Deep Covariance Descriptors for Facial Expression Recognition

Naima Otberdout[1]
naima.otberdout@um5s.net.ma
Anis Kacem[2]
anis.kacem@imt-lille-douai.fr
Mohamed Daoudi[2]
mohamed.daoudi@imt-lille-douai.fr
Lahoucine Ballihi[1]
lahoucine.ballihi@um5.ac.ma
Stefano Berretti[3]
stefano.berretti@unifi.it

[1] LRIT - CNRST URAC 29
Rabat IT Center,
Mohammed V University in Rabat
Faculty of sciences, Rabat, Morocco

[2] IMT Lille-Douai, Univ. Lille,
CNRS, UMR 9189 CRIStAL,
Lille, France

[3] Media Int. and Com. Center (MICC),
University of Florence,
Florence, Italy

Abstract

In this paper, covariance matrices are exploited to encode the deep convolutional neural networks (DCNN) features for facial expression recognition. The space geometry of the covariance matrices is that of Symmetric Positive Definite (SPD) matrices. By performing the classification of the facial expressions using Gaussian kernel on SPD manifold, we show that the covariance descriptors computed on DCNN features are more efficient than the standard classification with fully connected layers and softmax. By implementing our approach using the VGG-face and ExpNet architectures with extensive experiments on the Oulu-CASIA and SFEW datasets, we show that the proposed approach achieves performance at the state of the art for facial expression recognition.

1 Introduction

Automatic analysis of facial expressions has been attractive in computer vision research since long time due to its wide spectrum of potential applications that go from human computer interaction to medical and psychological investigations, to cite a few. Similarly to other applications, for many years facial expression analysis has been addressed by designing hand-crafted low-level descriptors, either geometric (*e.g.*, distances between landmarks) or appearance based (*e.g.*, LBP, SIFT, HOG, etc.), with the aim of extracting suitable representations of the face. Higher order relations, like the covariance descriptor, have been also computed on raw data or low-level descriptors. Standard machine learning tools, like SVMs, have then been used to classify expressions. Now, the approach to address this problem has changed quite radically with Deep Convolutional Neural Networks (DCNNs). The idea here is to make the network learn the best features from large collections of data during a training phase. However, one drawback of DCNNs is that they do not take into account the spatial relationships within the face. To overcome this issue, we propose to exploit globally and

locally the network features extracted in different regions of the face. This yields a set of DCNN features per region. The question is how to encode them in a compact and discriminative representation for a more efficient classification than the one achieved globally by classical softmax. In this paper, we propose to encode face DCNN features in a covariance matrix. These matrices have shown to outperform first-order features in many computer vision tasks [23, 24]. We demonstrate the benefits of this representation in facial expression recognition from static images or collections of static peak frames (*i.e.*, frames where the expression reaches its maximum). In doing this, we exploit the space geometry of the covariance matrices as points on the symmetric positive definite (SPD) manifold. Furthermore, we use a valid positive definite Gaussian RBF kernel on this manifold to train a SVM classifier for expression classification. Implementing our approach with different network architectures, *i.e.*, VGG-face [22] and ExpNet [6], and by a thorough set of experiments, we found that the classification of these matrices outperforms the classical softmax.

Overall, the proposed solution permits us to combine the geometric and appearance features enabling an effective description of facial expressions at different spatial levels, while taking into account the spatial relationships within the face. An overview of the proposed solution is illustrated in Figure 1. In summary, the main contributions of our work consist of: *(i)* encoding DCNN features of the face by using covariance matrices; *(ii)* encoding local DCNN features by local covariance descriptors; *(iii)* classifying facial expressions using Gaussian kernel on the SPD manifold; *(iv)* conducting an extensive experimental evaluation with two different architectures and comparing our results with state-of-the-art methods on two publicly available datasets.

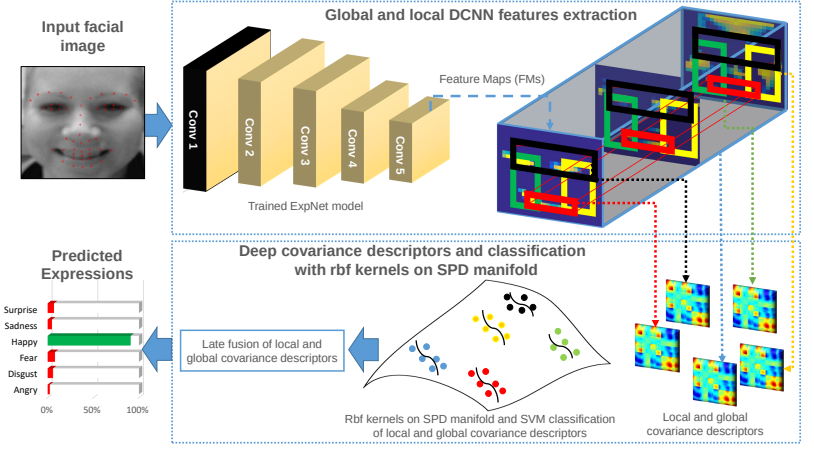

Figure 1: Overview of the proposed method.

The rest of the paper is organized as follows: In Section 2, we summarize the works that are most related to our solution, including facial expression recognition, and covariance descriptors; In Section 3, we present our solution for facial feature extraction and, in Section 4, we introduce the idea of DCNN covariance descriptors for expression classification. A comprehensive experimentation using the proposed approach on two publicly available benchmarks, and comparison with state-of-the-art solutions is reported in Section 5; Finally, conclusions and directions for future work are sketched in Section 6.

2 Related work

The approach we propose in this paper is mainly related to the works on facial expression recognition and those on DCNNs combined with covariance descriptors. Accordingly, we first summarize relevant works using DCNN for facial expression, then we present some recent works that use covariance descriptors in conjunction with DCNN.

DCNN for Facial Expression Recognition: Motivated by the success of DCNN models in facial analysis tasks, several papers proposed to use them for both static and dynamic facial expression recognition [12, 19, 20, 25]. However, the main reason behind the impressive performance of DCNNs is the availability of large-scale training datasets. As a matter of fact, in facial expression recognition, datasets are quite small, mainly for the difficulty of producing properly annotated images for training. To overcome such a problem, Ding et al. [6] proposed *FaceNet2ExpNet*, where a regularization function helps to use the face information to train the facial expression classification net of static images. Facial expression recognition from still images using DCNN was also proposed in [19, 20, 28]. All these methods use a similar strategy in the network architecture: multiple convolutional and pooling layers are used for feature extraction; fully connected ones, and softmax layers are used for classification. In [21], the authors proposed a method for dynamic facial expression recognition that exploits deep features extracted at the last convolutional layer of a trained DCNN. They used a Gaussian Mixture Model (GMM) and Fisher vector encoding on the set of extracted features from videos to get a single vector representation of the video, which is fed into a SVM classifier to predict expressions.

DCNN and Covariance Descriptors: Covariance features were first introduced by Tuzel et al. [23] for texture matching and classification. Bhattacharya et al. [3] constructed covariance matrices, which capture joint statistics of both low-level motion and appearance features extracted from a video. Dong et al. [7] constructed a deep neural network, which embeds high dimensional SPD matrices into a more discriminative low dimensional SPD manifold. In the context of face recognition from image sets, Wang et al. [27] presented a Discriminative Covariance oriented Representation Learning (DCRL) framework to learn better image representations, which can closely match the subsequent image set modeling and classification. The framework constructs a feature learning network (*e.g.*, a CNN) to project the face images into a target representation space. The network is trained with the goal of maximizing the discriminative ability of the set of covariance matrices computed in the target space. In the dynamic facial expression recognition method proposed by Liu et al. [18], deep and hand-crafted features are extracted from each video clip to build three types of image set models, *i.e.*, covariance matrix, linear subspace, and Gaussian distribution. Then, different Riemannian kernels are used, separately and combined, for classification.

To the best of our knowledge, compared to existing literature, our work is the first one that uses covariance descriptors in conjunction with DCNN for static expression recognition.

3 DCNN features

Given a set of n_f face images $\mathcal{F} = \{f_1, f_2, ...f_{n_f}\}$ labeled with their corresponding expressions $\{y_1, y_2, ...y_{n_f}\}$, our goal is to find a high discriminative face representation allowing an efficient matching between faces and their expression labels. Motivated by the success of DCNNs in automatic extraction of non-linear features that are relevant to the problem at hand, we opt for this technique in order to encode the facial expression into Feature Maps

(FMs). A covariance descriptor is then computed over these FMs and is considered for global face representation. We also extract four regions on the input face image around the eyes, mouth, and cheeks (left and right). By mapping these regions on the extracted deep FMs, we are able to extract local regions in these FMs that bring more accurate information about the facial expression. A local covariance descriptor is also computed for each local region.

The first step to our approach is the extraction of non-linear features that encode well the facial expression in the input face image. In this work, we use two DCNN models, namely, *VGG-face* [22] and *ExpNet* [6].

3.1 Global DCNN features

VGG-Face is a DCNN model that is commonly used in facial analysis tasks. It consists of 16 layers trained on 2.6M facial images of 2.6K people for face recognition in the wild. This model has been also successfully used for expression recognition [6]. However, the model was trained for face identification, so it is expected to also encode information about the identity of the persons that should be filtered-out in order to capture person-independent facial expressions. This may deteriorate the discrimination of the expression model after fine-tuning, especially when it comes to dealing with small datasets, which is quite common in facial expression recognition. To tackle this problem, Ding et al. [6] proposed ExpNet, which is a much smaller network containing only five convolutional layers and one fully connected layer. The training of this model is regularized by the VGG-face model.

Following Ding et al. [6], we first fine-tune the VGG-face network on expression datasets by minimizing the cross entropy loss. This fine-tuned model is then used to regularize the ExpNet model. Because we are interested in facial feature extraction, we only consider the FMs at the last convolutional layer of the ExpNet model. In what follows, we will denote the set of extracted FMs from an input face image f as $\Phi(f) = \{M_1, M_2, \cdots, M_m\}$, where $\{M_i\}_{i=1}^m$ are the m FMs at the last convolutional layer, and $\Phi(.)$ is the non-linear function induced by the employed DCNN architecture at this layer.

3.2 Local DCNN features

In addition to using the global feature map $\Phi(f)$, we focus on specific regions extracted from this global feature map that are relevant for face expression analysis. To do so, we start by detecting a set of landmark points on the input facial image using the method proposed in [2]. Four regions $\{R_j\}_{j=1}^4$ are then constructed around the eyes, mouth, and both cheeks using these points. By defining a pixel-wise mapping between the input face image and its corresponding FMs, we map the detected regions from the input face image to the global FMs. Indeed, a feature map M_i is obtained by convolution of the input image with a linear filter, adding a bias term and then applying a non-linear function. Accordingly, units within a feature map will be connected to different regions R_j on the input image. Based on this assumption, we can find a mapping between the coordinates of the input image and those of the output feature map. Specifically, for each point p of coordinates (x_p, y_p) in the input face image f, we associate a feature $\phi_p(f, M_i)$ in the feature map M_i such that,

$$\phi_p(f, M_i) = M_i(\overline{s \times x_p}, \overline{s \times y_p}) \,, \tag{1}$$

where s is the map size ratio with respect to input size, and $\overline{(.)}$ is the rounding operation. It is worth noting that for both models used in this work, the input image and output maps have

the same spatial extent. This is important to map landmarks position in the input image to the coordinates of convolutional feature maps. Using this pixel-wise mapping, we map each region R_j formed by r pixels $\{p_1, p_2, \cdots, p_r\}$ on the input image into the global FMs $\{M_i\}_{i=1}^{m}$ to obtain the corresponding local FMs $\Phi^{R_j}(f) = \{\phi_{p_1}(f, M_i), \phi_{p_2}(f, M_i), \cdots, \phi_{p_r}(f, M_i)\}_{i=1}^{m}$.

4 DCNN based covariance descriptors

Both our local and global non-linear features $\Phi(f)$ and $\{\Phi^{R_j}(f)\}_{j=1}^{4}$ can be directly used to classify the face images. However, motivated by the great success of covariance matrices in various recent works, we propose to compute covariance descriptors using these global and local features. In particular, a covariance descriptor is computed for each region R_j across the corresponding local FMs $\Phi^{R_j}(f)$ yielding four covariance descriptors. A covariance descriptor is also computed on the global FMs $\Phi(f)$ extracted from the whole face f. In this way, we encode the correlation between the extracted non-linear features within different spatial levels, which results in an efficient, compact and more discriminative representation. Furthermore, covariance descriptors allow us to select local features and focus on local facial regions, which is not possible with fully connected and softmax layers. We can also note that the covariance descriptors are treated separately, then lately fused in the classifier. In what follows, we describe the processing for the global features $\Phi(f)$; the same steps hold for the covariance descriptors computed over the local features.

The extracted features $\Phi(f)$ are arranged in a $(m \times w \times h)$ tensor, where w and h denote the width and height of the feature maps, respectively, and m is their number. Each feature map M_i is vectorized into a n-dimensional vector with $n = w \times h$ to transform the input tensor to a set of n observations stored in the matrix $[v_1, v_2, ..., v_n] \in \mathbb{R}^{m \times n}$. Each observation $\{v_i\}_{i=1}^{n} \in \mathbb{R}^m$ encodes the values of the pixel i across all the m feature maps. Finally, we compute the corresponding $(m \times m)$ covariance matrix,

$$C_{\Phi(f)} = \frac{1}{n-1} \sum_{i=1}^{n} (v_i - \mu)(v_i - \mu)^T \,, \tag{2}$$

where μ is the mean of the feature vectors such that $\mu = \frac{1}{n} \sum_{i=1}^{n} v_i$. Covariance descriptors are mostly studied under a Riemannian structure of the space of symmetric positive definite matrices $Sym^{++}(m)$ [10, 23, 27]. Several metrics have been proposed to compare covariance matrices on $Sym^{++}(m)$, the most widely used is the Log-Euclidean Riemannian Metric (LERM) [1] since it has excellent theoretical properties with simple and fast computations. Formally, given two covariance descriptors $C_{\Phi(f_1)}$ and $C_{\Phi(f_2)}$ of two images f_1 and f_2, their log-Euclidean distance $d : (Sym^{++}(m) \times Sym^{++}(m)) \to \mathbb{R}^+$ is given by,

$$d(C_{\Phi(f_1)}, C_{\Phi(f_2)}) = \| \log(C_{\Phi(f_1)}) - \log(C_{\Phi(f_2)}) \|_{\mathrm{F}} \,, \tag{3}$$

where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm, and $\log(.)$ is the matrix logarithm.

4.1 RBF Kernels for DCNN covariance descriptors classification

As discussed above, each face f is represented by its global and local covariance descriptors that lie on the non-linear manifold $Sym^{++}(m)$. The problem of recognizing expressions from facial images is then turned to classifying their covariance descriptors in $Sym^{++}(m)$. However, one should take into account the non-linearity of this space, where traditional machine

learning techniques cannot be applied in a straightforward way. Accordingly, we exploit the the log-Euclidean distance mentioned in Eq. (3) between symmetric positive definite matrices to define the Gaussian RBF kernel $K : (Sym^{++}(m) \times Sym^{++}(m)) \rightarrow \mathbb{R}^+$,

$$K(C_{\Phi(f_1)}, C_{\Phi(f_2)}) = \exp(-\gamma d^2(C_{\Phi(f_1)}, C_{\Phi(f_2)})) \,, \tag{4}$$

where $d(C_{\Phi(f_1)}, C_{\Phi(f_2)})$ is the log-Euclidean distance between $C_{\Phi(f_1)}$ and $C_{\Phi(f_2)}$. Conveniently for us, this kernel has been already proved to be a positive definite kernel for all $\gamma > 0$ [10]. This kernel is computed for the global covariance descriptor as well as for each local covariance descriptor yielding to five different kernels. Then, each kernel is fed, separately, to a SVM classifier that outputs a score per class. Finally, fusion is performed by multiplying or computing a weighted sum over the scores given by the different kernels.

5 Experimental results

The effectiveness of the proposed approach in recognizing basic facial expressions has been evaluated in constrained and unconstrained (*i.e.*, in-the-wild) settings using two publicly available datasets with different challenges:

Oulu-CASIA dataset [29]: Includes 480 image sequences of 80 subjects taken in a constrained environment with normal illumination conditions. For each subject, there are six sequences, one for each of the six basic emotion labels. Each sequence begins with a neutral facial expression and ends with the apex of the expression. For both training and testing, we use the last three peak frames to represent the video resulting in 1440 images. Following the same setting of the state-of-the-art, we conducted a ten-fold cross validation experiment, with subject independent splitting.

Static Facial Expression in the Wild (SFEW) dataset [5]: Consists of $1,322$ static images labeled with seven facial expressions (the six basic plus the neutral one). This dataset has been collected from real movies and targets spontaneous expression recognition in challenging, *i.e.*, in-the-wild, environments. It is divided into training (891 samples), validation (431 samples), and test sets. Since the test labels are not available, here we report results on the validation data.

5.1 Settings

As initial step, we performed some preprocessing on the images of both datasets. For Oulu-CASIA, we first detected the face using the method proposed in [26]. For SFEW, we used the aligned faces provided by the dataset. Then, we detected 49 facial landmarks on each face using the Chehra Face Tracker [2]. All frames were cropped and resized to 224×224, which is the input size of the DCNN models.

VGG fine-tuning: Since the two datasets are quite different, we fine-tuned the VGG-face model on each dataset separately. To keep the experiments consistent with [6] and [21], we conducted ten-fold cross validation on Oulu-CASIA. This results in ten different deep models, each of them is trained on nine splits with $9 \times 3 \times (480/10) = 1,296$ images. On the SFEW dataset, one model is trained using the provided training data. The training procedure for both datasets is executed for 100 epochs, with a mini-batch size of 64 and learning rate of 0.0001 decreased by 0.1 after 50 epochs. The momentum is fixed to be 0.9, and Stochastic Gradient Descent is adopted as optimization algorithm. The fully connected layers of the VGG-face model are trained from scratch by initializing them with a Gaussian distribution.

For data augmentation, we used horizontal flipping on the original data without any other supplementary datasets.

ExpNet training: Also in this case, a ten-fold cross validation is performed on Oulu-CASIA requiring the training of ten different deep models. The ExpNet architecture consists of five convolutional layers, each one followed by Relu activation and max pooling [6]. As mentioned in Section 3.1, these layers were trained first by regularization with the fine-tuned VGG model, then we appended one fully connected layer of size 128. The whole network is finally trained. All parameters used in the ExpNet training (learning rate, momentum, mini-batch size, number of epochs) are the same as in [6]. We conducted all our training experiments using the Caffe deep learning framework [11].

Features extraction: We used the last pooling layer of DCNN models to extract features from each face image. This layer provides 512 feature maps of size 7×7, which yields to covariance descriptors of size 512×512. For the local approach, to well map landmarks position in the input image to the coordinates of the feature maps, we resized all feature maps to 14×14, that allows us to correctly localize regions on the feature maps and minimize the overlapping between them. The detected regions in the input image were mapped to the feature maps using Eq. (1) with a ratio $s = 1/16$. Based on this mapping, we extracted features around eyes, mouth and both cheeks from each feature map. Finally, we used these local features to compute a covariance descriptor of size 512×512 for each region in the input image. It is worth noting that the extracted regions have different sizes in different images. However, the size of the resulting covariance matrices depends only on the number of feature maps (as results from Eq. (2)). This yields covariance matrices of the same size lying in the same SPD manifold $Sym^{++}(512)$, without the necessity of any resizing that can change our DCNN features. In Sections 1 and 2 of the supplementary material, we show images of the extracted global and local FMs and their corresponding covariance matrices.

Classification: For the global approach, each static image is represented by a covariance descriptor of size 512×512. In order to compare covariance descriptors in $Sym^{++}(512)$, it is empirically necessary to ensure their positive definiteness by using their regularized version, $C_{\Phi(f)} + \varepsilon I$, where ε is a small regularization parameter (set to 0.0001 in all our experiments), and I is the 512×512 identity matrix. To classify these descriptors, we used multi-class SVM with Gaussian kernel on the Riemannian manifold $Sym^{++}(512)$. For reproducibility, we choose parameters of the Gaussian kernel γ and SVM cost δ using cross validation with grid search in the following intervals: $\gamma \in [10^{-3}, 10^{-10}]$ and $\delta \in [10^{3}, 10^{8}]$. Concerning the local approach, each image was represented by four covariance descriptors, each regularized as stated for the global covariance descriptor. This resulted in four classification decisions that were combined using two late fusion methods: *weighted sum* and *product*. The best performance were achieved for weighted sum fusion with w_{global}, w_{eyes}, $w_{cheek-left}$, $w_{cheek-right}$ equal to 1 and $w_{mouth} = 0.2$, for the Oulu-CASIA dataset, and $w_{global} = 1$, and w_{eyes}, w_{mouth}, $w_{cheek-left}$, $w_{cheek-right}$ equal to 0.1 for the SFEW dataset. Note that we report the results of our local approach with only ExpNet model since it provides better results with the global approach than VGG-face model. SVM classification was obtained using the LIBSVM [4] package. Note that for testing the Oulu-CASIA dataset, we represented each video by its three peak frames as in Ding et al. [6]. Hence, to calculate the distance between two videos, we considered the mean of the distances between their frames. For softmax, we considered the video as correctly classified if its three frames are correctly recognized by the model.

5.2 Results and discussion

As first analysis, in Table 1, we compare our proposed global (G-FMs) and local (R-FMs) solutions with the baselines provided by the VGG-face and ExpNet models, without the use of the covariance matrix (*i.e.*, they used the fully connected and softmax layers). On Oulu-CASIA, the G-FMs solution improves by 3.7% and 1.26%, respectively, the VGG-face and ExpNet models. Though less marked, an increment of 0.69% for the VGG-face and of 0.92% for ExpNet has been also obtained on the SFEW dataset. These results prove that the covariance descriptors computed on the convolutional features provide more discriminative representations. Furthermore, the classification of these representations using Gaussian kernel on SPD manifold is more efficient than the standard classification with fully connected layers and softmax, even if these layers were trained in an end-to-end manner. Table 1 also shows that the fusion of the local (R-FMs) and global (G-FMs) approaches achieves a clear superiority on the Oulu-CASIA dataset surpassing by 1.25% the global approach, while no improvement is observed on the SFEW dataset. This is due to the failure of landmark detection skewing the extraction of the local deep features. In Section 3 of the supplementary material, we show some failure cases of landmark detection on this dataset.

Dataset	Model	FC-Softmax	ours (G-FMs)	ours (G-FMs and R-FMs)
Oulu-CASIA	*VGG Face*	77.8	81.5	–
	ExpNet	82.29	**83.55**	**84.80**
SFEW	*VGG Face*	46.66	47.35	–
	ExpNet	48.26	**49.18**	**49.18**

Table 1: Comparison of the proposed classification scheme with respect to the VGG-Face and ExpNet models with fully connected layer and Softmax.

In Table 2, we investigated the performance of the individual regions of the face for ExpNet. On both datasets, the right and left cheek provide almost the same score outperforming at a large extent the mouth score. Results for the eye region are not coherent across the two datasets: the eyes region is the best performing for Oulu-CASIA, but this is not the case on SFEW. We motivate this result by the fact that, in the wild acquisitions as for the SFEW dataset, the region of the eyes can be affected by occlusions, and the landmarks detection can be less accurate (see Section 3 of the supplementary material for failure cases of landmark detection in this dataset). Table 2 also compares different fusion modalities. We found consistent results across the two datasets, indicating the weighted sum fusion between G-FMs and R-FMs is the best approach.

The confusion matrices for ExpNet with weighted-sum are reported in Figure 2 left and right plots, respectively, for Oulu-CASIA and SFEW. For Oulu-CASIA, the happy and surprise expressions are better recognized over the rest. The happy expression is the best recognized one also for SFEW, followed by the neutral one, while surprise, disgust and fear expressions are harder to recognize. This is related to the unbalanced number of expression examples for the different classes included in this database as explained in [20].

As last analysis, in Table 3 we compare our solution with respect to state-of-the-art methods. Overall, on Oulu-CASIA, we obtained the second highest accuracy, outperforming several recent approaches. Furthermore, Ofodil et al. [21], who achieved the highest accuracy on this dataset, also used temporal information of the video. In addition, they did not report the frames used to train their DCNN model, which is indeed an important information to compare the two approaches. Note that, to compare our results with those of Ding et al. [6],

Region	Oulu-CASIA	SFEW
Eyes	84.59	38.05
Mouth	70.00	38.98
Right Cheek	83.96	43.16
Left Cheek	83.12	42.93
R-FMs product fusion	83.66	42.92
G-FMs and R-FMs product fusion	84.05	45.24
R-FMs weighted-sum fusion	84.59	43.85
G-FMs and R-FMs weighted-sum fusion	**84.80**	**49.18**

Table 2: Overall accuracy (%) of different regions and fusion schemes on the Oulu-CASIA and SFEW datasets for the ExpNet model.

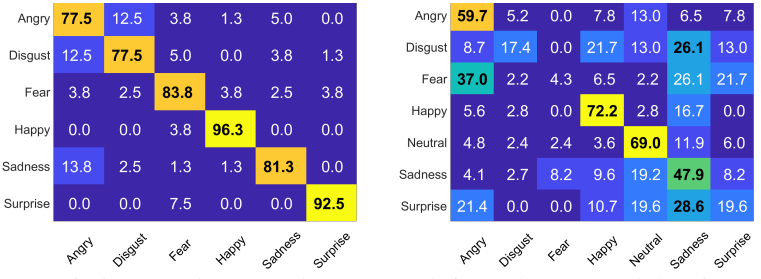

Figure 2: Confusion matrix on Oulu-CASIA (left) and SFEW (right) for ExpNet with weighted-sum fusion.

Method	Oulu-CASIA	SFEW
*Kacem et al. [13] **	83.13	–
*Jung et al. [12] **	74.17	–
Liu et al. [16]	–	26.14
Levi et al. [15]	–	41.92
Mollahosseini et al. [19]	–	47.70
Ng et al. [20]	–	48.50
Yu et al. [28]	–	52.29
Ding et al. [6]	82.29	48.29
*Liu et al. [17] **	74.59	–
*Guo et al.[9] **	75.52	–
*Zhao et al. [30] **	84.59	–
*Jung et al. [12] **	81.46	–
*Ofodil et al. [21] **	89.60	–
ours (ExpNet + G-FMs)	**83.55**	**49.18**
ours (ExpNet + G-FMs and R-FMs fusion)	**84.80**	**49.18**

Table 3: Comparison with state-of-the art solutions on Oulu-CASIA and SFEW. Geometric, appearance and hybrid solutions are reported in the first three groups of methods; Our solutions are given in the last row. (*) Dynamic approaches.

which was reported per frames, we reproduced the results for their approach on a per video basis, considering that the video is correctly classified if the three frames of the video are correctly recognized. On the SFEW dataset, the global approach achieves the second highest accuracy, surpassing various state of the art methods with significant gains. Moreover, the

highest accuracy reported by [28] is obtained using a DCNN model trained on more than 35,000 additional data provided by the FER-2013 database [8]. As reported in [6], this data augmentation can improve results on SFEW from 48.29% to 55.15%.

6 Discussion and Conclusions

In this paper, we have proposed the covariance matrix descriptor as a way to encode DCNN features in facial expression recognition. In the general approach, DCNNs are trained to automatically identify the patterns that characterize each class in the input images. For the case of facial expression recognition, these patterns correspond to high-level features that are related to Facial Action Units [14] (in the supplementary material, we have shown some examples of the extracted features at the top of convolutional layers of a trained DCNN model). Following a standard classification scheme in DCNN models, these features are firstly flattened using fully connected layers by performing a set of linear combinations of the input features followed by a softmax activation for predicting the expression. By contrast, in this work, we discard the fully connected layers and use covariance matrices to encode all the linear correlations between the activated non-linear features at the top convolutional layers. This is achieved both globally and locally by focusing on specific regions of the face. By doing so, we exploit locally and globally both first-order statistics information (deep features) and second-order information (covariance), which results in a more discriminative representation. More particularly, the covariance matrix belongs to the set of symmetric positive-definite (SPD) matrices, thus laying on a special Riemannian manifold. We have shown the classification of these representations using Gaussian kernel on the SPD manifold is more efficient than the standard classification with fully connected layers and softmax. By implementing our approach using different architectures, *i.e.*, VGG-face and ExpNet, in extensive experiments on the Oulu-CASIA and SFEW datasets, we have shown that the proposed approach achieves state-of-the-art performance for facial expression recognition. As future work, we aim to include the temporal dynamics of the face in the proposed model.

Acknowledgements

The authors thank the high performance computing center of the Université de Lille for computational facilities.

References

[1] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1859–1866, 2014.

[3] Subhabrata Bhattacharya, Nasim Souly, and Mubarak Shah. Covariance of motion and appearance featuresfor spatio temporal recognition tasks. *CoRR*, abs/1606.05355, 2016.

[4] Chih-chung Chang and Chih-jen Lin. Libsvm: A library for support vector machines. software available at http://www. csie. ntu. edu. tw/cjlin/libsvm. 2001.

[5] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ACM Int. Conf. on Multimodal Interaction*, pages 423–426. ACM, 2015.

[6] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *IEEE Int. Conf. on Automatic Face Gesture Recognition (FG)*, pages 118–126, 2017.

[7] Zhen Dong, Su Jia, Chi Zhang, Mingtao Pei, and Yuwei Wu. Deep manifold learning of symmetric positive definite matrices with application to face recognition. In *AAAI Conf. on Artificial Intelligence*, pages 4009–4015, 2017.

[8] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Int. Conf. on Neural Information Processing (NIPS)*, pages 117–124. Springer, 2013.

[9] Yimo Guo, Guoying Zhao, and Matti Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *European Conf. on Computer Vision (ECCV)*, volume 7573 of *Lecture Notes in Computer Science*, pages 631–644. Springer, 2012.

[10] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(12):2464–2477, 2015.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Int. Conf. on Multimedia*, pages 675–678. ACM, 2014.

[12] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision, ICCV*, pages 2983–2991, 2015.

[13] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, and Juan Carlos Álvarez Paiva. A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3199–3208, 2017.

[14] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, pages 19–27, 2015.

[15] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *ACM Int. Conf. on Multimodal Interaction*, pages 503–510. ACM, 2015.

[16] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[17] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1749–1756, 2014.

[18] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *ACM Int. Conf. on Multimodal Interaction*, pages 494–501. ACM, 2014.

[19] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[20] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ACM Int. Conf. on Multimodal Interaction*, pages 443–449. ACM, 2015.

[21] Ikechukwu Ofodile, Kaustubh Kulkarni, Ciprian Adrian Corneanu, Sergio Escalera, Xavier Baro, Sylwia Hyniewska, Juri Allik, and Gholamreza Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061*, 2017.

[22] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conf. (BMVC)*, pages 41.1–41.12. BMVA Press, 2015.

[23] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conf. on Computer Vision (ECCV)*, pages 589–600, 2006.

[24] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30 (10):1713–1727, Oct. 2008.

[25] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, Dec 2017. doi: 10.1109/JSTSP.2017.2764438.

[26] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal on Computer Vision*, 57(2):137–154, 2004.

[27] Wen Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Discriminative covariance oriented representation learning for face recognition with image sets. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5599–5608, 2017.

[28] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *ACM Int. Conf. on Multimodal Interaction*, pages 435–442, 2015.

[29] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9): 607–619, 2011.

[30] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European Conf. on Computer Vision (ECCV)*, pages 425–442. Springer, 2016.