2022

Facial Landmark Feature Fusion in Transfer Learning of Child Facial Expressions

Megan A. Witherow
Old Dominion University, mwith010@odu.edu

Manar D. Samad

Norou Diawara
Old Dominion University, ndiawara@odu.edu

Khan M. Iftekharuddin
Old Dominion University, kiftekha@odu.edu

Original Publication Citation

Witherow, M. A., Samad, M. D., Diawara, N., & Iftekharuddin, K. M. (2022). Facial landmark feature fusion in transfer learning of child facial expressions. *Proceedings of SPIE*, *12227*, 1-6, Article 122270P. https://doi.org/10.1117/12.2641898

Facial Landmark Feature Fusion in Transfer Learning of Child Facial Expressions

Megan A. Witherow[a], Manar D. Samad[b], Norou Diawara[c], Khan M. Iftekharuddin[a*]
[a]Vision Lab, Dept. of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA USA; [b]Dept. of Computer Science, Tennessee State University, Nashville, TN USA; [c]Dept. of Mathematics and Statistics, Old Dominion University, Norfolk, VA USA

ABSTRACT

Automatic classification of child facial expressions is challenging due to the scarcity of image samples with annotations. Transfer learning of deep convolutional neural networks (CNNs), pretrained on adult facial expressions, can be effectively finetuned for child facial expression classification using limited facial images of children. Recent work inspired by facial age estimation and age-invariant face recognition proposes a fusion of facial landmark features with deep representation learning to augment facial expression classification performance. We hypothesize that deep transfer learning of child facial expressions may also benefit from fusing facial landmark features. Our proposed model architecture integrates two input branches: a CNN branch for image feature extraction and a fully connected branch for processing landmark-based features. The model-derived features of these two branches are concatenated into a latent feature vector for downstream expression classification. The architecture is trained on an adult facial expression classification task. Then, the trained model is finetuned to perform child facial expression classification. The combined feature fusion and transfer learning approach is compared against multiple models: training on adult expressions only (adult baseline), child expression only (child baseline), and transfer learning from adult to child data. We also evaluate the classification performance of feature fusion without transfer learning on model performance. Training on child data, we find that feature fusion improves the 10-fold cross validation mean accuracy from 80.32% to 83.72% with similar variance. Proposed fine-tuning with landmark feature fusion of child expressions yields the best mean accuracy of 85.14%, a more than 30% improvement over the adult baseline and nearly 5% improvement over the child baseline.

Keywords: Facial expression recognition, transfer learning, feature fusion, facial landmarks, child facial expressions

1. INTRODUCTION

Facial expressions are a salient communication medium encoding rich psychophysiological information throughout the lifespan. In children, facial expressions provide a lens into social and emotional development. For example, children with Autism Spectrum Disorder (ASD) produce and perceive facial expressions differently than typically developing children[1,2]. Automated facial expression classification may help clinicians assess and monitor social and emotional development in children, including children with ASD.

Until recently classification performance for child facial expressions suffered from the use of models trained with inappropriate ground truth data, i.e., adult facial expressions, and lack of annotated ground truth data sets for children. The past few years have seen a growth of new annotated data sets for child facial expression classification[3-5]. However, many commercial and research solutions have continued to use models based on adult ground truth data[6-8]. Such models show poor generalization of child facial expressions due to changes in facial structure as a child grows and improvements in facial motor skills with age[9,10]. Recently, transfer learning in convolutional neural networks (CNNs) has shown promise for classifying child facial expressions through fine-tuning models pretrained on adult facial expression datasets[11,12].

While the classification of child facial expressions is a relatively new task due to the only recent availability of ground truth data, facial age estimation (FAE)[13] and age-invariant face recognition (AIFR)[14] tasks are well-established in the literature. Feature fusion has been used in state-of-the-art methods for both FAE and AIFR to combine geometric and

*kiftekha@odu.edu

Applications of Machine Learning 2022, edited by Michael E. Zelinski, Tarek M. Taha, Jonathan Howe, Proc. of SPIE Vol. 12227, 122270P
 · 0277-786X · doi: 10.1117/12.2641898

texture features, including those extracted via deep learning for more discriminative representation learning[15-17]. Furthermore, AIFR methods have used statistical latent variable models to decompose feature sets into subsets correlated with identity and age[14,15]. Ref. 18 takes inspiration from FAE and AIFR to propose a novel deep domain adaptation algorithm for child/adult-invariant facial expression classification fusing geometric landmark features that are significantly correlated with expression class. While Ref. 18 aims to jointly optimize classification performance on both child and adult expression images, we hypothesize that combining such landmark feature fusion with transfer learning may also improve knowledge transfer to maximize child expression classification performance. Thus, we propose adult-to-child transfer learning fusing geometric landmark features for classification of child facial expressions. We compare the proposed approach against multiple baselines including transfer learning and training on either adult or child expression images.

The remainder of this paper is organized as follows. Section 2 introduces relevant background information. Section 3 describes methodology and experiments. Section 4 provides results and discussion. Section 5 concludes.

2. BACKGROUND

In deep learning, transfer learning refers to reusing trainable model parameters learned while solving a 'source' task to a new but related 'target' task. The typical setting begins with a model pretrained on the source task. If the output space of the target task is different than that of the source task, the last layer(s) of the model may be removed and replaced with layer(s) appropriate for the target output space. Then, the last few layers are trained on the target data while holding the weights of early layers fixed. The layers with fixed weights are referred to as 'frozen'. This layer freezing aims to preserve the early layer representations, which are expected to be generalizable across the related source and target tasks. Following this initial training phase, some of the frozen model layers closest to the output are unfrozen. With a reduced learning rate, training continues to 'fine tune' the final layer representations for the target task. Transfer learning has shown promise for the classification of child expressions across multiple studies[11,12]. In Ref. 12, transfer learning with weighted categorical cross entropy loss improves child expression classification performance by nearly 30% over the baseline model trained with adult groundtruth.

Feature fusion refers to combining features originating from different data or extraction methods. Fusion of texture and geometric landmark features has proven effective across multiple tasks involving facial images[15,16]. Ref. 18 extracts geometric landmark features, including pairwise distances between landmarks and the areas and internal angles of triangles formed by landmark triplets. Fusion of these geometric landmark features with CNN-extracted features has improved performance on child and adult facial expression classification[18].

3. METHODS

3.1 Data sets

We consider two data sets, one source and one target, for training and evaluating our proposed method. For the source data set, we consider the Extended Cohn-Kanade (CK+)[19,20] data set. For the target data set, we consider the Child Affective Facial Expression (CAFE) set. CK+ and CAFE consist of prototypical facial expressions posed by adult and child subjects, respectively. For both data sets, we consider the following basic expression classes: 'anger', 'disgust', 'fear', 'happy', 'neutral', 'sad', and 'surprise'. For both data sets, we follow the sample selections established in past studies[11,12,18]. Thus, CK+ consists of 1254 images: 135 'anger', 177 'disgust', 75 'fear', 207 'happy', 327 'neutral', 84 'sad', and 249 'surprise'. CAFE consists of 707 images: 119 'anger', 96 'disgust', 79 'fear', 120 'happy', 129 'neutral', 62 'sad', and 102 'surprise'. We preprocess all data following Ref. 12 to yield cropped, grayscale, 256 by 256-pixel images normalized to the range [0,1].

We consider an input space $\mathcal{X}$ to represent the set of all possible input images and features. Output space $\mathcal{Y} = \{0, \dots, K\}$ denotes the set of K class labels. Then, we represent the source data set as $D_S = \{(x_i^S, y_i^S) \mid x_i^S \in \mathcal{X}, y_i^S \in \mathcal{Y}\}_{i=1}^{N_S}$ where N_S is the total number of source samples. We denote the target data set as $D_T = \{(x_i^T, y_i^T) \mid x_i^T \in \mathcal{X}, y_i^T \in \mathcal{Y}\}_{i=1}^{N_T}$ where N_T is the total number of target samples.

3.2 Facial landmark features

We use the dlib (http://dlib.net/) library to extract 68 landmark points on each facial image and follow Ref. 18 to compute the landmark features. These include Euclidean distances between pairs of landmarks and facial triangles defined by landmark triplets. Each facial triangle represents four geometric features: three internal angles and the area of the triangle.

Examples of these features are shown in Figure 1. We use the same feature selection reported by Ref. 18 for the CK+ and CAFE data sets to yield a d-dimensional landmark feature vector.

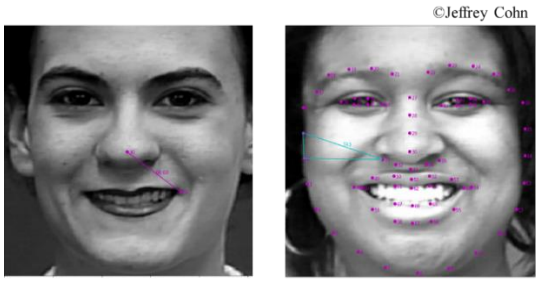

Figure 1. Examples of landmark features: (a) Euclidean distances between pairs of landmarks and (b) areas and internal angles of triangles defined by triplets of landmarks.

3.3 Feature fusion model architecture

Our model architecture, shown in Figure 2, integrates two input branches: a CNN branch for extracting features from the facial expression images and a fully connected branch that learns an embedding of the d geometric landmark features in high dimensional feature space. We define the input as a tuple $X = (U, V) \in \mathcal{X}$, where $U \in \mathcal{U} = \mathbb{R}^{256 \times 256}$ is the input to the CNN branch and $V \in \mathcal{V} = \mathbb{R}^d$ is the input to the fully connected branch. The CNN branch consists of three convolutional layers with 16, 32, and 64 feature maps, respectively. A kernel size of 3x3 is used in all convolutional layers and each convolutional layer is followed by 2x2 maximum pooling. We use the ReLU activation function in the convolutional layers. The fully connected layer has 128 hidden nodes with ReLU activation. Each branch learns a 128-dimensional feature vector. These are concatenated to yield a 256-dimensional latent feature space. A final fully connected layer with K hidden units and softmax activation is used to yield predicted class labels $Y \in \mathcal{Y}$.

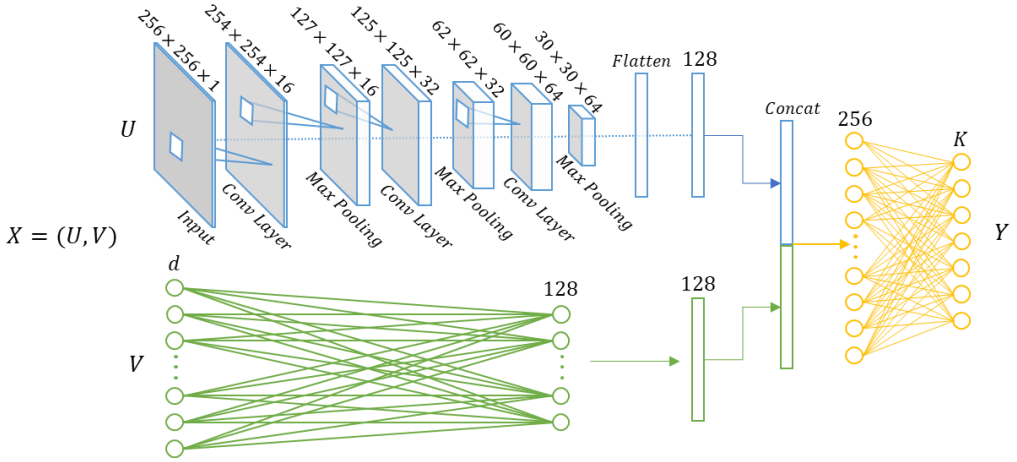

Figure 2. Model architecture fusing CNN-extracted features with geometric landmark features for classification of child facial expressions.

3.4 Transfer learning

For transfer learning, we define the source task, adult facial expression classification, as T_S. T_S corresponds to learning a model $f(\cdot)$ for optimally predicting class labels Y when inputs $X = (U, V)$ represent adult facial expressions. The goal of transfer learning is to use knowledge learned from T_S to solve related target task T_T, child facial expression classification.

We begin by training model $f(\cdot)$, initialized with uniform weights, on training data from D_S to learn a set of weights for T_S. Next, we freeze all model layers except for the final classification layer such that the weights of frozen layers are not updated during training. Then, we train the classification layer on the training data from D_T with a learning rate of

1×10^{-3}. To fine-tune $f(\cdot)$ for T_T, we unfreeze the last convolutional layer and continue training on D_T with a reduced learning rate of 5×10^{-4}.

We supervise the optimization for all training sessions using the categorical cross-entropy loss. We use a batch size of 32 and biased batch sampling to ensure that all classes are equally represented in each batch. We determine the number of training epochs using a hold-out validation set.

3.5 Experiments

To evaluate our proposed approach, we partition both data sets into train and test sets via subject-independent 10-fold cross validation[11,12]. This yields ten 90% train, 10% test sets. We then partition the 90% train set into 80% train and 10% validation. We tune the number of epochs for training using the validation set. After that, we recombine the 80% train and 10% validation sets and train the model on the full 90% train data for the number of epochs determined in the previous step. We evaluate the resulting trained model on the 10% test set.

We evaluate the performance of our proposed transfer learning and facial landmark fusion model against six comparison models: (1) CNN trained on source data, (2) CNN fusing landmark features trained on source data, (3) CNN trained on target data, (4) CNN fusing landmark features trained on target data, (5) CNN with transfer learning, and (6) Ref. 12.

4. RESULTS AND DISCUSSION

We first train our proposed model using the 80% train set and observe the 10% validation set performance. Figure 3 shows representative training and validation plots for our proposed transfer learning and landmark feature fusion model. The model training is well behaved showing a decreasing validation loss curve(Figure 3(b)). The learning rate is decreased for fine tuning starting from epoch 16, followed by continued improvement in validation accuracy (Figure 3(a)).

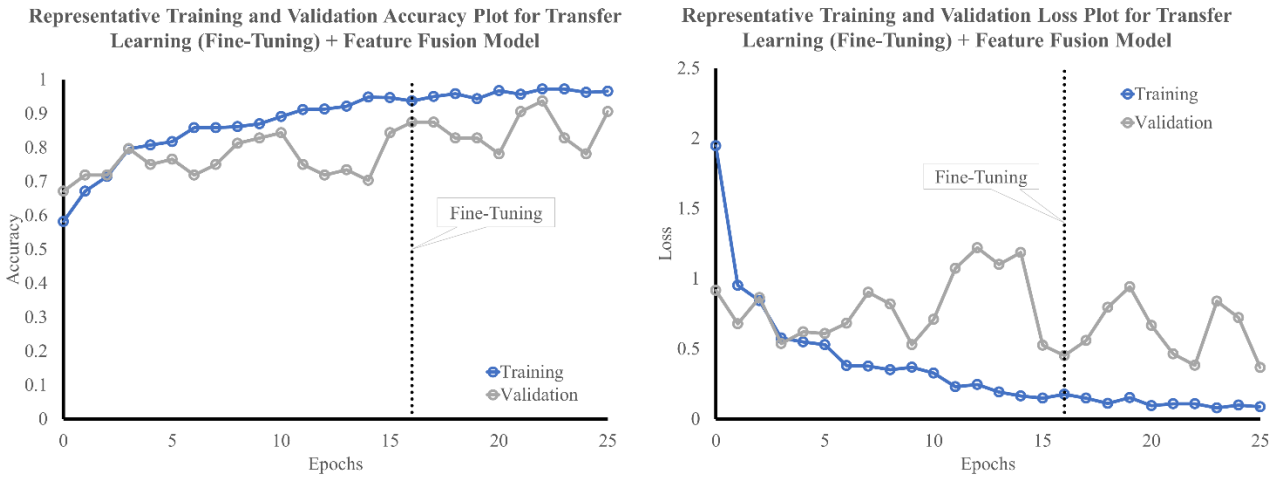

Figure 3. Representative training and validation plots for (a) accuracy and (b) loss produced during the training of our proposed transfer learning and facial landmark feature fusion model.

The performance of our proposed transfer learning and facial landmark feature fusion model and comparison models is reported in Table 1. Our proposed model outperforms all comparison models, achieving the best mean 10-fold cross validation accuracy of 85.14% ± 3.81%. This accuracy represents an improvement of 2.41% over transfer learning without feature fusion. Furthermore, we find that the CNN models fusing facial landmark features outperform CNN models without feature fusion resulting in 1.42% and 3.40% higher accuracies on average for CNNs trained on source and target data, respectively. These results suggest that the fusion of facial landmark features provides additional discriminative information to augment child facial expression classification.

Table 1. Comparison of proposed and comparison models on 10-fold cross validation mean accuracy for 7-class child facial expression classification; Source: CNN trained on adult facial expressions; Target: CNN trained on child facial expressions

Model	**Mean Accuracy**
Source CNN	54.15% ± 5.97%
Source CNN + facial landmark feature fusion	55.57% ± 5.42%
Target CNN	80.32% ± 4.08%
Target CNN + facial landmark feature fusion	83.72% ± 5.11%
Transfer learning	82.73% ± 4.60%
Ref. 12: Witherow et al. (2019)	76.03% ± 7.06%
Transfer learning + facial landmark feature fusion	**85.14% ± 3.81%**

Figure 4 plots representative confusion matrices for our proposed transfer learning and facial landmark feature fusion model and the transfer learning comparison model. The most confusing pair of expressions for the transfer learning model is 'anger'/'disgust', likely due to the visual similarity of these two expressions. By contrast, our proposed model augmented with the landmark features confuses fewer samples of this challenging expression pair.

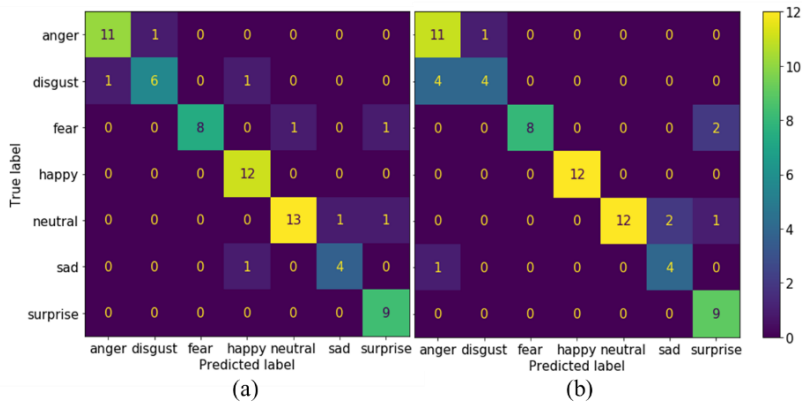

Figure 4. Representative confusion matrices for (a) proposed transfer learning and facial landmark feature fusion and (b) transfer learning.

5. CONCLUSION

This work proposes transfer learning by fusing facial landmark features for the classification of child facial expression images. Our proposed approach outperforms all comparison models, including transfer learning, CNN trained on adult expression data, and CNN trained on child expression data. Furthermore, we find that performance improvements offered by feature fusion are not limited to only the transfer learning setting and may also benefit models trained from uniform weights initialization. In future work, we plan to investigate the effect of facial landmark feature fusion on more complex tasks in facial expression analysis, such as the classification of facial action units from the Facial Action Coding System. We hope that continued improvements in facial expression analysis for children will facilitate systems for automatic and ongoing assessments of social and emotional development in children to aid clinicians in treating conditions such as ASD.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1753793 and by the Research Computing clusters at Old Dominion University under National Science Foundation Grant No. 1828593.

REFERENCES

[1] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, "A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data," IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(2), 353-361 (2018).

[2] M. D. Samad, N. Diawara, J. L. Bobzien, C. M. Taylor, J. W. Harrington, and K. M. Iftekharuddin, "A pilot study to identify autism related traits in spontaneous facial actions using computer vision," Research in Autism Spectrum Disorders, 65, 14-24 (2019).

[3] V. LoBue, and C. Thrasher, "The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults," Frontiers in Psychology, 5, (2015).

[4] V. LoBue, and C. Thrasher, "The Child Affective Facial Expression (CAFE) set. Databrary. 2014," Series The Child Affective Facial Expression (CAFE) set. Databrary. 2014. (2014).

[5] J. G. Negrão, A. A. C. Osorio, R. F. Siciliano, V. R. G. Lederman, E. H. Kozasa, M. E. F. D'Antino, A. Tamborim, V. Santos, D. L. B. de Leucas, P. S. Camargo, D. C. Mograbi, T. P. Mecca, and J. S. Schwartzman, "The Child Emotion Facial Expression Set: A Database for Emotion Recognition in Children," Frontiers in Psychology, 12, (2021).

[6] Noldus Information Technology bv, "FaceReader," Series FaceReader. 2022(2022).

[7] iMotions A/S, "Facial Expession Analysis," Series Facial Expession Analysis. 2022(2022).

[8] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," Series Openface 2.0: Facial behavior analysis toolkit. 59-66 (2018).

[9] C. Grossard, L. Chaby, S. Hun, H. Pellerin, J. Bourgeois, A. Dapogny, H. Ding, S. Serret, P. Foulon, and M. Chetouani, "Children facial expression production: influence of age, gender, emotion subtype, elicitation condition and culture," Frontiers in psychology, 446 (2018).

[10] P. Burke, and C. Hughes-Lawson, "The growth and development of the soft tissues of the human face," Journal of anatomy, 158, 115 (1988).

[11] M. Witherow, W. Shields, M. Samad, and K. Iftekharuddin, "Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning," Series Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning. 11511(2020).

[12] M. Witherow, M. Samad, and K. Iftekharuddin, "Transfer learning approach to multiclass classification of child facial expressions," Series Transfer learning approach to multiclass classification of child facial expressions. 11139(2019).

[13] P. Punyani, R. Gupta, and A. Kumar, "Neural networks for facial age estimation: a survey on recent advances," Artificial Intelligence Review, 53(5), 3299-3347 (2020).

[14] K. Baruni, N. Mokoena, M. Veeraragoo, and R. Holder, "Age Invariant Face Recognition Methods: A Review," Series Age Invariant Face Recognition Methods: A Review. 1657-1662 (2021).

[15] L. Meng, C. Yan, J. Li, J. Yin, W. Liu, H. Xie, and L. Li, "Multi-Features Fusion and Decomposition for Age-Invariant Face Recognition," Series Multi-Features Fusion and Decomposition for Age-Invariant Face Recognition. 3146–3154 (2020).

[16] S. Taheri, and Ö. Toygar, "On the use of DAG-CNN architecture for age estimation with multi-stage features fusion," Neurocomputing, 329, 300-310 (2019).

[17] M. S. Shakeel, and K.-M. Lam, "Deep-feature encoding-based discriminative model for age-invariant face recognition," Pattern Recognition, 93, 442-457 (2019).

[18] M. A. Witherow, M. D. Samad, N. Diawara, H. Y. Bar, and K. M. Iftekharuddin, "Deep Adaptation of Adult and Child Facial Expressions Fusing a Selection of Facial Landmark Measurements," arXiv preprint, (2022).

[19] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," Series Comprehensive database for facial expression analysis. 46-53 (2000).

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," Series The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 94-101 (2010).