Expression Recognition with Deep Features Extracted from Holistic and Part-based Models

S L Happy, Antitza Dantcheva, Francois Bremond

Expression Recognition with Deep Features Extracted from Holistic and Part-based Models

S L Happy, Antitza Dantcheva, François Bremond

INRIA Sophia Antipolis, France
e-mails: s-l.happy@inria.fr;antitza.dantcheva@inria.fr;francois.bremond@inria.fr

Abstract

Facial expression recognition aims to accurately interpret facial muscle movements in affective states (emotions). Previous studies have proposed holistic analysis of the face, as well as the extraction of features pertained only to specific facial regions towards expression recognition. While classically the latter have shown better performances, we here explore this in the context of deep learning. In particular, this work provides a performance comparison of holistic and part-based deep learning models for expression recognition. In addition, we showcase the effectiveness of skip connections, which allow a network to infer from both low and high-level feature maps. Our results suggest that holistic models outperform part-based models, in the absence of skip connections. Finally, based on our findings, we propose a data augmentation scheme, which we incorporate in a part-based model. The proposed multi-face multi-part (MFMP) model leverages the wide information from part-based data augmentation, where we train the network using the facial parts extracted from different face samples of the same expression class. Extensive experiments on publicly available datasets show a significant improvement of facial expression classification with the proposed MFMP framework.

Keywords: Facial expression recognition, Convolutional neural networks, Part-based face representation, Data augmentation

1. Introduction

Facial expressions constitute a pertinent human nonverbal channel for communicating emotions [1]. Ekman and Friesen have broadly analyzed this channel from psychological point of view and have postulated the universality of neutral and six prototypical human facial expressions, namely: anger, disgust, fear, happiness, sadness, and surprise [2]. A growing literature has addressed the recognition of these facial expressions in the contexts of human computer interaction, biometrics, digital entertainment, health-care, as well as robotics [3]. The performance of an automated facial expression recognition (FER) system has been a function of the underlying representations. While early approaches were based on hand-crafted features representing the whole face, these were gradually replaced by representations of salient facial regions, improving the FER-performance [4, 5, 6, 7, 8]. Two major observations drove this development: (a) facial expressions are predominantly reflected in certain regions of face, i.e., lip and periocular regions; (b) facial expressions can be associated with action units (AUs), representing movement of facial muscles, as described in the facial action coding system (FACS) [9]. Thus, each basic expression comprises the joint movement of multiple AUs. The AUs occurring around lips and periocular region region have primarily been utilized to recognize emotions [10].

Emerging deep learning based approaches have significantly advanced research and the related performances in FER [11, 12, 13, 14]. While deep architectures proposed for FER have focused on the representation of the whole face, part-based models have received limited attention. Motivated by the above, we here investigate and compare part-based deep learning models vs. holistic models. While it is agreed that part-based models outperform holistic hand-crafted models, the selection of suitable facial regions for FER has not been consented [4, 6, 7, 8]. However, the most prominent local regions have been lips, eyes, and eyebrows. Specifically, we here select the periocular and mouth regions, which have been shown to be prominent for FER [3, 4], and proceed to investigate the performance of a deep two-part model for FER. Further, we analyze the effect of skipped connections and the benefits of facial patch-based architectures based on the established VGG-Face model [15] for FER. While the first few layers of CNN architectures encode low-level features, the layers towards the end encode high-level information. Skip connections allow the network to infer both low and high-level feature maps, which as we show in this work, has a significant impact on expression recognition performance.

Finally, this paper proposes a data augmentation scheme that we incorporate in a part-based model, which we refer to as multi-face multi-part (MFMP) model. Data augmentation and specifically MFMP seeks to overcome a challenge related to the limited dataset size, which is known to negatively affect the learning process of deep models. While translation, rotation, skew, flip, and perturbation has been often utilized in data augmentation [16], we here propose to train the network using the facial parts extracted from different face samples of the same class. To be specific, during training of our two part model, we use the periocular region from one face-sample, while the mouth region comes from another face-sample of the same ex-

pression class. Thus, the MFMP network is trained with different facial parts, mitigating spatial correlation on the input side. This allows the model to learn the expression representation by finding the commonality between different parts of different instances, improving the performance. Such data augmentation scheme can only be utilized in part-based models, thereby leveraging the part-specific information of face to predict the expression.

Contributions This paper has following contributions.

- We compare the performance of a set of variants of VGG-16-based holistic and part-based models.
- We analyze the effect of skip connections in the analyzed models, allowing for combinations of low and high-level feature maps.
- We propose a novel data augmentation scheme MFMP, which can be instrumental in other object classification frameworks where object alignment can be achieved.
- Extensive experiments were carried out on in-the-lab (CK+, RaFD, FACES, lifespan) and in-the-wild (RAF, AffectNet) datasets, which validates the effectiveness of skipped connections and the proposed data augmentation scheme.

2. Related Work

A growing FER literature suggests that research in FER is on the rise. We revisit here briefly hand-crafted methods and proceed to review deep feature extraction, focusing on most recent transfer learning approaches employed in FER for improving accuracy. Then, we discuss the part-based classification methods in deep learning domain and their pertinence for face representation.

Facial expression recognition (FER) Feature extraction methodologies and classifiers used for FER to encode the appearance and geometrical changes of expressions have been broadly reviewed in a survey by Sariyanidi *et al.* [3]. In summary, while spatial feature extractors can be categorized as high-level, low-level, and hierarchical features, high-level representations, such as NMF [17] and sparse coding [18, 19], aim to encode semantically interpretable traits.On the other hand, low-level features – like local binary patterns, histogram of oriented gradients, local phase quantization, local directional patterns etc., [8, 20, 21, 22, 23] – generally encode the local edge patterns into a global representation by pooling local histograms of each region. Though low-level representations are relatively less sensitive to illumination variation and registration errors, they can be affected by identity and demographic bias. Hierarchical feature representation benefits from both low and high-level representation and is more robust to registration errors, partial occlusions, and bias factors [3].

Deep learning methods for FER CNNs have excelled hierarchical feature extraction and while hand-crafted features dominated FER previously, recently researchers have been focusing on deep learning methods for FER [16, 24, 25, 26, 27, 28, 29, 30].Notably, Jung *et al.* [26] attempted to encode temporal appearance and geometry features in a CNN framework. Boosted Deep Belief Network [31] improved the expression recognition performance by jointly learning the feature representation, feature selection, and classification. Peak-piloted deep network [32] implicitly embedded facial representation of both, peak and non-peak expression frames. Identity-aware CNN [33] jointly learned expression and identity related features to improve person independent recognition performance. Conditional convolution neural networks enhanced random forest was proposed [13], which used salient features from salient facial patches for FER in unconstrained scenarios including pose variations and occlusions.

Transfer-learning The lack of sufficient annotated data has been a major challenge in expression recognition. Early developed datasets, such as CK+ and JAFFE [34], have less than 400 samples per six basic expression categories. Irrespective of the difficulty, manual labour, and time-consuming nature of emotion database creation, databases with more than a thousand images have emerged [35, 36, 37]. However, with such limited data, machine learning algorithms, especially the deep learning models failed to train well. Transfer learning [38] has proved to be an efficient choice in such scenarios by which the parameters learned with large data of related task was used as the network initialization. Ng *et al.* [39] observed the improvement in expression classification by using the model weights trained on ImageNet. FaceNet2ExpNet [11] fine-tuned the FaceNet model to capture high level expression semantics. For expression recognition, the models trained with faces for face recognition are more suitable. The same work suggests the use of pre-trained VGG-Face model [15] achieves promising performance due to transfer learning. VGG-Face uses a VGG-16 architecture and was trained with 2.6 million face images for face recognition application. Instead of learning the model parameters from scratch, the pre-trained weights of VGG-Face is a popular choice of researchers [11, 12, 13, 40] for FER applications.

Part-based classification Object alignment is a critical factor in object classification. In part-based classification, each part of the object is treated independently for global representation, which makes the system less sensitive to registration issues. Recognizing fine-grained categories (like bird species) is challenging because of the lack of suitable learning method to discriminate the fine-grained local features. Using parts of an object instead of the whole image [41, 42, 43, 44] has substantially improved the performance of visual recognition tasks. Lin *et al.* [45] proposed a fine-grained object recognition system by incorporating localization and alignment of object parts in a single deep learning framework. Attention mechanism is used in [46] to recursively learn the discriminative regions followed by region-based feature representation for classification. A weakly supervised method for obtaining part attention mechanism is proposed in [47]. Both object-level and part-level features are integrated in [48] to boost the recognition performance. Similar part-based models are also popular in face detection, face recognition [49, 50], head pose estimation [51] and gender classification [7, 51].

Part-based face representation Due to the success of part-based learning, a research trend involved part-based models [4, 5, 7, 8, 52] in FER. Deviating from encoding faces as a whole, part-based representation processes each registered facial regions independently and combines the final representation from multiple parts as the global face descriptor. Thus, such models are robust against head pose variation and registration errors. The efforts of [5, 22, 52], and [53] involve the division of face into multiple blocks and encoding face as the concatenation of individual block representation. In FACS, each facial muscle movements was associated with an AU and the combination of certain AUs was considered for expression classification. Enhancing and cropping network was proposed by Li *et al.* [54] which provided attention to facial regions based on facial landmarks and individual CNN layers were used to learn the patterns of each facial region separately to recognize action units. Specific facial regions were targeted for recognizing these AUs based on domain knowledge [10]. This implied that the facial expression could be inferred directly by processing these areas without AU classification. The works in [4, 6, 7, 8, 55] were focused on extracting features from facial regions dynamically around the mouth and eyes, where the emotion change was prominently observed. Particularly the salient expressive regions were defined based on domain knowledge (FACS) and they were localized based on the detected facial landmarks. Patch-Gated CNN [14] inspects several patches from the intermediate feature map using the landmarks and assigned a weight to each patch according to its importance, thereby making the system aware of facial occlusions.

3. Holistic and Part-based Deep Face Models

We investigate the effectiveness of the features of the whole face (holistic features) and part-based features in deep learning scenario for FER. During our experiments, the skipped connections found to be an important technique for performance improvement. Therefore, this section describes four CNN models, which are VGG-16 variants to evaluate the effectiveness of holistic/part-based models with/without skipped connections in FER.

Model M1 To achieve transfer learning, many researchers [11][12][40] use pre-trained VGG-Face model [15] for expression recognition. VGG-Face has 13 convolutional layers and three fully connected layers followed by the softmax layer. Since its weights are trained for face recognition, its face representation supposedly perform well for any face analysis application including expression classification. In our experiments, the last fully connected layer is replaced by the expression classification layer with the number of units equivalent to the number of expression classes as shown in Fig. 1a. The following conventions are used in the network definition: "3×3 conv, 64" : 64 convolutional filters of size 3×3; "pool $(2 \times 2)/2$: pooling filters of size 2×2 with stride 2; and "fc, 4096" : fully connected layer with 4096 units.

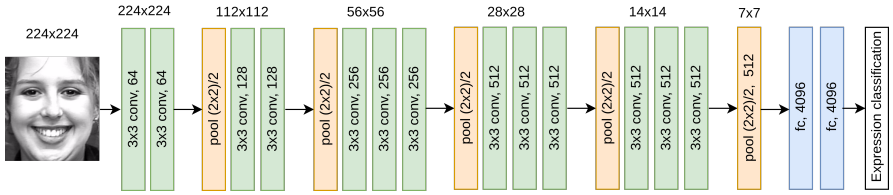
Model M2 Different literature suggests extracting features from different facial regions [7, 4, 8, 6] of different sizes for accurate expression recognition. These regions are mostly around the eyes and lips, which is also supported by FACS [9]. Thus, instead of investigating the size, number, and location of facial regions, we use large facial regions around both lips and eyes so that most of the features are retained for expression classification. In our experiments, peri-ocular region and mouth region are used for part-based model construction.

The eyebrows move with the presence of AU 1, 2 and 4. Thus, selecting a region around the eyebrow corners does not make sense as its position is not intact. Rather the eye corners are fixed and can be considered as reference points. The relative position of eyebrows from these reference points can be encoded in a feature descriptor for improved expression recognition. Similarly, the raised cheek (AU 6 and 12) can also be identified with reference to the position of eye. Therefore, we choose the eye region along with the eyebrow and cheek portion as one region of interest. Similarly, we cropped a large region around lips including lips corners, chin, and some portion of cheek, which contain most information pertaining to expression classes.

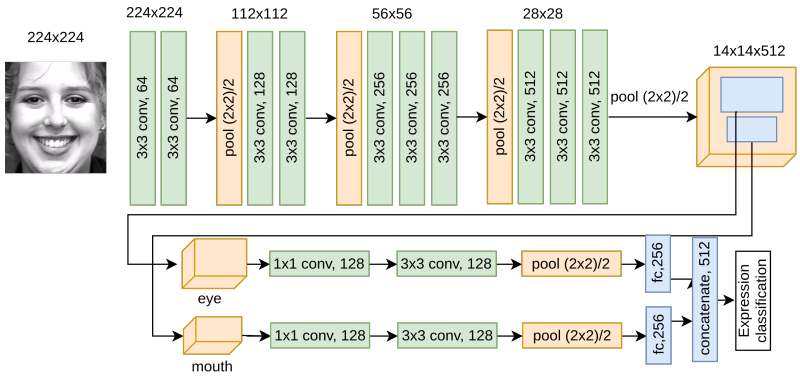
These regions can be cropped and processed independently with a CNN model for feature extraction. To do so, we need to train a new CNN architecture from scratch based on the shape of these regions, without the benefit of transfer learning that we expect from a pre-trained model. Therefore, we processed the whole image through the VGG-Face model and cropped the feature representation of the corresponding facial regions from the output feature map. In order to accurately obtain the feature representations in higher order layers of CNN, we first aligned each face image using the position of eyes and nose. Both the eyes were positioned at a fixed distance parallel to the horizontal axis using image transformations and then re-sized to 224×224 resolution facilitating the input to the pre-trained VGG-face model. Since the location of eyes and nose does not change with expressions, we assume that the facial region representations are the higher layer feature maps in corresponding fixed positions. As can be seen in Fig. 1b, the feature maps from the fourth pooling layer are used for further processing of individual patch.

The CNN architecture in M2 (refer Fig. 1b) is based on the VGG-16 with two parallel cropped branches (for mouth and peri-ocular region) from the fourth pooling layer. Further, each branch is passes through 1×1 and 3×3 convolutional layer each with 128 filters followed by a max polling and fully connected layer. The 1×1 layer helps in reducing the number channels of the feature maps. Since the 3×3 weights are not shared between the parallel branches, they supposedly learn different activation patterns for different facial regions. The pooled features are flattened and fed to a fully connected layer with 256 units. In this network, we use Rectified Linear Unit (ReLU) as the activation function. The output from each stream is further concatenated and fed to the classification layer.

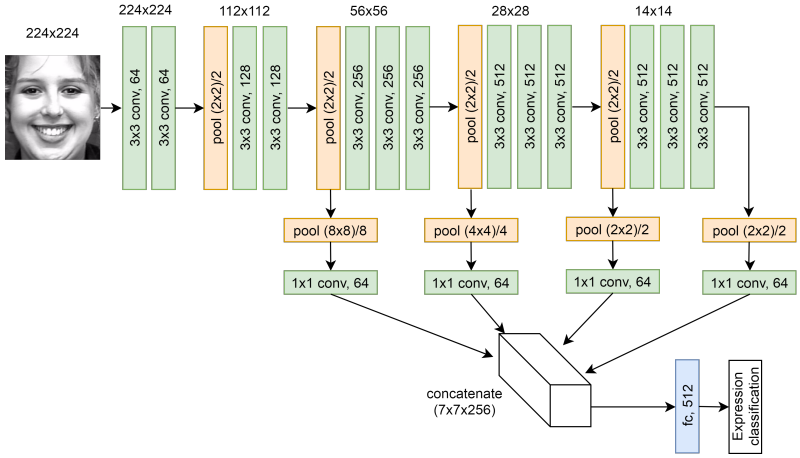
Model M3 Being an hierarchical feature extraction method, CNN learns low-level features in the early layers while high level features are learned in the layers toward the end of the pipeline. The local feature extraction techniques were very popular before the deep learning era which are similar to low-level features. Though perception of expressions from low-level fea-



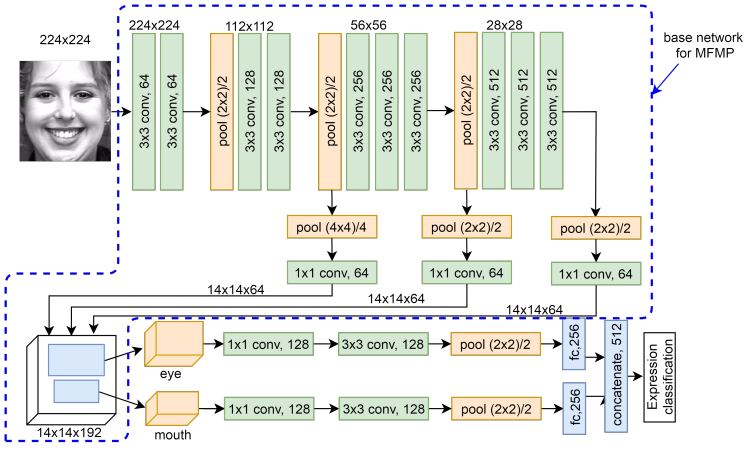
(a) M1



(b) M2



(c) M3



(d) M4

Figure 1: Different deep learning architectures considered in this work. (a) M1: The original VGG-Face model with suitable softmax layer for expression classes, (b) M2: Two facial patch (eye and mouth) based architecture with separate weights for each branch, (c) M3: Skipped connection based architecture with features pooled from both lower and higher order layers, (d) M4: Architecture with skipped connections followed by two parallel branches with separate weights to extract features from two facial patches.

tures such as local edges is quite inefficient, its contribution towards learning overall expression representation cannot be ignored. We used skip connections to make the model learn both high-level and low-level features simultaneously. Here the skip connections allow the concatenation of output of different layers to bring both high (early layers) and low (later layers) level features to the same vector space. The architecture of M3 is shown in Fig. 1c.

Since the dimension of feature maps for different convolution blocks are different, it is difficult to directly fuse these features. As can be seen from Fig. 1c, the skip connections from pooling layer of different blocks are further processed through pooling and 1×1 convolution to reduce the feature maps to a particular size (7×7). Note that this pooling operation is different for different blocks based on their feature map size. Finally, the outputs are concatenated channel-wise forming a feature map of size 7×7 with 256 channels. A fully connected layer with 512 units is used before the final classification layer.

Model M4 In our experiments, we observe that the skip connections in M3 are more effective than the part-based model of M2. To further validate the effect of skip connections along with part-based feature processing, we propose another model which takes the advantage of both the models. The fourth model (M4) uses the the skip connections of M3 and part-based architecture of M2 as shown in Fig. 1d. Different to M3, the skipped connections are processed with pooling and 1×1 convolutional layers to reduce the feature map dimension to 14×14, which corresponds to the dimension of feature maps in M2 for part based branching. The parallel branches for feature extraction of facial parts follow the same architecture as in M2.

Network Training of M1-M4 Given the input image x, the classification network learns to accurately predict the relevance expression scores $p(k|x, \theta)$, where θ are the network parameters and $k \in \{1, 2, ..., K\}$ represents K classes. For the soft-max layer, we have $p(k|x, \theta) = \frac{\exp(z_k)}{\sum_{i=1}^{K} \exp(z_i)}$, where z_i are the unnormalized log probabilities. The ground-truth distribution $q(k|x)$ is used to train the network parameters (θ) by minimizing the cross-entropy loss function

$$\mathcal{L} = -\sum_{k=1}^{K} \log(p(k|x, \theta))q(k|x). \tag{1}$$

Usually one-hot encoding is used in classification tasks, which takes the form $q(y|x) = 1$ and $q(k|x) = 0, \forall k \neq y$ for a sample x having class label y. However, the expression classes are highly correlated and two or more expressions can occur simultaneously as compound expressions. One-hot vectors impose for the CNN to predict one of the class labels with high confidence, i.e., with probability 1. We observe that learning the expression classifier with one-hot encoding prohibits the model to learn low intensity and mixed expressions. We use label smoothing [56] to allow the CNN model to adapt to the low intensity expressions to some extent. In label smoothing, the smoothed label information is used instead of 0-1 targets which imposes the model to be less confident about its predictions and the model is learned without pursuing of hard probabilities while encouraging the correct classification. We implement the label smoothing as,

$$q(k|x) = \begin{cases} 1 - \epsilon, & k = y \\ \frac{\epsilon}{K-1}, & k \neq y, \end{cases} \tag{2}$$

where $\epsilon \in [0, 1]$ is the label smoothing hyperparameter. While setting $\epsilon = 0$ refers to one-hot encoding, setting ϵ a large value might result in learning a poor performing model.

4. Multi-Face Multi-Part (MFMP) Framework

Data augmentation improves the performance in problems with low sample size. In our expression recognition framework, the pre-processing step involves face alignment. Once the faces are aligned, image augmentation with large translation or rotation is not effective. We used a simple strategy to augment data for part-based models. Instead of using periocular region and mouth region of the same face sample during network training, we augment the data by using these two regions from different face samples of the same expression class. We here exploit the improvement due to the augmentation with multiple regions sampled from multiple instances.

The multi-face multi-part (MFMP) architecture is shown in Fig. 2. It accepts two inputs, one for processing the peri-ocular region and another for processing the mouth region. As discussed earlier (in the description of M2), the feature maps of the facial regions in higher order layers are used instead of learning the weights for these regions from scratch. Here the network model M4 up to the concatenation layer is considered as the *base network*. The response of this base network to a whole face is used for further processing of each facial part. Specifically, both the inputs are first passed through the base network of M4 which produces a feature map of size $14 \times 14 \times 192$. As can be seen in Fig. 2, the base networks for both input stream share the same weights. The peri-ocular region is extracted from the upper branch, whereas the mouth region is extracted from the feature map of lower branch. Further, the extracted features are passed through the network similar to M4 resulting a 512 dimensional face embedding which is used for expression classification.

This architecture benefits from huge data augmentation. In a K class classification scenario with n_k samples in kth class and p number of patches, MFMP scheme increases the sample size from $\sum_k n_k$ to $\sum_k \binom{n_k}{p}$ by part-based augmentation. The improvement in classification accuracy with MFMP on different datasets are discussed in the result section.

During test time, we need to predict the emotion probabilities for each sample. Since the proposed MFMP accepts two inputs, the testing scheme needs to be defined thoroughly. Two methods can be followed to obtain the emotion probabilities of a sample: 1) The same image can be fed to both of the inputs in order to process the periocular region and mouth region in parallel followed by the feature concatenation and further recognition; 2) Since the weights are shared between both base networks, we can only use one of the parallel branches with a single input image. The latter method is equivalent to training the MFMP model with two inputs followed by transferring the

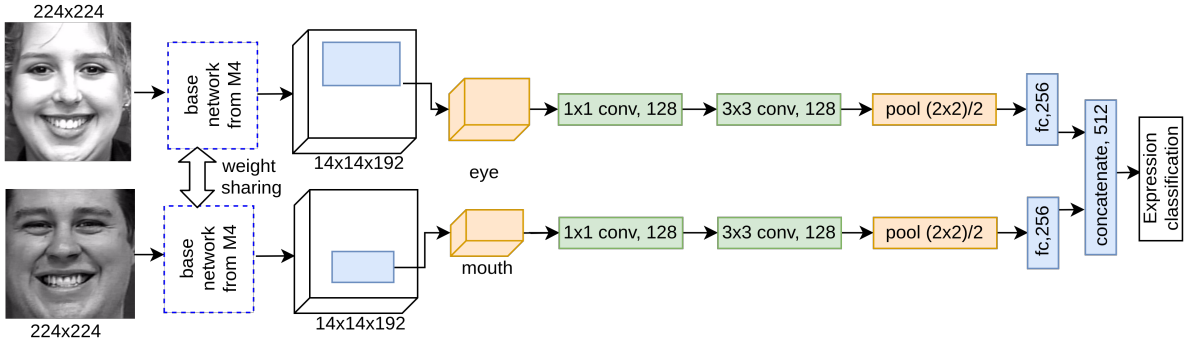


Figure 2: Proposed MFMP model for FER. The weight-sharing indicates that the weight parameters of all the convolution layers in the convolution blocks in for both branches are shared with each other.

weights of base network and the following layers to M4 architecture, thereby using a single input during testing phase.

Though the base networks share the same weight, the branch networks for each part learn separate weights. Thus, MFMP can be considered as a late fusion method with separate weight scheme for different facial parts [57]. With separate-weights, each branch treats input feature maps of facial parts differently and learns the high-level patterns associated to a specific part of face. Thereby, the network leverages the huge data augmentation possibility and learns the weights for a specific part in order to represent that region in the best possible way, which helps in improving the model performance.

MFMP+ : On the top of training with MFMP, an additional strategy is used in our experiments. We fine-tune the model trained with MFMP using different parts of the same face sample, which we refer as MFMP+. In other words, MFMP+ is firstly trained with facial parts from multiple faces (like MFMP) followed by fine-tuning the weights by training the model with parts of the same face sample. This process is similar to learning the model weights with MFMP first followed by using M4 for further fine-tuning the network weights. This process learns with a few more samples $(\sum_k \binom{n_k}{p} + \sum_k n_k)$ apart from the possible set from MFMP. Usually, the high-level feature map reveals the relation between different areas of the input. Since the face parts used in MFMP come from dissimilar faces, the weights learned by the parallel branches are less related. We hypothesize that once the weights of part-based parallel branches are learned with MFMP, they benefited from the positional relation of facial parts which is learned by the high-level feature map.

5. Experiments

Experiments are carried out on a number of data sets to validate the effectiveness of the part-based models. In this section, we first describe the data sets and the evaluation metrics we have used in this paper. Then, the implementation details of the CNN models are discussed followed by the experimental results from individual and cross-database evaluation. Finally, the results obtained in our method are compared to the state-of-the-art methods.

5.1. Data Set Description

Both in-the-lab (CK+[34], RaFD[35], FACES[37], lifespan[36]) and in-the-wild (RAF[58], AffectNet[59]) datasets are used in our experiments. The in-the-wild expression datasets contain data collected from uncontrolled environments and thus cover real-world expressions with various facial poses, illuminations, emotion intensity, occlusion, and other factors. Whereas the in-the-lab datasets are mostly posed in a controlled environment and contain exaggerated expressions of frontal faces.

Most of these datasets contain annotation for six basic expressions, namely: anger, disgust, fear, happy, sad, and surprise. Datasets like AffectNet and RAF are collected from internet by using certain emotional terms in various search engines. RAF [58] contains manually annotated images, out of which 12271 are listed in training set and 3068 samples in testing set. AffectNet [59] contains around 400000 labelled data for training from 10 categories and 5000 samples for validation. The agreement between two annotators in AffectNet is found to be 60%, which explains the complexity and subtlety of the expressions in this dataset. Following the experimental settings of [14] and [60], we only used the seven classes (six basic expressions and neutral). For both these datasets, we used the aligned images provided by the respective authors.

Among the in-the-lab datasets, CK+[34] is the smallest one with 327 sequences annotated with 7 expression labels (anger, disgust, fear, happiness, sad, surprise, and contempt). The contempt class was excluded in our experiments as we aim for cross-database evaluation and comparison purposes [31]. The image sequences in CK+ start from a neutral face and end with a peak of the respective expression. Following the literature [31][11][33], we used the 10-fold cross validation while using the first frame as neutral and the last three frames of each sequence as the corresponding expression label. Thus, our seven class framework on CK+ uses the classes: anger, disgust, fear, happiness, neutral, sad, and surprise; and in 6 class classification, it excludes the 'neutral' class. The frontal faces (1407 samples) were used from RaFD dataset for our experiments, which includes neutral and six basic expressions of 67 models. Lifespan database is a challenging dataset with subjects of various age groups. We followed the experimental settings of [61],

[62] and [63], and conducted experiments on four class (happy, neutral, sad, surprise: 1137 images) and two class (neutral and happy: 995 images) settings.

5.2. Implementation Details

The faces from the dataset images were detected using MTCNN [64] followed by face alignment by positioning both eyes at a fixed distance parallel to the horizontal axis. The aligned faces were re-sized to 224×224 resolution and feed to the CNN models. The initial model weights were obtained from the pretrained VGG-Face model for the layers that were borrowed from VGG-Face. The rest of the weights were initialized with Xavier uniform initialization. The network was trained with mini-batch gradient descent method with batch size of 32. We normalized the training data to zero mean and unit variance in all experiments. Data augmentation was carried out with random horizontal flip, small rotation, and scaling. We use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of 1e–4 for all the models. The label smoothing parameter ϵ was empirically set to 0.1 in all our experiments. The experiments were carried out using NVIDIA 1080 GPU with CUDA to improve the speed.

We perform a series of experiments on M1 by freezing the lower-order layer weights while fine-tuning the higher-order layer weights to learn the suitable model for FER purpose. From our experiments, we conclude that fine-tuning the last fully connected layers achieves the best performance for expression classification in low sample size datasets, whereas fine-tuning two or three more layers achieves the best performance for large datasets. The best results for M1 in our experiments are reported in this paper.

5.3. Evaluation Strategy

The in-the-wild datasets (AffectNet and RAF) have train and validation sets defined by the database developers (for RAF, the test set is used for validation). While we train our models on a train set, we validate it on a validation set. For other datasets (CK+, RaFD, FACES, lifespan), we employ 5-fold cross-validation, as often reported in literature [11], [33], [62], [65]. To compare our model performance with the state-of-the-art methods, we use macro average accuracy, which is computed by averaging over the accuracy scores of each expression category.

Cross-dataset Evaluation The reliability of data annotation is a major concern in facial expression datasets. We adopted rigorous cross-dataset evaluation protocol to validate the effectiveness of the proposed method. In cross-dataset evaluation, the models are trained on samples from one dataset while tested on other, thereby demonstrating the generalization ability of the model. For RAF and AffectNet, the models were trained or tested on train and validation split respectively. For the rest datasets, the models were trained or tested on whole data for cross-dataset validation.

Table 1: Comparison of average classification accuracy on different databases for VGG variants. The best accuracy is reported in **bold** text.

Models	CK+	RaFD	lifespan	AffectNet
M1	97.34	97.23	92.98	52.83
M2	95.8	89.82	78.42	46.23
M3	**98.54**	**98.93**	91.67	**53.68**
M4	98.18	98.29	**93.86**	52.75

5.4. Performance Comparison of M1, M2, M3, and M4

The results obtained with models M1, M2, M3, and M4 are presented in this section. The performance of different models on different datasets are provided in Table 1. Though it is a popular belief that the use of patch-based features always enhances the expression recognition performance, our experimental results differ from it for deep learning models. As can be seen from Table 1, the performance of M2 is considerably lower than the other models. While the other models achieve close performances on different datasets, M2 lags by a huge margin to each of them. In other words, features from regions beyond eyes and mouth are also beneficial in expression recognition.

Table 1 indicates that the model with skip connections (M3) performs well across the datasets. Comparing the accuracy M1 and M3, we observe the performance improvement by fusing both low and high-order features. Since VGG-Face was initially trained for person identification, the receptive fields of higher order layers tend to learn the identity instead of the expressions. However, by pooling information from lower order layers enhances the expression classification rate.

The performance of M3 and M4 are very close. Though M4 achieves the best recognition accuracy in lifespan dataset, M3 outperforms M4 in the rest. One of the likely reasons for reduced performance of M4 is the loss of holistic information from the face during part-based analysis. Since the performance gap between M3 and M4 is very small, we can infer that the performance is more robust after pooling the features from both low and high-order deep layers. Moreover, the huge performance gap between M2 and M4 implicates that the use of skip connections improved the accuracy, even though both models use similar part-based feature maps.

The following conclusions can be drawn from the observations: (1) the use of mouth and periocular regions for feature extraction reduces the performance in CNN models, i.e. some important information is lost that resides on other facial parts; (2) the skip connections prove to be effective with or without facial part based models. We will discuss in the following Section that the performance of MFMP is slightly better than both M3 and M4. Therefore, following the above observations, the rest of the paper reports the performance comparison of M4 and MFMP along with the other methods in the literature, while neglecting the performance of other models.

5.5. Performance of MFMP/MFMP+ and Comparisons

The MFMP architecture has the advantage of data augmentation as discussed in Section 4. The performance of MFMP and MFMP+ is compared with M4 and other state-of-the-art

Table 2: Comparison of average classification accuracy on CK+ database. The best two accuracies are shown in **bold** and underlined text respectively.

Methods	Validation settings	Accuracy
STM-ExpLet [66]	7 class	94.19
LOMo [67]	7 class	95.1
IACNN [33]	7 class	95.37
DTAGN [26]	7 class	97.25
Lopes *et al.* [29]	7 class	**98.8**
Proposed M4	7 class	97.17
Proposed MFMP	7 class	97.09
Proposed MFMP+	7 class	97.5
BDBN[31]	6 class[a]	96.7
PPDN [32]	6 class	97.3
facenet2expnet [11]	6 class	98.6
Lopes *et al.* [29]	6 class	**98.92**
Proposed M4	6 class	98.18
Proposed MFMP	6 class	98.65
Proposed MFMP+	6 class	98.85

[a] Eight-fold cross-validation is performed.

Table 3: Comparison of seven class classification accuracy on RaFD database.

Methods	Validation settings	Accuracy
Metric learning[68]	10 fold	95.95
W-CR-AFM [69]	train-test split	96.27
BAE-BNN-3[65]	5 fold	96.93
TLCNN+FOS[70][b]	4 fold	97.75
Carcagni *et al.*[71]	10 fold	98.5
Proposed M4	5 fold	98.29
Proposed MFMP	5 fold	98.6
Proposed MFMP	10 fold	98.64
Proposed MFMP+	5 fold	99.07
Proposed MFMP+	10 fold	**99.1**

[b] Six classes considered. Neutral class was not included.

Table 4: Comparison of expression recognition accuracy on lifespan database.

Methods	Validation settings	Accuracy
Guo *et al.*[61]	2 class	91.05
Joint-Learn [62]	2 class	93.91
Proposed M4	2 class	93.86
Proposed MFMP	2 class	94.29
Proposed MFMP+	2 class	**95.64**
Wu *et al.* [63]	4 class	82.55
Proposed M4	4 class	81.66
Proposed MFMP	4 class	**89.3**
Proposed MFMP+	4 class	88.45

Table 5: Comparison of expression recognition accuracy on FACES database.

Methods	Accuracy
Guo *et al.* [61]	84.68
Joint-Learn [62]	92.19
Wu *et al.* [63]	94.12
Proposed M4	96.1
Proposed MFMP	96.49
Proposed MFMP+	**96.78**

methods in Table 2–6. The best two accuracies in each table are shown in bold and underlined text respectively. For CK+, some literature reports the accuracy for 6 class classification, whereas some other report the 7 class accuracy. Therefore, we report both the cases. Similarly, we report the accuracy in two and four class settings for lifespan dataset.

From Table 2, we observe that the method proposed by Lopes *et al.* [29] achieves the best accuracy of 98.8% for seven class classification. However, the accuracy of the proposed MFMP+ is the second best and for six class classification, MFMP+ performs close to the state-of-the-art accuracy. MFMP+ and MFMP achieved the top accuracies (99.07% and 98.6% respectively) in RaFD (see Table 3) outperforming the other state-of-the-art methods. In both lifespan and FACES, the proposed methods outperform the methods in literature with a margin of 2% approximately. As can be seen from Table 4, the best accuracy is obtained with MFMP for 4 class classification, whereas MFMP+ achieves the best accuracy for 2 class classification. Similarly, in FACES dataset, MFMP+ achieves an accuracy of 96.78% and outperforms the state-of-the-art results by a margin of 2.5%. For three in-the-lab datasets, MFMP and MFMP+ achieved the top two performances. Further, MFMP+ achieves better accuracy than MFMP in all in-the-lab datasets. The lack of sufficient samples might be the reason that prevents MFMP to learn a suitable generalization, whereas MFMP+ uses the same faces at both inputs for training once the model learns the suitable weights with the proposed augmentation method, thus learning a better expression generalization.

On the other hand, MFMP outperforms MFMP+ in case of in-the-wild datasets, where the number of training samples is large. As can be seen in Table 6, the performance of MFMP is 3% higher than that of MFMP+. However, the proposed methods could not achieve state-of-the-art performances in RAF. IPA2LT [72] achieves the best on RAF as it is trained with both AffectNet and RAF train data. We believe the presence of occlusion, pose variations, lower resolution, and facial painting are the major reasons for the failure of the model. However, these limitations can be overcome with a larger dataset like AffectNet. As can be seen in Table 7, the highest accuracy of 58.93% was obtained with MFMP for AffectNet.

The following observations can be outlined from the experimental results: (1) Both MFMP and MFMP+ perform consistently better than M4 model; (2) Different from other works in literature we use only two facial regions to obtain good performance; (3) The consistent trend of MFMP and MFMP+ on all posed expression datasets confirms the advantage of facial

Table 6: Comparison of expression recognition accuracy on RAF database.

Methods	Accuracy
CAKE [60]	68.9
DLP-CNN [58]	74.2
Vielzeuf *et al.* [73]	80
PG-CNN [14]	83.27
IPA2LT [72][c]	**86.77**
Proposed M4	72.54
Proposed MFMP	83.15
Proposed MFMP+	80.26

[c] Trained with both AffectNet and RAF train set.

Table 7: Comparison of expression recognition accuracy on AffectNet database.

Methods	Accuracy
PG-CNN [14]	55.33
IPA2LT [72][c]	57.31
CAKE [60]	58.1
AlexNet [59]	58
Proposed M4	52.75
Proposed MFMP	**58.93**
Proposed MFMP+	58.86

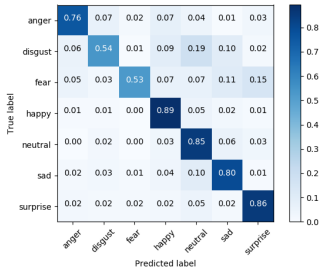
[c] Trained with both AffectNet and RAF train set.

patch augmentation. The performance of proposed methods achieve higher accuracy or close to the state-of-the-art methods; (4) A benefit of MFMP+ is observed in most in-the-lab expression datasets as they contain a small number of samples. However, MFMP method seems to perform better in large datasets. This indicates that the generalization ability of our architecture is improved with large dataset. Notice that the use of more advanced architecture (such as ResNet) could further improve the model performance.

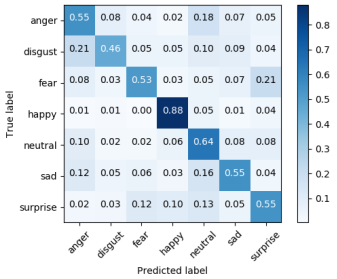
5.6. Results and Analysis

Fig. 3a and 3b respectively show the confusion matrices of RAF and AffectNet. *Happiness* is detected with highest accuracy among all the expression classes in both the datasets. In the RAF dataset, the accuracy of *neutral* and *surprise* is more than 85%. However, *disgust* and *fear* are the lowest performing classes in both the datasets. And it is observed in both datasets that *fear* is classified as *surprise* among most of the false classifications. Similarly, *sad* is mainly misclassified as *neutral*. The misclassified *disgust* samples are mostly assigned to *neutral* and *anger* classes in RAF and AffectNet respectively. The subtle variation among classes such as anger, sad, and neutral is probably the primary reason behind misclassification among these classes.

To further demonstrate the effectiveness of our method, we report the class-wise ROC and the macro-ROC for RAF and AffectNet datasets in Fig. 4a and 4b respectively. The ROC curve is obtained by plotting the true-positive rates against the false-positive rates while varying the decision threshold of the prediction scores of the softmax layer. Similar to observations in Fig. 3b, we can see that the ROC of happiness class is better than the rest in AffectNet. The AUC for each class is also shown for better comprehension. The proposed model achieves an AUC macro average of 0.96 and 0.90 in RAF and AffectNet respectively.



(a) RAF



(b) AffectNet

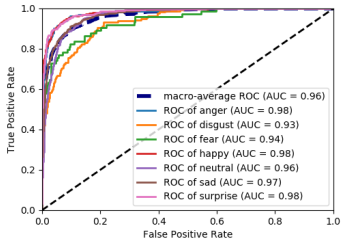
Figure 3: Confusion matrices for RAF and AffectNet with the proposed MFMP+ model.

Fig. 5 shows the validation accuracies with respect to the number of epochs during training. We can observe that the curve of MFMP+ is above the other models with a significant margin. Since the pretrained weights of VGG-Face are used for the low order layers, each model learns the expression patterns very quickly in less than 50 epochs. However, the MFMP+ model learns faster than other models and achieves the best accuracy.

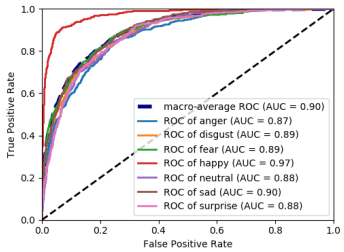
The T-SNE representation of the feature embedding of M3 and MFMP is shown in Fig. 6. Since both the models achieve close performances, the projection patterns on two-dimensional plane for these two models are almost similar. However, one can see that the samples of surprise are projected to two clusters and much more separated from other classes in case of MFMP. Similarly, the samples from sad are scattered in M3 projection, whereas these samples are more distinguished in MFMP projection.

5.7. Cross-database Evaluation

Table 8 reports the cross-dataset performance for classification of seven expression classes. The other datasets are not included in the evaluation due to the unavailability of all seven class samples in those datasets. The best performing model weights were used in these experiments. The datasets presented in the table are in the increasing order of number of samples in



(a) RAF



(b) AffectNet

Figure 4: Class-wise ROC and the macro-ROC for RAF and AffectNet with the proposed MFMP+ model. Comparing the class-wise ROCs, *happy* class performs the best in both the datasets.

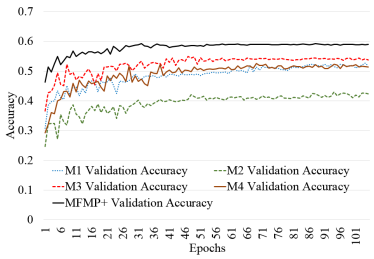


Figure 5: The accuracy curves of different models on AffectNet validation set. MFMP+ model learns faster than other models and achieves the best accuracy.

Table 8: Cross-dataset accuracy results for seven expression classes.

TrainTest	CK+	RaFD	RAF	AffectNet
CK+	-	52.48	24.64	21.57
RaFD	63.41	-	27.65	22.64
RAF	70.91	68.45	-	42.17
AffectNet	81.96	86.92	69.89	-

the dataset (CK+ <RaFD <RAF <AffectNet). One interesting trend that can be observed from Table 8 is that, when a dataset with less samples is used for training, the model performance is low on datasets with larger samples, and vice versa. This has also been observed in [31] and [29], where this trend is considered as the effect of less variations in training data due to low sample size. Furthermore, the occlusion, pose and illumination variations are absent in-the-lab datasets; thus, it is difficult to train the complex model architecture without causing significant overfitting. As can be seen in Table 8, the best accuracies are obtained when the models are trained on AffectNet. The cross-dataset performance of RaFD is better than CK+ when the model is trained on AffectNet; this probably happens as the expressions in RaFD are more exaggerated than the samples of CK+ and the model trained with AffectNet learns the optimal weight for high intensity expression recognition.

6. Conclusions

In this paper, (a) we compared the performance of holistic and part-based deep models for expression recognition, (b) we explored skip connections in the context of studied models and (c) we proposed a data augmentation scheme to improve the performance of part-based models.

With respect to the experiments, the results suggest that (a) holistic models outperform part-based models and hence pertinent information is lost when only facial parts are being analyzed. Further, we showed that (b) skip connections can improve the accuracy of part-based models substantially, which attests the importance of low-level feature maps in expression recognition. The proposed data augmentation scheme uses multiple facial parts, each from different samples of the same expression category, thereby increasing the number of samples to a great extent. MFMP model that uses the data augmentation scheme (c) improved additionally the performance of part-based models. Since the spatial relationship is lost during MFMP training, the model is able to learn to infer the best out of a specific face region w.r.t. expression recognition. Further, with separate weights for each facial part, the network learned patterns associated to specific parts of the face and represented that region efficiently, which benefited the model performance. The proposed data augmentation framework can be instrumental in other part-based deep models, where data alignment plays a critical role.

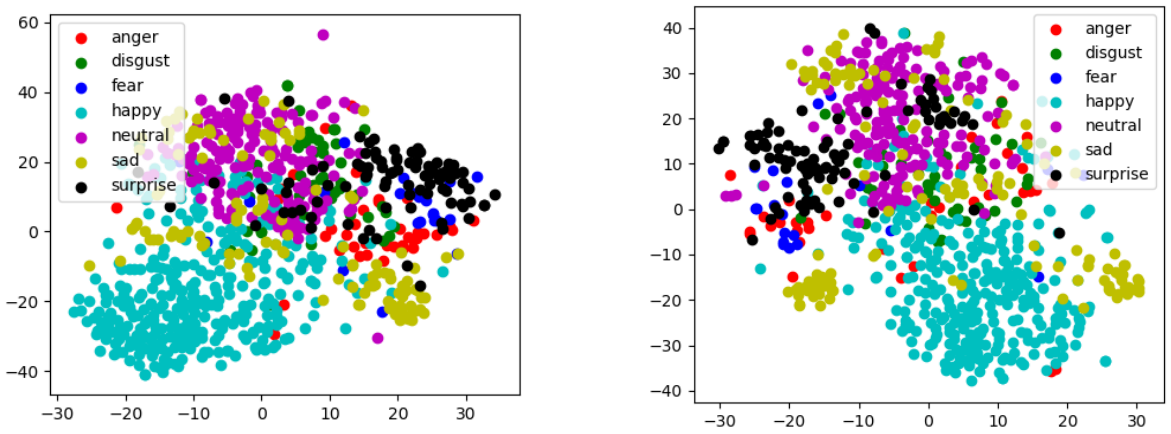
7. References

[1] J. F. Cohn, P. Ekman, Measuring facial action, The new handbook of methods in nonverbal behavior research (2005) 9–64.

[2] P. Ekman, W. V. Friesen, Facial action coding system, Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[3] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: A survey of registration, representation, and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015) 1113–1133.

[4] L. Zhong, Q. Liu, P. Yang, J. Huang, D. N. Metaxas, Learning multiscale active facial patches for expression analysis, IEEE Transactions on Cybernetics 45 (2015) 1499–1510.



(a) M3

(b) MFMP

Figure 6: The two dimensional T-SNE representation of the feature embedding of M3 and MFMP on a few samples of RAF test set. The projections of *sad* and *surprise* samples are less scattered in MFMP framework compared to the projections in M3.

[5] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Computer Vision and Image Understanding (CVIU) 115 (2011) 541–558.

[6] Q. Mao, Q. Rao, Y. Yu, M. Dong, Hierarchical bayesian theme models for multipose facial expression recognition, IEEE Transactions on Multimedia 19 (2017) 861–873.

[7] S. Azzakhnini, L. Ballihi, D. Aboutajdine, Combining facial parts for learning gender, ethnicity, and emotional state based on rgb-d information, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (2018) 19.

[8] S. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE transactions on Affective Computing 6 (2015) 1–12.

[9] P. Ekman, E. L. Rosenberg, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Oxford University Press, USA, 1997.

[10] S. Jaiswal, M. Valstar, Deep learning the dynamic appearance and shape of facial action units, IEEE Winter Conference on Applications of Computer Vision (WACV) (2016) 1–8.

[11] H. Ding, S. K. Zhou, R. Chellappa, Facenet2expnet: Regularizing a deep face recognition net for expression recognition, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (2017) 118–126.

[12] D. Acharya, Z. Huang, D. Pani Paudel, L. Van Gool, Covariance pooling for facial expression recognition, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018) 367–374.

[13] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recognition (PR) 84 (2018) 251–261.

[14] Y. Li, J. Zeng, S. Shan, X. Chen, Patch-gated cnn for occlusion-aware facial expression recognition, International Conference on Pattern Recognition (ICPR) (2018).

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition., BMVC 1 (2015) 6.

[16] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, ACM International Conference on Multimodal Interaction (ICMI) (2015) 435–442.

[17] H.-W. Kung, Y.-H. Tu, C.-T. Hsu, Dual subspace nonnegative graph embedding for identity-independent expression recognition, IEEE Transactions on Information Forensics and Security (TIFS) 10 (2015) 626–639.

[18] Y. Guo, G. Zhao, M. Pietikäinen, Dynamic facial expression recognition with atlas construction and sparse representation, IEEE Transactions on Image Processing (TIP) 25 (2016) 1977–1992.

[19] S. H. Lee, K. N. K. Plataniotis, Y. M. Ro, Intra-class variation reduction using training expression images for sparse representation based facial expression recognition, IEEE Transactions on Affective Computing 5 (2014) 340–351.

[20] X. Huang, G. Zhao, W. Zheng, M. Pietikainen, Spatiotemporal local monogenic binary patterns for facial expression recognition, IEEE Signal Processing Letters 19 (2012) 243–246.

[21] B. Ryu, A. R. Rivera, J. Kim, O. Chae, Local directional ternary pattern for facial expression recognition, IEEE Transactions on Image Processing (TIP) 26 (2017) 6006–6018.

[22] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Transactions on Affective Computing 9 (2018) 38–50.

[23] A. Dhall, A. Asthana, R. Goecke, T. Gedeon, Emotion recognition using phog and lpq features, IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011) (2011) 878–883.

[24] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, M. Pantic, Deep structured learning for facial action unit intensity estimation, Computer Vision and Pattern Recognition (CVPR) (2017) 5709–5718.

[25] A. Mollahosseini, D. Chan, M. H. Mahoor, Going deeper in facial expression recognition using deep neural networks, IEEE Winter Conference on Applications of Computer Vision (WACV) (2016) 1–10.

[26] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, International Conference on Computer Vision (ICCV) (2015) 2983–2991.

[27] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni, Expnet: Landmark-free, deep, 3d facial expressions, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018) 122–129.

[28] P. Rodriguez, G. Cucurull, J. Gonzalez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, F. X. Roca, Deep pain: Exploiting long short-term memory networks for facial expression classification, IEEE Transactions on Cybernetics (2017) 1–11.

[29] A. T. Lopes, E. de Aguiar, A. F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognition (PR) 61 (2017) 610–628.

[30] G. Pons, D. Masip, Supervised committee of convolutional neural networks in automated facial expression analysis, IEEE Transactions on Affective Computing 9 (2018) 343–350.

[31] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1805–1812.

[32] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, European confer-

ence on computer vision (ECCV) (2016) 425–442.
[33] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (2017) 558–565.
[34] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, IEEE Computer Vision and Pattern Recognition - Workshops on human communicative behavior analysis (2010) 94–101.
[35] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, A. van Knippenberg, Presentation and validation of the radboud faces database, Cognition and Emotion 24 (2010) 1377–1388.
[36] M. Minear, D. C. Park, A lifespan database of adult facial stimuli, Behavior Research Methods, Instruments, & Computers 36 (2004) 630–633.
[37] N. C. Ebner, M. Riediger, U. Lindenberger, Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation, Behavior research methods 42 (2010) 351–362.
[38] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1717–1724.
[39] H.-W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, ACM international conference on multimodal interaction (2015) 443–449.
[40] S. Albanie, A. Vedaldi, Learning grimaces by watching TV, British Machine Vision Conference (BMVC) (2016).
[41] D. Lin, X. Shen, C. Lu, J. Jia, Deep lac: Deep localization, alignment and classification for fine-grained recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 1666–1674.
[42] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, D. Metaxas, Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1143–1152.
[43] G.-S. Xie, X.-Y. Zhang, W. Yang, M. Xu, S. Yan, C.-L. Liu, Lg-cnn: From local parts to global discrimination for fine-grained recognition, Pattern Recognition (PR) 71 (2017) 118–131.
[44] S. Huang, Z. Xu, D. Tao, Y. Zhang, Part-stacked cnn for fine-grained visual categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1173–1182.
[45] A. Gonzalez-Garcia, D. Modolo, V. Ferrari, Do semantic parts emerge in convolutional neural networks?, International Journal of Computer Vision (IJCV) 126 (2018) 476–494.
[46] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2 (2017) 3.
[47] Y. Peng, X. He, J. Zhao, Object-part attention model for fine-grained image classification, IEEE Transactions on Image Processing (TIP) 27 (2018) 1487–1500.
[48] T. Sun, L. Sun, D.-Y. Yeung, Fine-grained categorization via cnn-based automatic extraction and integration of object-level and part-level features, Image and Vision Computing (IVC) 64 (2017) 47–66.
[49] H. Li, G. Hua, Hierarchical-pep model for real-world face recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 4055–4064.
[50] S. Gao, K. Jia, L. Zhuang, Y. Ma, Neither global nor local: Regularized patch-based representation for single sample per person face recognition, International Journal of Computer Vision (IJCV) 111 (2015) 365–383.
[51] F. H. d. B. Zavan, O. R. Bellon, L. Silva, G. G. Medioni, Benchmarking parts based face processing in-the-wild for gender recognition and head pose estimation, Pattern Recognition Letters (PRL) (2018).
[52] S. Happy, A. George, A. Routray, A real time facial expression classification system using local binary patterns, International Conference on Intelligent Human Computer Interaction (2012) 1–5.
[53] X. Fan, T. Tjahjadi, A dynamic framework based on local zernike moment and motion history image for facial expression recognition, Pattern Recognition (PR) 64 (2017) 399–406.
[54] W. Li, F. Abtahi, Z. Zhu, L. Yin, Eac-net: Deep nets with enhancing and cropping for facial action unit detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2018).
[55] S. Happy, A. Routray, Robust facial expression classification using shape and appearance features, International Conference on Advances in Pattern Recognition (2015) 1–5.
[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2818–2826.
[57] J. Shao, C. C. Loy, K. Kang, X. Wang, Crowded scene understanding by deeply learned volumetric slices, IEEE Transactions on Circuits and Systems for Video Technology 27 (2017) 613–623.
[58] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, pp. 2584–2593.
[59] A. Mollahosseini, B. Hasani, M. H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, IEEE Transactions on Affective Computing (2017).
[60] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, F. Jurie, Cake: Compact and accurate k-dimensional representation of emotion, Image Analysis for Human Facial and Activity Recognition (BMVC Workshop) (2018).
[61] G. Guo, R. Guo, X. Li, Facial expression recognition influenced by human aging, IEEE Transactions on Affective Computing 4 (2013) 291–298.
[62] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, T. Gevers, Expression-invariant age estimation using structured learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 40 (2018) 365–375.
[63] S. Wu, S. Wang, J. Wang, Enhanced facial expression recognition by age, IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 1 (2015) 1–6.
[64] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503.
[65] W. Sun, H. Zhao, Z. Jin, An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks, Neurocomputing 267 (2017) 385–395.
[66] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1749–1756.
[67] K. Sikka, G. Sharma, M. Bartlett, Lomo: Latent ordinal model for facial analysis in videos, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 5580–5589.
[68] B. Jiang, K. Jia, Robust facial expression recognition algorithm based on local metric learning, Journal of Electronic Imaging 25 (2016) 013022.
[69] B.-F. Wu, C.-H. Lin, Adaptive feature mapping for customizing deep learning based facial expression recognition model, IEEE Access 6 (2018) 12451–12461.
[70] Y. Zhou, B. E. Shi, Action unit selective feature maps in deep networks for facial expression recognition, International Joint Conference on Neural Networks (IJCNN) (2017) 2031–2038.
[71] P. Carcagnì, M. Coco, M. Leo, C. Distante, Facial expression recognition and histograms of oriented gradients: a comprehensive study, SpringerPlus 4 (2015) 645.
[72] J. Zeng, S. Shan, X. Chen, Facial expression recognition with inconsistently annotated datasets, European conference on computer vision (ECCV) (2018) 222–37.
[73] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, F. Jurie, An occam's razor view on learning audiovisual emotion recognition with small training sets, International Conference on Multimodal Interaction (ICMI) (2018) 589–593.