업을 동시에 효율적으로 수행합니다. 모범적 인 MTL 방법에서는 서로 다른 작업에 대한 모든 레이블을 함께 포함하는 데이터 집합을 구성하는 것을 고려해야합니다. 그러나 대상 작업이 다각화되고 복잡해지면 더 강력한 레이블을 가진 과도하게 큰 데이터 세트가 필요할 수 있습니다. 따라서 원하는 라벨 데이터를 생성하는 비용은 종종 장애물이며 특히 다중 작업 학습의 경우 장애가됩니다. 따라서 우리는 가버 필터와 캡슐 기반 네트워크 (MTL) 및 데이터 증류를 기반으로하는 다중 작업 학습에 기반한 새로운 반 감독 학습 방법을 제안한다.

주요어: 얼굴 관련 작업, Capsule Net, 데이터 증류, 다중 태스크 학습, 도메인 적응
학번: 2017-26727