Domain Separation for Person-Specific Aspects in Facial Expression Recognition

Cheong Ren Hann
rhc37@cam.ac.uk
University of Cambridge, Department of Computer Science
United Kingdom

ABSTRACT

The ability for machines to identify facial expressions has many varied applications spanning from mental health assessment to engagement recognition. Recent advances have focused on using facial action units as a means of detecting changes of expression and have been proved to be strongly correlated to specific emotional states. However, due to the large variability in the sample pool induced by gender, age, race, and personality, obtaining a model to accurately reflect expressions from unseen faces is challenging. This paper proposes using the idea of domain separation to separate features that are person-specific to those that represent facial expression. The aim is to train better models AU recognition by removing the spurious inherent features that are not important for the classification. Our results show that our model is able to outperform models that do not incorporate the separation of features and have extended the model to detect AUs on specific within facial landmarks.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Machine learning**; • **Social and professional topics** → *User characteristics*;

KEYWORDS

expression recognition, domain separation, neural networks

ACM Reference Format:
Cheong Ren Hann. 2021. Domain Separation for Person-Specific Aspects in Facial Expression Recognition. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Facial expressions have been able to accurately provide insight into an individual's affective state [11] as it has been widely acknowledged as a core means for social interaction. Automatic analysis of these expressions has wide applications ranging from mental health to psychological studies. The recent advances in Neural Networks have allowed for models to be trained quickly and effectively, though these models are often unable to reconcile the large variance in the human population such as age, gender, race, and personality which prevent accurate predictions especially in faces not seen during training. This results in varying levels of accuracy among different tested groups which can introduce bias to pipelines in which these models are used. In order to avoid this situation, it is key to be able to accurately select features that are relevant while also rejecting features that are unimportant to the task at hand.

One way to adjust for these parameters is through applying the technique called Domain Adaptation [5]. The typical objective is to train a model given supervised data for a task in a source domain to 'adapt' the model to perform in an target domain without supervised data. A common approach to Domain Adaptation is to learn feature representations whose distribution is the same on the test and training domains. Generative Adversarial Networks [7] are typically used to achieve this by simultaneously training a discriminator to distinguish whether a feature vector is derived from samples on the source or target domains and a representation encoder network that learns a mapping from the data to a feature vector so that the resulting features are indistinguishable to the discriminator. The training iterates this process such that the mapping provided by the target representation network can be used for any classifier trained on the source mapping. This results in obtaining an 'invariant' feature representation to the domain.

Domain Separation proposed by Bousmalis et al. (2016) [2] is a technique that is related to Domain Adaptation to tackle this problem using Domain Separation Networks (DSN), where a neural network is trained to learn separate feature encodings for both the source and target domain separately as well as a shared encoding for both the domains. This is achieved by introducing loss functions that measure how similar and different the features the encodings produce. The losses encourage the individual domain encoding to be distinct from the shared encoding while the shared encodings are encouraged to be domain-independent.

Inspired by this, this paper proposes using a modified Domain Separation Network with distinct encoders for person-specific and expression features. This allows a classifier to be trained on the expression features whilst also accounting for individual differences in facial features between different subjects. Initial experiments conducted have shown that this is a viable approach that is able to produce better results compared to a baseline classifier. Considering that a typical facial dataset includes multiple different faces, we can apply Domain Separation Networks which attempts to learn domain-invariant features whist also retaining unique information within each domain. This is important for facial expression classification as these domains are strongly distinct and traditional domain methods would likely ignore these individual characteristics.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

2 RELATED WORKS

The Facial Action Units (AUs) [4] is a set of 32 facial muscle actions and miscellaneous such as eye gazes, head poses, etc., actions formalized by P. Ekman et al. (2002) which allows every facial action to be described as a combination of the AUs. According to K. Scherer and P. Ekman (1982) [10] more than 7000 combinations of AUs have been observed in everyday life. Due to research [11] [1] showing the relation between facial expression and physiological and psychological states, AUs have been typically used to perform automated expression classification from facial data.

There have been many varied ways of training models to evaluate expression from AUs ranging from training Convolutional Neural Networks [6], Support Vector Machines [12] and Hidden Markov Modelling [9]. However, person-specific differences of subjects are not typically analysed when considering the feature extracted in these models. This would be an significant issue, especially when these models are trained on homogeneous/imbalanced data [8] and would likely introduce bias into the models which could hinder its applicability to the real-world.

In this project, we seek to organize the breadth of features obtained and consider the features into relevant features for expression classification and those that are specific to individual subjects. Through this process, we seek to develop models that can reduce bias while simultaneously improving classification performance by reducing spurious features.

3 METHODOLOGY

The core idea behind this approach is to consider the different faces as different domains and to separate the prediction of facial expression from the domain. These faces contains specific features that vary from person to person, these are denoted as person-specific features. By separating these features, we can also obtain the expression features which will then be used for classification.

The classification problem used for experiments in this paper will be derived from identifying AUs from cropped front-facing face images from the BP4D dataset [14]. This will form a multi-label classification problem where training samples will be taken from various subjects of images that have been annotated with AU labels and the testing samples will be of different subjects not seen in the training set.

An extension to the project adds classifies AUs in specific areas of the face, namely lower and upper. This is done by first pre-processing the images using Multitask Cascaded Convolutional Networks (MTCNN) [13] to obtain the facial landmarks, with which the images are cropped to the area used for the AU.

4 MODEL

The structure of the model is shown in Figure 1. This is a modification of the Domain Separation Network by Bousmalis et al. (2016). Instead of an private encoder for each domain, there is a single encoder $\mathbf{E}_p$ that maps all person-specific features and another that maps all expression features $\mathbf{E}_e$. The input image is first mapped into these two encoders and the feature maps are obtained, these two maps are combined and passed onto a decoder $\mathbf{D}$ to reconstruct the original image similar to an autoencoder. A separate classifier $\mathbf{C}$ retrieves the expression encoding and predicts the AUs labels without the person-specific features. Using this method, a face can be reconstructed via its person-specific and expression components (Face = Person + Expression).

There are 3 component losses in this model: $\mathcal{L}_{class}$, $\mathcal{L}_{recon}$ and $\mathcal{L}_{diff}$. This is combined to form the the overall loss $\mathcal{L}$ in the formula:

$$\mathcal{L} = \mathcal{L}_{class} + \alpha\mathcal{L}_{recon} + \beta\mathcal{L}_{diff} \tag{1}$$

where α and β are hyper-parameters.

Given N samples $\mathbf{x}_i$, $i \in \{1, \cdots, N\}$, $\mathcal{L}_{class}$ represents the classification loss of the predicted labels which is the average negative log-likelihood of the ground truth labels:

$$\mathcal{L}_{class} = -\frac{1}{N}\sum_{i=0}^{N} \mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i \tag{2}$$

where $\mathbf{y}_i$ represent the ground truth label in one-hot encoding for sample i and $\hat{\mathbf{y}}_i = \mathbf{C}(\mathbf{E}_e(x_i))$ is the prediction of the model of sample i.

The reconstruction error denoted $\mathcal{L}_{recon}$ represents the difference of the reconstructed image to the original. For this, we use a Scale-Invariant Mean Square Error (SIMSE) [3] as follows:

$$\mathcal{L}_{recon} = \sum_{i=0}^{N} \frac{1}{k_i}||\mathbf{x} - \hat{\mathbf{x}}||_2^2 - \frac{1}{k_i^2}([\mathbf{x} - \hat{\mathbf{x}}] \cdot \mathbf{1}_{k_i})^2 \tag{3}$$

where k_i is the number of pixels in $x_i, \forall i \in \{1, \cdots, N\}$, $\mathbf{1}_{k_i}$ is a vector containing ones of length k_i and $||\cdot||_2^2$ is the squared L_2 norm. SIMSE penalizes difference between pairs of pixels which compared to traditional Mean Squared Error penalizes differences up to a absolute scaling term. This allows the model to reproduce the shape and structure of the original image without being too focused on the absolute intensity and color of the image.

The difference loss $\mathcal{L}_{diff}$ allows the model to learn to choose non-overlapping features in both encoders from the input. This is given as a soft orthogonality constraint between the two feature representations. Given $\mathbf{H}_p = (\mathbf{E}_p(\mathbf{X}))_{ij}$ and $\mathbf{H}_e = \mathbf{E}_e(\mathbf{X}))_{ij}$ be the matrices whose rows are the feature representation of the samples $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$ after passing through the respective encoders.

The difference loss is thus given by the equation:

$$\mathcal{L}_{diff} = ||\mathbf{H}_p^T\mathbf{H}_e||_F^2 + ||\mathbf{H}_e^T\mathbf{H}_p||_F^2 \tag{4}$$

where $||\cdot||_F^2$ is the squared Frobenius norm.

The base model classifies AUs that span over the entire face, though an extension involves training separate models on AUs on specific areas of the face such as those on the upper area (eyes, forehead) or lower area (mouth, nose). This can be achieved by first pre-processing the input data to crop the face to the relevant areas by using a pre-trained MTCNN model to identify different facial landmarks. For this project, we will evaluate on the upper and lower areas of the face and these are defined as the above and below the mid-point of the eyes and nose respectively. To simplify the pre-processing, the landmarks are calculated in advance for each subject using samples and these landmarks will determine the crop of the image used for the model.

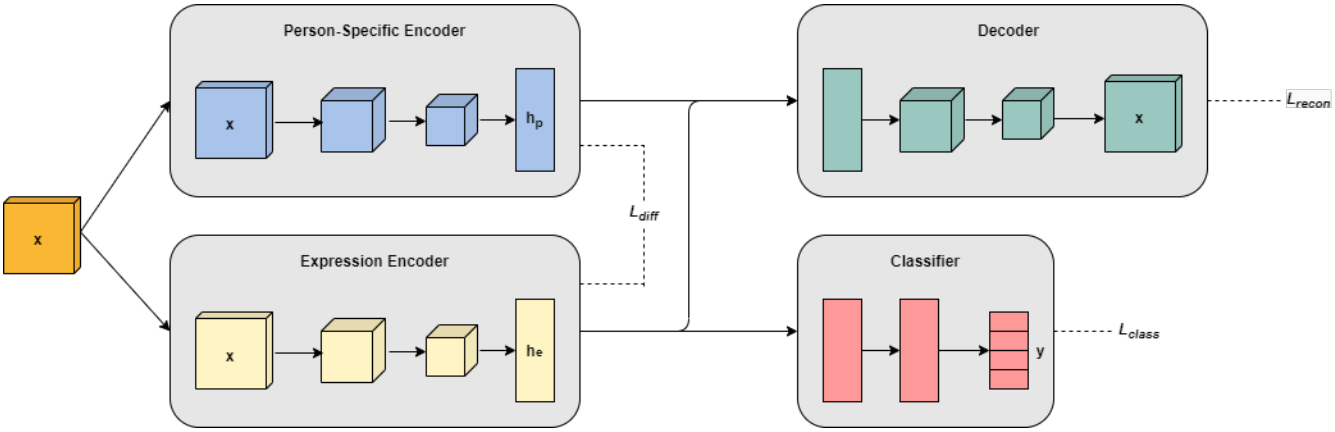

Figure 1: Structure of the Domain Separation Model with indicating loss functions.

Table 1: Summary of model results over all AUs

Model	Accuracy (%)	F1 (%)
DSN	70.2	62.1
Single Encoder	64.5	52.2

Table 2: Individual AU accuracy results

Model	AU1	AU6	AU12	AU14
DSN (Multi-label)	40.9	53.8	90.4	61.1
DSN (Single Class)	59.0	73.4	96.9	61.3

Table 3: Individual AU F1 results

Model	AU1	AU6	AU12	AU14
DSN (Multi-label)	51.5	45.7	95.1	75.3
DSN (Single Class)	65.0	82.4	97.9	76.1

5 RESULTS

For the purposes of this paper, the parameters α and $beta$ have been set to 0.01 and 0.05 respectively through empirical testing. This structure also provides a baseline to compare by removing the person-specific encoder and the $\mathcal{L}_{diff}$ loss.

The data is extracted from the BP4D dataset, which contains the frames of cropped front-facing subjects and annotations of AUs. The model is trained and evaluated using a total of 5 subjects chosen randomly split into 3/1/1 train/validation/testing sets. The model is then trained over 20 epochs with the final model model state being the one that has the best validation classification loss over all epochs.

Table 1 summarises the results for training a classifier over the entire face for the following AUs, (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24). We can see that the model is able to outperform the baseline single encoder model which suggests that the additional constraint of separating the features is able to allow the classifier to effectively learn.

Furthermore, the results for the classification task on individual AUs in Tables 2 and 3 have empirically shown that individual training of classifier on the relevant cropped areas of the AUs are able to better classify AUs compared to the joint classification done in the earlier experiments. However, this increases the amount of computation required for training as individual models need to be trained independently. Further work is needed to evaluate the effectiveness on the remaining AUs as well as the impact of using specific landmarks compared to the entire face.

Overall, we can see that at minimum the DSN model performs no worse than the baseline autoencoder network while providing potential areas for improvement and further research such as in the effectiveness of using facial landmarks.

6 CONCLUSION

Overall, applying domain separation for facial expression is able to yield promising results compared to the baseline. In addition, classifying each AU individually using facial landmarks have been shown to improve the performance of the classification though it require more training compared to a multi-label approach.

Further areas of interest could be to evaluate how changes in the hyper-parameters, in particular β, affects the training and to optimize separation of both person-specific and expression features.

REFERENCES

[1] Zara Ambadar, Jonathan W. Schooler, and Jeffrey F. Cohn. 2005. Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. *Psychological Science* 16, 5 (2005), 403–410. https://doi.org/10.1111/j.0956-7976.2005.01548.x arXiv:https://doi.org/10.1111/j.0956-7976.2005.01548.x PMID: 15869701.

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain Separation Networks. arXiv:1608.06019 [cs.CV]

[3] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. arXiv:1406.2283 [cs.CV]

[4] P. EKMAN. 2002. Facial Action Coding System (FACS). https://ci.nii.ac.jp/naid/10025007347/en/

[5] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2020. A Brief Review of Domain Adaptation. arXiv:2010.03978 [cs.LG]

[6] Pedro Carvalho Gerardo and Paulo Menezes. 2019. Classification of FACS-Action Units with CNN Trained from Emotion Labelled Data Sets. , 3766-3770 pages.

https://doi.org/10.1109/SMC.2019.8914238

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

[8] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2019. Deep Imbalanced Learning for Face Recognition and Attribute Prediction. arXiv:1806.00194 [cs.CV]

[9] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. 2000. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems* 31, 3 (2000), 131–146. https://doi.org/10.1016/S0921-8890(99)00103-7

[10] R. L. Schiefelbusch. 1983. Handbook of methods in nonverbal behavior research. Klaus R. Scherer and Paul Ekman (Eds.). Cambridge University Press, 1982. Pp. x 598. *Applied Psycholinguistics* 4, 3 (1983), 289–291. https://doi.org/10.1017/S0142716400004641

[11] Amanda Williams. 2002. Facial expression of pain: An evolutionary account. *The Behavioral and brain sciences* 25 (09 2002), 439–55; discussion 455. https://doi.org/10.1017/S0140525X02000080

[12] Li Yao, Yan Wan, Hongjie Ni, and Bugao Xu. 2021. Action unit classification for facial expression recognition using active learning and SVM. https://doi.org/10.1007/s11042-021-10836-w

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

[14] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706. https://doi.org/10.1016/j.imavis.2014.06.002 Best of Automatic Face and Gesture Recognition 2013.