Facial Expression Recognition via Joint Deep Learning of RGB-Depth Map Latent Representations

Oyebade K. Oyedotun[1], Girum Demisse[1], Abd El Rahman Shabayek[1,2],
Djamila Aouada[1], Björn Ottersten[1]
[1]Interdisciplinary Centre for Security, Reliability and Trust (SnT),
University of Luxembourg, L-1855 Luxembourg
[2]Computer Science Department, Faculty of Computers and Informatics,
Suez Canal University, Egypt
{oyebade.oyedotun, girum.demisse, abdelrahman.shabayek, djamila.aouada
bjorn.ottersten}@uni.lu

Abstract

Humans use facial expressions successfully for conveying their emotional states. However, replicating such success in the human-computer interaction domain is an active research problem. In this paper, we propose deep convolutional neural network (DCNN) for joint learning of robust facial expression features from fused RGB and depth map latent representations. We posit that learning jointly from both modalities result in a more robust classifier for facial expression recognition (FER) as opposed to learning from either of the modalities independently. Particularly, we construct a learning pipeline that allows us to learn several hierarchical levels of feature representations and then perform the fusion of RGB and depth map latent representations for joint learning of facial expressions. Our experimental results on the BU-3DFE dataset validate the proposed fusion approach, as a model learned from the joint modalities outperforms models learned from either of the modalities.

1. Introduction

Facial expressions have been used successfully and effectively by humans for communicating their emotional states. The six basic facial expressions include *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. Our excellent ability to correctly read-off facial expressions from people usually lead to a host of corresponding responses. For example, when a teacher read-offs an expression of surprise on the faces of students during a lecture, he may decide to restate his point in a more comprehensible way for the students. In general, other areas where facial expression recognition play critical roles for improving communication and responses include counselling, interrogation, reward games, etc.

Interestingly, tasking computers to perform facial expression recognition (FER) is an active research problem in the human-computer interaction domain [1]. A major motivation for this is to allow a computer adapt its response based on the facial expression of the user; this consequently fosters better human-computer interaction. However, recognizing facial expressions of subjects with reasonably high accuracy is not an easily achievable task for computers. Several works have studied FER and proposed different approaches for improving the performance of FER systems. Many of these works rely on learning facial expression features via appearance, fiducial landmarks or detecting action units first, and then subsequent classification [2][3]. For feature based FER systems, learning highly discriminative and robust facial expression is very critical. This approach is challenging as the features that discriminate the different facial expression are quite subtle [4]; the usually more abundant features pertaining to the identity of the subject could easily dominate the discriminative features for the different facial expressions.

In recent times, deep learning has become very useful for many computer vision applications [5][6]. One of the main motivation behind deep learning is that learning several hierarchical levels of feature abstractions fosters the disentanglement of the different aspects in the training dataset, as the different levels in the model can represent different aspects in the training dataset [7][8]. Moreover, deep neural networks are powerful function approximators, with the capacity for learning complex and highly varying functions [9].

In this paper, we approach FER via appearance based representation using deep learning and the fusion of RGB and depth map latent representations. Our motivation for the proposed approached lies in learning jointly from depth map and RGB modalities. Since, we propose an appearance based modelling of FER, we employ deep convolutional neural network (DCNN) for learning the *subtle* features that discriminate the six basic different facial expressions. Particularly, we posit that learning jointly from depth map and RGB modalities improves the robustness of discriminative features learned for FER as against learning singly from either of the two modalities. We take inspiration from the demerits of employing either of the two modalities separately for learning facial expression features that are highly discriminative. RGB data can effectively capture very rich information about facial expressions; however, RGB data are quite susceptible to severe corruption due to illumination variations. In fact, RGB data may be rendered considerably ineffective for learning very discriminative features. Conversely, depth map data are modestly tolerant of illumination changes and therefore features learned from such data are more stable and robust. In addition, depth map data carry important information describing geometric relations in an image. However, in well setup and stable environmental settings, depth map data may be less rich in information as compared to RGB data. The contributions of this paper are as follows:

1. We posit that learning jointly from depth map and RGB modalities result in a more robust classifier for FER as against learning singly from either of the two modalities. Our extensive experiments and results on the BU-3DFE dataset validates our position.

2. We construct and describe a pipeline based on transfer learning and deep convolutional neural network (DCNN) that are later used to learn and fuse several hierarchical levels of feature representations from RGB data and depth map.

3. We provide a comparison between our method and state-of-the-art approaches on BU-3DFE dataset. Particularly, the result of our fusion approach is better than many state-of-the-art results.

The rest of this paper is organized as follows. Section 2 discusses related works along with their merits and demerits for FER. Section 3 describes the necessary background and the problem formulation followed by the proposed approach, in Section 4. In section 5, we give the details of experiments and results along with discussions. In section 6, we conclude the paper by highlighting our key findings.

2. Literature review

Several studies have been carried out on facial expression recognition (FER). Many of these works essentially rely on two stages of information processing: (1) features extraction [10](2) classification of extracted features [11][12]. Occasionally, there is a feature selection stage as an intermediate stage between the feature extraction stage and classification [13]. Additionally, FER systems can be developed based on geometric [14] or appearance [15] features modelling. For geometric based FER modelling, [16] relied on Haar-based features and the Viola-Jones' algorithm for detecting the eye regions; the work emphasized the challenges of accurately detecting the eyes and other facial regions from which features for FER are extracted; for classification a total of 26 geometric features where extracted and fed into a self-organizing map for classification. In [17], estimation of principal curvature was employed for labelling the vertices of 3D facial surface individual models. The work [17] then segmented facial surfaces into expressive regions and applied histogram statistics on the segmented regions. Finally, the features obtained from the histogram statistics were used to train a linear discriminant classifier for FER. In [18], a FER system was proposed based on the combination of Bayesian Belief Net (BBN) and statistical facial features model. The work described an approach where manual landmarking for facial features extraction is eliminated. Another interesting work [19] proposed using 2D and 3D features for FER. The work employed incremental Parallel Cascade of Linear Regression (iParCLR) for simultaneously localizing fiducial landmarks from 2D and 3D scans. For extracting features from the 2D scans, a novel Histogram of Second Order Gradients (HSOG) and first-order gradient based SIFT descriptor were used. For extracting features from the 3D scans, Histogram of mesh Gradients (meshHOG) and Histogram of mesh Shape index were used. Subsequently, the features obtained from both 2D and 3D scans were used to train a support vector machine (SVM); further improvement in result was reported by fusing 2D and 3D modalities. In [20], an elaborate FER system was proposed composing low pass filtering, eye, nose, lip corner and eyebrow corner detection; features were extracted from detected regions of interests using local binary pattern (LBP) encoding. Later, support vector machine (SVM) based classifier is trained on a feature space estimated with Principal Component Analysis (PCA) from LBP encoding. One obvious demerit of the geometric based modelling is the elaborate arrangement for the detection of regions of interest in face images.

Alternatively, other works pursue appearance based modelling for FER. Here, arrangement for elaborate detection of regions of interest in face images is eliminated; instead, through highly robust learning schemes, facial expression features are learned (or extracted) from face im-

ages with no or minimal processing. For example, [21] proposed a boosted deep belief network for performing feature learning, selection and classification all together as a unified framework. In another work [22], a deep neural network was employed for FER, and they report state-of-the-art results on extended Cohn-Kanade (CK+) dataset and the Toronto Face Dataset (TFD). More interestingly, they show that the features learned by the deep network are akin to facial action units (FAUs) for facial expressions.

Another approach for FER which have been studied in several works relies on the detection of specific FAUs based on the facial action coding system (FACS) [23]. In [24], FAUs were explored for composing new expressions which they referred to as compound expressions, while [25] proposed a FER system based on the detection of some specific FAUS in face images.

3. Problem formulation

Learning robust discriminative features is critical to the performance of appearance based approaches for FER. However, learning robust discriminative features directly from raw data is not at all trivial. Deep neural networks have large representation capacity allowing the learning of complex target functions, and therefore seems a natural choice for modelling such a daunting task. For example, [26] employed an ensemble of deep neural networks based on transfer learning; firstly, they trained their models on a much larger dataset (i.e. 2013 facial Expression Recognition (FER) Challenge dataset), then fine-tuned the models on the target dataset-Static Facial Expressions in the Wild (SFEW) dataset.

However, training DCNNs require massive datasets as deep models typically compose lots of parameters; with small datasets, deep models have the tendency to quickly over-fit [27]. One solution which has been explored in many works that employed deep models for small datasets [27] [28] is transfer learning, where the features learned by a deep model on a generic and very large dataset are transferred to another task. Here, the hope is that the features learned by the pre-trained model are considerably general enough for other tasks; this is especially the case when the dataset on which the model has been pre-trained considerably shares similarity with the target dataset. Otherwise, one cannot guarantee the performance of feature transfer when the target dataset is quite different from the one on which the model has been pre-trained; in this case, the suitability of feature transfer can only be determined via experimentation [27]. Since, the dataset used in this work is relatively small, we experiment with transfer learning to determine the scenario (i.e. data modality) where it is useful for improving feature learning.

Considering the aforementioned strengths of depth map and RGB data, we formulate an approach for extracting more interesting features for facial expression recognition. Namely, the proposed approach is aimed at solving the following problems:

- Feature learning from a small dataset in cases where transfer learning is not effective (the depth map case).
- More effective learning of facial expression features from both depth map and RBG modalities via the fusion of their latent representations.

4. Proposed facial expression framework

In this section, we relate how the proposed approach addresses the problems that we propose to solve as mentioned in section 3. We give the details of the dataset used in this paper, data pre-processing, proposed learning pipeline, model architectures, and considerations for the different pipeline constructions. The highlights of the proposed facial expression recognition framework and the problems that are tackled are as follows:

- We can leverage both depth map and RGB modalities for extracting more robust discriminative features for FER. The aim is that the individual strength of the modalities can be used compliment each other via the fusion of extracted latent representations.
- We consider the usefulness of hierarchical feature representations to construct a pipeline that allows us to learn several levels of latent representations and perform fusion of latent representations of RGB and depth map modalities; such a pipeline should allow the different levels learn different and interesting aspects of the training data.
- We rely on the state-of-the-art pre-trained models (i.e. ResNet50 [29] and VGG19 [30]) where they are helpful for learning several levels of latent representation; otherwise, we train our model from scratch and rely on other training schemes such as batch normalization [31] to improve optimization and generalization.

4.1. Dataset

For validating the proposed approach in this paper, we use the BU-3DFE dataset [17] that is composed of 100 subjects (56 female and 44 males) with a total of 2500 textured 3D scans along with corresponding 2D (RGB) data. Furthermore, each subject perform the six basic facial expressions at 4 different intensity levels; see Figure.1 [32].

4.2. Data pre-processing

The BU-3DFE dataset offers both 3D and 2D (RGB) data for all expressions contained therein. Since, the conventional DCNN can be employed directly only for structured data such as images, we obtain depth map images

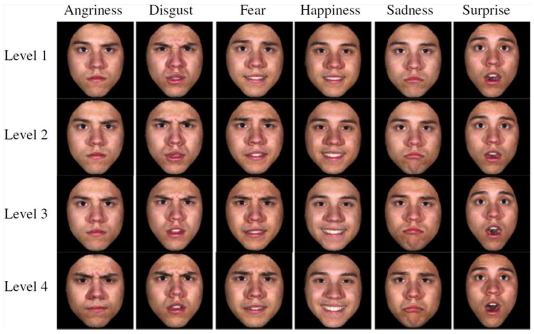

Figure 1. BU-3DFE dataset facial expressions [32]

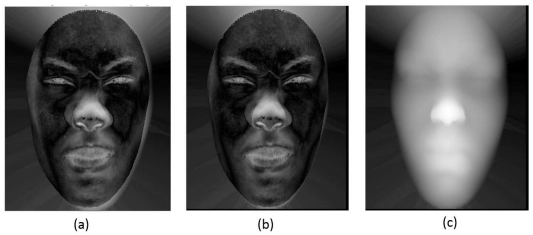

Figure 2. BU-3DFE depth map pre-processing

from the 3D data and then perform registration using the corresponding RGB images via maximization of mutual information. The registration of depth map images is meant to improve the joint learning of facial features from both depth map and RGB data; this is straightforward since we want the DCNN to strictly capture the same aspect of the facial images during training. For example, Fig.2(a) shows the superposition of the obtained and unregistered depth map over the grayscale transform of the RGB data for an expression; Fig.2(b) shows the superposition of the registered depth map over the grayscale transform of the RGB data for the same expression; Fig.2(c) shows the registered depth map. Furthermore, we employ an algorithm that captures the face images in bounding boxes, eliminating the redundant background data; this should make learning facial expression features more concise.

4.3. Learning pipeline

Since, we propose to show that learning jointly from depth map and RGB data allows us to learn more robust features with higher discriminative power than learning singly from either of the two modalities, we construct three different pipelines to validate the argument presented in this paper. We refer to the three different pipelines as PL-depth map, PL-RGB and PL-fusion; where, PL-depth map, PL-RGB and PL-fusion denote the constructed pipelines for depth map data, RGB data and their fusion, respectively. For the depth map and RGB data, we experiment with transfer learning using ResNet50 [29] and VGG19 [30] models that were pre-trained on the imagenet dataset. The three pipelines are discussed below.

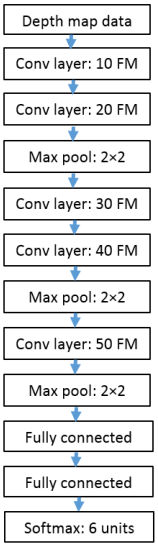

Figure 3. Learning pipeline for depth map modality: PL-fusion

4.3.1 PL-depth map

We experiment with transfer learning using ResNet50 and VGG19 for the pre-processed depth map images. However, we observe that employing the pre-trained models for the depth map data resulted into poor performance. Our explanation is that the features learned by the pre-trained models are considerably unuseful for the depth map images. This is not surprising, as the target data is quite dissimilar to the one that the pre-trained models were trained on; the target data, depth maps, are range images while the data (i.e. imagenet dataset) used for pre-training are natural colour images.

Therefore, we opt for training a DCNN from scratch on the depth map images. Firstly, we resize the depth map images from the original size of 512×512 pixels to 64×64 pixels to reduce training time and computational requirement. We then construct a deep model with 5 convolution layers and 3 max pooling layers. The architecture of the DCNN is as follows: {input:depth map}-{conv layer:10 FM}-{conv layer:20 FM}-{max pool:2×2}-{conv layer:30 FM}-{conv layer:40 FM}-{max pool:2×2}-{conv layer:50 FM}–{max pool:4×4}-{FC:300 units}-{FC:300 units}-{softmax:6 units}; where, FM and FC denotes feature maps and fully connected layers, respectively; all convolution operations use filters of size 6×6. The constructed pipeline is shown in Fig.3.

4.3.2 PL-RGB

For the RGB data, we find that the pre-trained networks ResNet50 and VGG19 work well, hence we do not construct models that are trained from scratch as carried out for the depth map data. We take the pre-trained models, remove the fully connected layers including the softmax and stack

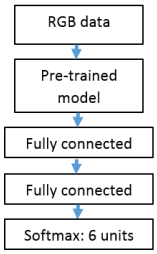

Figure 4. Learning pipeline for RGB modality: PL-RGB

two newly initialized fully connected layers and a 6-way softmax layer for training to classify the six basic facial expressions. The full architecture of the PL-RGB pipeline is as follows: {input:RGB data}-{pre-trained model}-{FC1}-{FC2}-{softmax:6 units}; where, 'pre-trained model' is either ResNet50 or VGG19; for ResNet50, FC1 and FC2 have 1000 and 700 units respectively; while for VGG19, FC1 and FC2 have 500 units each. The constructed pipeline is shown in Fig.4. Furthermore, for the ease of referral, we tag the pipeline with pre-trained ResNet50 as *PL-RGB-ResNet50*, and the other pipeline with pre-trained VGG19 as *PL-RGB-VGG19*.

4.3.3 PL-fusion

Here, we construct a pipeline for the joint learning of latent representations obtained from the depth map and RGB modalities. We take the following considerations in constructing the pipeline to achieve two important goals: (1) learning several hierarchical representations for both data modalities (2) fusing the separately learned high level representations and then learning jointly from the new and much richer representations. The constructed pipeline is shown in Fig.5. The depth map and RGB pipelines before the fusion of latent representations have the same construction as though they operate singly; this scenario can be seen as a form of late fusion. Note that for the fusion of depth map and RGB modalities, we concatenate the latent representations obtained from both depth map and RGB data, and then learn two new levels of latent representations before feeding into a 6-way softmax classifier layer.

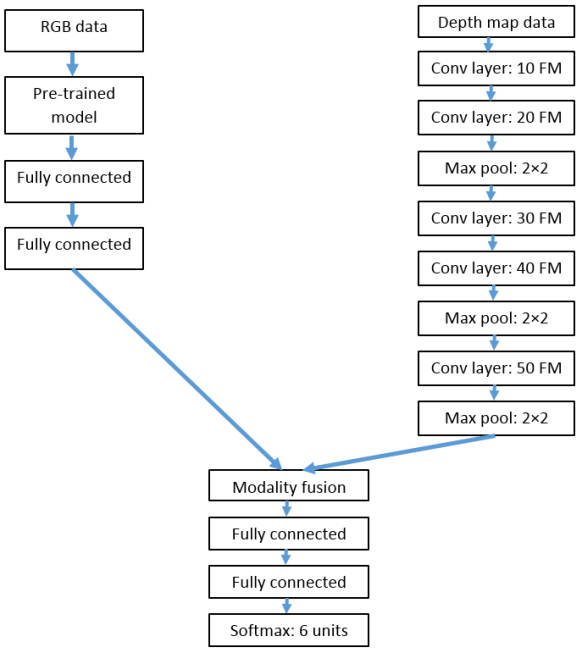

Figure 5. Depth map and RGB latent representation fusion pipeline: PL-fusion

5. Experiments

5.1. Pipeline training

For all the pre-training scenarios reported in this paper, we perform global average pooling at the last layer to reduce spatial dimensionality of data before feeding into the fully connected layers. Since we employ batch normalization [31] for improving generalization and optimization, we initialize all the newly added layers via He initialization [33] to improve convergence rates of models during training. Our observation aligns with [31] on the use of dropout and batch normalization together for training; in our experiments, using dropout along with batch normalization did not bring any improvement in learning. Hence, we did not employ dropout for all newly added layers reported in this paper. Since, our pipelines are based on several stacks of feature representations, all the pipelines in this paper use hidden units with the rectified linear function for activation; again, this is aimed at tackling units saturation and improving gradients condition for fast convergence during training. Furthermore, we employ mini-batch gradient descent optimization for training all models. Particularly, we rely on the adaptive moment estimation (adam) technique as in [18] for mini-batch gradient descent optimization so that we can be less careful about the set learning rate for model training. Nevertheless, we set the learning rate for all models in this paper to a considerably small value of 0.003, and training all the different pipelines for a maximum of 1000 epochs.

5.2. Experimental setup

As mentioned earlier, we follow the common protocol as in [17][34][18][35] and many other works for choosing training and testing data. We randomly select 60 subjects from the BU-3DFE dataset; from these selected subjects, we randomly select 54 subjects for building the training data and the remaining 6 subjects for building the test data; this is akin to a 10-fold cross validation training scheme. In addition, we use only the third and fourth intensity levels of facial expressions. Again, in conformity with the training

and model evaluation protocol found in [17][34][18][35] and many other works, we fix the randomly selected 60 subjects, and perform for a total of 200 experiments for the random selection of 54 subjects as training data and remaining 6 subjects as test data similar to a 10-fold cross-validation scheme; this is meant to reduce sampling bias on training and testing data for model training and evaluation. All the experiments and results given in this paper are based on this training and testing protocol.

However, we note that some other works [34] among others follow a different training and evaluation protocol where the selected 60 subjects are not fixed, but are randomly selected for each experiment. In this paper, we do not consider such a training and evaluation experimental setup.

5.3. Results and discussion

We report in Table 1 the results of the experiments performed for the different learning pipelines discussed in section 4. We observe that the depth map and RGB modalities achieve promising results when they are used separately for FER. When the depth maps only are used for training a DCNN from scratch (i.e. PL-depth map) as described in section 4.3.1, we obtain a test accuracy of 84.72%. Also, when the RGB data are used for obtaining features from pre-trained models, we obtain a test accuracy of 82.92% for ResNet50 and 81.25% for VGG19. Interestingly, combining both depth map and RGB modalities, we are able to reach a test accuracy of 87.08% using pre-trained ResNet50 for the RGB data (i.e. PL-fusion-ResNet50 in Table 2) and 89.31% using pre-trained VGG19 for the RGB data (i.e. PL-fusion-VGG19 in Table 2); for both fusion experiments, DCNNs are trained from scratch on the depth map data.

We include in Table 1 the results of parallel experiments obtained using pre-trained models ResNet50 and VGG19 on depth maps; these are the first two results given in Table 1 with asterisk. Particularly, we observe poor performance in this scenario as features do not seem transferable from the imagenet pre-trained models to depth maps. When we rely on features obtained from the pre-trained ResNet50 for the depth map data, we obtain a poor test accuracy of 61.11%; using the pre-trained VGG19 in this setting, a much worse test accuracy of 28.06% is obtained. Here, we note that the results are not at all surprising as there is a large dissimilarity between the imagenet data on which the ResNet50 and VGG19 were pre-trained and the target data in this paper which are depth maps; features obtained from the pre-trained models in this setting are far from the optimal representations of the target data. In fact, we note that [27] also acknowledged that the transferability of features decreases with the increase in disparity between the base task (i.e. pre-training data) and target task (i.e. target data). Hence, we did not consider these models useful enough for constructing any of the pipelines described in this paper; rather, we chose to train a DCNN from scratch on the depth map data. Table 2 shows the results (i.e. test accuracy in %) for the different pipelines constructed in this paper, along with state-of-the-art results reported in other works that used a similar training and testing protocol to ours on the BU-3DFE dataset. Firstly, we observe that our result based on modality fusion, outperforms many state-of-the-art results on the BU-3DFE dataset as reported in Table 2. For example, [38] employed deep representation, explored geometric and photometric attributes for obtaining 6 different types of 2D facial maps, after which it relied on three pre-trained VGG-Nets for obtaining feature representations from the facial maps; the outputs of the VGG-Nets were then used for training several SVMs. Finally, [38] performed score fusion based on the trained SVMs for FER. We note that despite the elaborate framework proposed in [38], our less complex pipeline based on modality fusion at latent representation level significantly outperforms their reported result; see Table 2.

Furthermore, we observe that [39] reported a test accuracy of 92% based on the combination of 2D and 3D features trained on DCNNs from scratch. However, [39] did not follow any of the well known protocols for training and evaluation on the BU-3DFE dataset; that is, using a set of randomly selected 60 subjects that are either fixed or not in a 10-fold cross validation scheme for training and evaluation. Instead, they consider the whole 100 subjects for training and evaluation, using 90 subjects for training and the remaining 10 subjects for testing. In addition, [39] failed to

Approach	Test acc. (%)
Depth map+pre-trained ResNet50*	61.11
Depth map+pre-trained VGG19*	28.06
Depth map: PL-depth map	84.72
RGB+ pre-trained ResNet50: PL-RGB-ResNet50	82.92
RGB+ pre-trained VGG19: PL-RGB-VGG19	81.25
Depth map+RGB: PL-fusion-ResNet50	87.08
Depth map+RGB: PL-fusion-VGG19	**89.31**

Table 1. Experimental results for the different pipelines

Approach	Test acc. (%)
3D geometric shape model+LDA [17]	83.60
Bayesian Belief net+statistical facial features [35]	82.30
Distance+slopes+SVM [19]	87.10
2D+3D features fusion+SVM [36]	86.32
Geometric scattering representation+SVM [37]	84.80
Geometric+photometric attributes+VGG19 [38]	84.87
Depth map+RGB: PL-fusion-ResNet50 (ours)	**87.08**
Depth map+RGB: PL-fusion-VGG19 (ours)	**89.31**

Table 2. Our best experimental result via depth map and RGB fusion along with state-of-the-art results for comparison

report that they carried out cross validation in their experimental setup; this naturally raises two concerns: (1) a larger dataset has been used for training the proposed model (2) sampling bias for the training and testing data in view of the reported result. However, using the fusion approach that we propose in this paper and the well established training and evaluation protocol for the BU-3DFE dataset (i.e. a smaller training data), we reach a competitive test accuracy of 89.31%.

6. Conclusion

Facial expression recognition is an important domain in human-computer interaction. Endowing computers with the ability to analyse the facial expressions of users should improvement their utility; as such, a computer can adapt its response given the facial expression of the user. In this paper, motivated by the individual strengths of depth map and RGB data for representation information, we propose a facial expression recognition pipeline for joint learning of more robust features. We posit that learning jointly from both depth map and RGB modalities would result in learning more discriminative features as against singly learning from either modalities. We construct three different learning pipelines for learning the BU-3DFE dataset; experimental results validate the effectiveness of the fusion approach over learning facial expression features separately from depth map or RGB data.

Acknowledgments

This work was funded by the National Research Fund (FNR), Luxembourg, under the project reference R-AGR-0424-05-D/Bjorn Ottersten. This work was also supported by the European Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947.

References

[1] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Frame-based facial expression recognition using geometrical features," *Advances in Human-Computer Interaction*, vol. 2014, p. 4, 2014.

[2] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognition*, vol. 48, no. 10, pp. 3191–3202, 2015.

[3] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[4] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of face recognition*. Springer, 2011, pp. 487–519.

[5] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[6] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[9] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[10] O. Beaudry, A. Roy-Charland, M. Perron, I. Cormier, and R. Tapp, "Featural processing in recognition of emotional facial expressions," *Cognition & emotion*, vol. 28, no. 3, pp. 416–432, 2014.

[11] Y. Luo, C.-m. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of pca and lbp with svm," *Optik-International Journal for Light and Electron Optics*, vol. 124, no. 17, pp. 2767–2770, 2013.

[12] E. Owusu, Y. Zhan, and Q. R. Mao, "A neural-adaboost based facial expression recognition system," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3383–3390, 2014.

[13] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[14] N. Sarode and S. Bhatia, "Facial expression recognition," *International Journal on computer science and Engineering*, vol. 2, no. 5, pp. 1552–1557, 2010.

[15] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[16] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition*, vol. 47, no. 3, pp. 1282–1293, 2014.

[17] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.

[18] H. Tang and T. S. Huang, "3d facial expression recognition based on automatically selected features," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[19] ——, "3d facial expression recognition based on properties of line segments connecting facial feature points," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[20] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015.

[21] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[22] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 19–27.

[23] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2207–2216.

[24] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[25] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[26] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[32] F. Ayatollahi, A. A. Raie, and F. Hajati, "Expression-invariant face recognition using depth and intensity dual-tree complex wavelet transform features," *Journal of Electronic Imaging*, vol. 24, no. 2, pp. 023 031–023 031, 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[34] J. Wang, L. Yin, X. Wei, and Y. Sun, "3d facial expression recognition based on primitive surface feature distribution," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1399–1406.

[35] X. Zhao, D. Huang, E. Dellandrea, and L. Chen, "Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3724–3727.

[36] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen, "An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition," *Computer Vision and Image Understanding*, vol. 140, pp. 83–92, 2015.

[37] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–6.

[38] H. Li, J. Sun, D. Wang, Z. Xu, and L. Chen, "Deep representation of facial geometric and photometric attributes for automatic 3d facial expression recognition," *arXiv preprint arXiv:1511.03015*, 2015.

[39] X.-P. Huynh, T.-D. Tran, and Y.-G. Kim, "Convolutional neural network models for facial expression recognition using bu-3dfe database," in *Information Science and Applications (ICISA) 2016*. Springer Singapore, 2016, pp. 441–450.