Fusing Aligned and Non-Aligned Face Information
for Automatic Affect Recognition in the Wild: A Deep Learning Approach

Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, Soo-Young Lee

Computational NeuroSystems Laboratory (CNSL)
Korea Advanced Institute of Science and Technology (KAIST)
{bokyeong1015, suhyeon.dong}@gmail.com, {rohleejh, gmkim90, sylee}@kaist.ac.kr

Abstract

Face alignment can fail in real-world conditions, negatively impacting the performance of automatic facial expression recognition (FER) systems. In this study, we assume a realistic situation including non-alignable faces due to failures in facial landmark detection. Our proposed approach fuses information about non-aligned and aligned facial states, in order to boost FER accuracy and efficiency. Six experimental scenarios using discriminative deep convolutional neural networks (DCNs) are compared, and causes for performance differences are identified. To handle non-alignable faces better, we further introduce DCNs that learn a mapping from non-aligned facial states to aligned ones, alignment-mapping networks (AMNs). We show that AMNs represent geometric transformations of face alignment, providing features beneficial for FER. Our automatic system based on ensembles of the discriminative DCNs and the AMNs achieves impressive results on a challenging database for FER in the wild.

1. Introduction

Facial expression is a primary means for understanding human emotions and has been actively studied over two decades [see 1-3 for survey]. Rapid progress has been made on technologies for automatic facial expression recognition (FER), particularly in controlled laboratory settings. Nevertheless, FER still remains a challenge in uncontrolled real-life situations in which a FER system must handle unpredictable variability in head poses, lighting conditions, occlusions, and subjects. To resolve this issue, researchers have collected large volumes of data "in the wild" [4-7], have held many grand challenges [8-11], and have presented excellent approaches notably with deep learning techniques [12-18].

Face alignment commonly performed in preprocessing modules is essential for achieving good FER performance
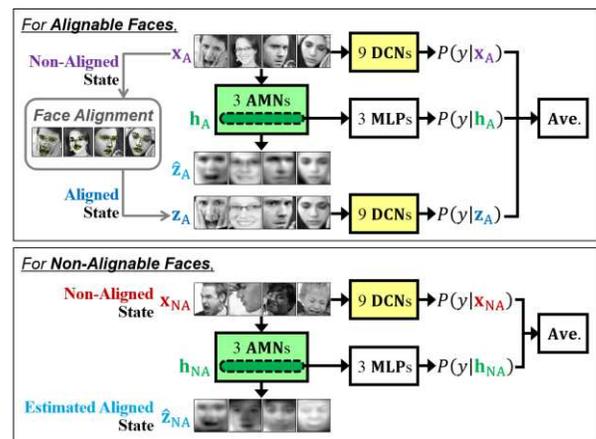
Figure 1: Overview of our approach. Our automatic FER system contains several DCNs to fuse information of non-aligned and aligned facial states.

[6, 16, 18]. However, many studies have applied some "manual" steps when this preprocessing fails. For real-life applications, automatic FER systems in which human intervention is not necessary are preferred, and this demands automatic face alignment. Recently, such alignment techniques have been extensively studied by using holistic deformable models (DMs) [19, 20] and parts-based DMs [21, 22], by taking advantage of both of the DMs [23, 24], and by applying regression-based methods [25, 26]. However, when alignment errors or failures occur "in the wild" and propagate to later stages of FER systems, final performance declines.

In this paper, we propose a framework based on an ensemble of deep convolutional neural networks (DCNs) toward automatic FER in the wild. We begin with the following realistic assumption (validated in Section 3):

■ Face alignment under real-world conditions is assumed to be not always successful.
■ Consequently, face images are either "alignable", *i.e.*, capable of being aligned or "non-alignable", *i.e.*, not capable of being aligned.

Our FER system is designed to separately deal with alignable faces and non-alignable faces, as shown in Figure 1. For alignable faces, we combine information from both of the "non-aligned state" and the "aligned state". For non-alignable faces, we initially obtain class predictions using the "non-aligned state". In order to determine the method to train discriminative DCNs, we assess several experimental scenarios for a better information fusion (Section 4). Then, in order to estimate aligned states of non-alignable faces, we introduce DCNs which learn a *mapping* from the non-aligned facial state to the aligned one (Section 5), called alignment-mapping networks (AMNs). Experimental results demonstrate that these AMNs represent similarity transformations performed in face alignment and yield hidden features beneficial for FER. To summarize, our empirical findings in developing the automatic FER system show that

■ In the learning phase of individual discriminative DCNs, it is beneficial to use the merged training dataset of alignable and non-alignable faces for efficiency in evaluation time as well as for effectiveness in FER performance.
■ In the testing phase for alignable faces, combining information of both non-aligned and aligned states from discriminative DCNs improves FER accuracy.
■ In the testing phase for non-alignable and alignable faces, it is better to add decision-level information from classifying the hidden features of AMNs.

We evaluate the proposed framework on the facial expression recognition 2013 (FER-2013) database [7], a challenging benchmark collected from the Web. Our final ensemble-based FER system achieves great performance on this in-the-wild database.

2. Related Work

Diversity and ensembles of neural networks. Combining decisions of multiple artificial neural networks is a method with a long history [27-29] in the field of classifier ensemble. Here, employing "diversity" of individual networks has been shown to improve ensemble performance, by providing uncorrelated, different, and thus informative decisions [30, 31]. For high diversity, individual networks are built using several diversification strategies [28, 29], *e.g.*, by altering network architectures and random weight initializations as well as by using variously normalized and differently divided datasets. Such strategies are now applied to ensembles of *deep* neural networks, resulting in remarkable successes in image classification [32-34]. In this work, our DCNs are also built towards high diversity, specifically by changing input normalization and weight

initialization.

Deep learning for facial expression recognition. In recent years, similar to other computer vision problems, applying deep learning techniques to FER has attracted considerable attention. For disentangling latent factors in face images, Rifai *et al.* [12] gradually separated discriminative expression information from non-discriminative pose and morphology factors based on convolutional networks and auto-encoders, while Reed *et al.* [13] modeled higher-order interactions of expression and pose manifolds based on a restricted Boltzmann machine. In addition, Reed *et al.* [14] used a bootstrapping-based approach regarding prediction consistency, resolving a problem of noisy- and subjective labels in FER. Liu *et al.* [15] used convolutional kernels for capturing facial action units relevant to facial expressions and extracted higher-level features based on deep belief networks.

To improve affect recognition in the wild, there are several ensemble-based deep learning approaches, which inspire our work. Kahou *et al.* [16] proposed a multimodal framework including a DCN and a deep belief network for extracting visual and audio features, respectively, and applied a random-search-based weighted decision fusion. Moreover, in order to combine multiple DCNs using late fusion schemes, Yu *et al.* [17] applied a hinge-loss-based weighted fusion in a single-level committee, while Kim *et al.* [18] used an exponentially weighted fusion in a hierarchical committee. These three studies used the FER-2013 database described in Sect. 3 to pre-train their DCNs for transfer learning or feature extraction. However, they did not cope with the failure in face alignment nor consider its influence on FER performance.

Converting pose states of face images. Many deep models for face recognition aim to learn a mapping from non-frontal faces to frontal ones, which share the similar goal with our AMNs. Kan *et al.* [35] used a stacked auto-encoder network to model gradual change in poses, while Zhu *et al.* [36] designed a deep convolutional network including locally-connected layers to extract pose- and illumination-robust features. In order to model the continuous pose space, Zhu *et al.* [37] used stochastic neurons in their multi-view perceptron which disentangled pose and identity factors. In addition, Yim *et al.* [38] employed a multi-task learning scheme to convert head poses while concurrently maintaining identity information.

One limitation of the aforementioned deep models is that they focused on changes only in yaw rotation. In addition, they dealt with data only captured under controlled lab settings, and 2-D aligned faces were used as inputs and outputs of their models. In contrast, our work attempts to handle all rotations of yaw, pitch, and roll as well as cropping and resizing operations. We deal with faces in the wild, and the faces before and after 2-D alignment are used
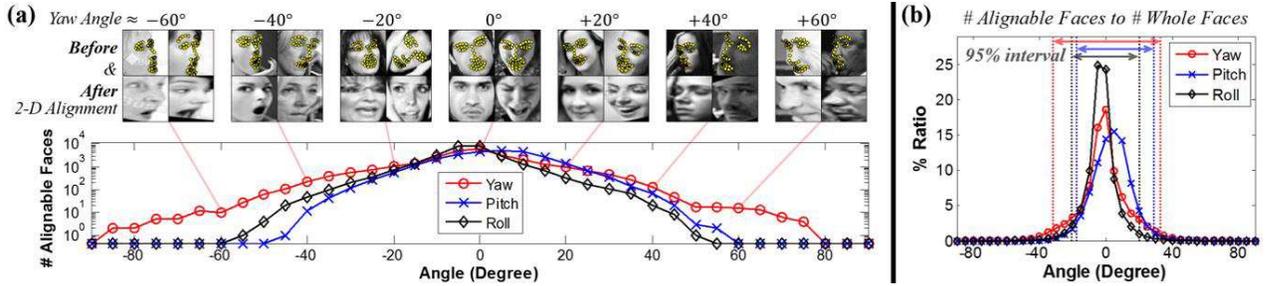
Figure 2: Statistics of head pose and examples of 2-D aligned result on the FER-2013 DB. The graph (a) shows the number of alignable faces as a function of the pose angle of yaw, pitch, and roll on a log scale, and the (b) shows the percent ratio of aliganble faces for each pose angle on a linear scale. The ranges of (-31.7, 32.7) degree for yaw rotation, (-17.5, 29.1) for pitch, and (-20.5, 20.3) for roll contain 95% of the alignable faces, indicating in-the-wild conditions of the FER-2013 DB.

as inputs and target outputs of our AMNs, respectively.

Very recently, interesting approaches have been proposed to yield frontal facial poses in unconstrained settings. After localizing facial landmarks, Hassner *et al.* [39] used a 3-D reference model common to all faces and efficiently estimated visible facial parts, improving performance. On the other hand, Sagonas *et al.* [40] employed an effective statistical model to localize landmarks and convert facial poses at the same time, notably using only frontal faces. Although these studies did not use deep learning techniques, excellent results on in-the-wild datasets were obtained particularly due to rigorous mathematical derivation.

3. Data and Face Alignment

This section describes the FER-2013 database used in our work. To investigate the validity of assumption in Sect. 1, we present face alignment results on this database. Finally, experimental datasets which we construct are introduced.

3.1. The FER-2013 database

The FER-2013 database was collected from the Web, and most images were captured under real-world imaging conditions. This database has been reported to include some noisy or confusing annotation and show a low human FER accuracy of approximately 65% [11, 41, 42]. This is possibly due to the method of its construction [11]. Specifically, more than 100 fine-grained emotion keywords searching out images were clustered into seven target classes: anger, disgust, fear, happiness, sadness, surprise, and neutral expression. Here, clustering errors could result in noisy and subjective labels.

The original FER-2013 was introduced for the sub-challenge of ICMLW'13 [11], and consists of 35,887 detected faces in grayscale at a size of 48-by-48 pixels:

FER-2013 DB		Number of Faces (Ratio)		
		Non-Alignable	Alignable	Total
Data	Train	3764 (15.0 %)	21346 (85.0 %)	25110
	Valid	537 (15.0 %)	3051 (85.0 %)	3588
	Test	543 (15.1 %)	3046 (84.9 %)	3589
State	Non-Aligned	X_{NA}	X_A	X
	Aligned	-	Z_A	-

Table 1: Alignment performance and summary of experimental datasets according to aligned facial states.

28,709 faces for training, 3,589 for public test, and 3,589 for private test. In the present study, after removing 11 non-number-filled training images, we randomly divide the training data into two parts: 25,110 faces (about seven eighths) for training our models and 3,588 faces for validation. For fair comparison with previous studies on this database, we only use the private test data for evaluation, and do not use any public test data for validation or evaluation.

3.2. Face alignment

For face registration, we conduct a conventional 2-D alignment using IntraFace [25, 43], a publicly-available landmark detector. After localizing the 49 predefined facial landmarks, face images are automatically rotated and cropped based on the eye coordinates, and finally resized into 48-by-48 pixels. Notice that manual preprocessing steps are not employed, *i.e.*, erroneously-aligned and non-alignable faces are not removed, and so are used in later processing stages.

Alignment performance on the FER-2013 is reported in Table 1. Here, 85 percent of faces are "alignable", meaning that IntraFace can provide the set of landmarks with the corresponding confidence score and pose angle of pitch, yaw and roll. Figure 2 shows histograms of the number of alignable faces as a function of the pose angle, some aligned examples, and their localized landmarks. Due to failed

landmark detection, 15 percent of faces are "non-alignable". Most of such failures are under extreme pose, occlusion, and/or bad illumination conditions. Examples of non-alignable faces are depicted in Figure 7. Clearly, face alignment in real-world situations is not always successful, posing a specific challenge for FER research.

3.3. Formation of experimental datasets

We form experimental datasets based on the alignment results in Sect 3.2. More specifically, after the data division described in Sect 3.1, face alignment with IntraFace is performed on each of training, validation, and testing data. Here, depending on whether IntraFace can localize the facial landmarks or not, face images are able to be aligned (alignable) or not to be aligned (non-alignable).

Non-aligned and aligned facial states thus constitute distinct classes, and we form our experimental datasets accordingly. Let $X = \{\mathbf{x}\}$ be a set of *whole* faces before the 2-D alignment, which are in the *non-aligned* state. We further divide this set X into two disjoint subsets: $X_{NA} = \{\mathbf{x}_{NA}$ is a non-alignable $\mathbf{x}\}$, a set of *non-alignable* faces in the *non-aligned* state, and $X_A = \{\mathbf{x}_A$ is an alignable $\mathbf{x}\}$, a set of *alignable* faces in the *non-aligned* state. Furthermore, let $Z_A = \{\mathbf{z}_A$ is obtained from $\mathbf{x}_A$ by face alignment$\}$ be a set of *alignable* faces in the *aligned* state. Hence, a one-to-one correspondence exists between X_A and Z_A. See Table 1 for summary of the notations.

4. Information Fusion of Non-Aligned and Aligned Facial States

This section discusses four scenarios with information fusion of non-aligned and aligned states and for comparison two scenarios without information fusion. Next, we present individual deep models and their ensembles used in the considered scenarios. We then move onto the experimental result and analysis.

4.1. Scenarios of information fusion

As shown in Sect. 3.2, face alignment during preprocessing is not always successful in uncontrolled environments. Therefore, both alignable faces (AF) and non-alignable faces (N-AF) exist in real-world situations, and FER researchers are faced with the question of using either aligned or non-aligned state information, or using both kinds of information during system development. To answer this question, we consider six scenarios depicted in Figure 3. Firstly, the scenarios are categorized according to three types of *training dataset* for deep models, *i.e.*, X (in non-aligned facial states), Z_A (in aligned facial states), and $X+Z_A$. We mark them with red, blue, and yellow color,
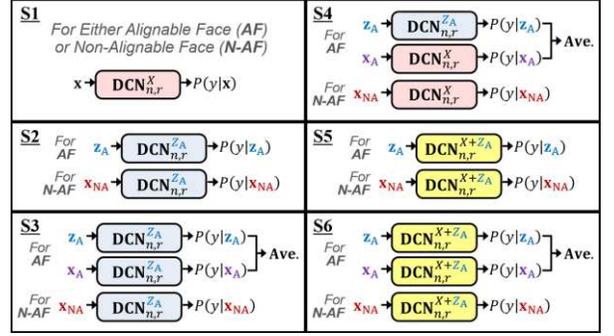
Figure 3: Scenarios to combine information of aligned and non-aligned facial states. See Sect. 3.3 and Sect. 4.2 for the notations of datasets and deep models, respectively.
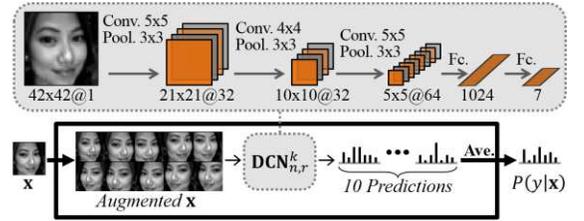
Figure 4: Architecture of the discriminative DCN and evaluation procedure with data augmentation.

respectively, in Figure 3. The dataset of $X+Z_A$ is constructed by *merging* X and Z_A into one set. The expectation is that models trained using $X+Z_A$ could extract some features representing the *fused knowledge* about non-aligned and aligned states, since these models simultaneously handle faces in both states during learning.

Moreover, they are categorized according to whether the AF are evaluated with a *decision-level* fusion. The scenarios with the fusion compute an average of posterior class probabilities estimated from deep models, $P(y|\mathbf{x}_A)$ and $P(y|\mathbf{z}_A)$, for combining information of non-aligned and aligned facial states. On the other hand, the scenarios without the fusion use either $P(y|\mathbf{x}_A)$ or $P(y|\mathbf{z}_A)$. Now, we introduce the details of each scenario.

Scenarios without information fusion. The first two scenarios, S1 and S2 in Figure 3, do not use any information fusion and are designed for comparison. The models in S1 are trained and evaluated using X without considering whether data are AF or N-AF. In other words, FER systems in S1 do not apply face alignment in preprocessing steps, and they have been *commonly* used for the FER-2013 database. In contrast, the models in S2 are trained only using Z_A. Then, the AF are tested in the aligned state $\mathbf{z}_A$ to maintain a consistency with training, while the N-AF are evaluated inevitably in the non-aligned state $\mathbf{x}_{NA}$.

Scenarios with information fusion 1. The next two scenarios, S3 and S4, fuse information of non-aligned and

Scenario			Individual DCN models										Ensemble	
			n = Raw			n = iNor			n = cEnh			Mean	Maj.	Ave.
No.	Train Set	Test Set (AF, N-AF)	r = 1	r = 2	r = 3	r = 1	r = 2	r = 3	r = 1	r = 2	r = 3		Vote	Rule
S1	X	(X_A, X_{NA})	70.38	69.24	69.55	69.63	70.08	69.30	69.18	69.49	69.16	69.56	72.39	72.47
S2	Z_A	(Z_A, X_{NA})	65.48	64.84	63.95	64.17	64.86	65.09	65.84	64.78	65.00	64.89	66.76	67.46
S3	Z_A	(Z_A+X_A, X_{NA})	63.92	63.97	63.39	63.14	63.81	63.97	64.45	64.20	63.69	63.84	65.98	66.93
S4	X, Z_A	(Z_A+X_A, X_{NA})	70.88	70.74	**70.99**	**70.80**	70.88	70.38	**71.08**	71.22	**71.25**	**70.91**	72.97	72.81
S5	$X +Z_A$	(Z_A, X_{NA})	70.66	70.41	69.80	69.07	69.16	69.49	69.30	70.66	69.99	69.84	72.33	72.42
S6	$X +Z_A$	(Z_A+X_A, X_{NA})	**71.86***	70.88	70.83	70.02	70.33	**70.69**	70.63	**71.75**	71.22	**70.91**	**73.31**	**73.31**

Table 2: Classification accuracy (%) of individual deep models and their ensemble for each experimental scenario on the FER-2013 DB. For a given model or an ensemble (each column), the highest accuracy indicating the best scenario is written in bold. The asterisk* denotes the best single DCN. Note that test sets of all scenarios contain exactly the same face images, indicating a fair comparison. The evaluation strategies for test sets differ in how to deal with alignable faces (AF) and non-alignable faces (N-AF).

aligned states in evaluating the AF. S3 is similar to S2, except that it combines $P(y|\mathbf{x}_A)$ and $P(y|\mathbf{z}_A)$ for the AF. In S4, we consider two types of models, which are trained using X and Z_A, respectively. Then, the N-AF are tested with the models trained using X in the same way as S1. The AF are tested with information fusion of $\mathbf{x}_A$ and $\mathbf{z}_A$. Here, $P(y|\mathbf{x}_A)$ and $P(y|\mathbf{z}_A)$ in S4 are obtained from the two types of models respectively, whereas these two class probabilities in S3 are obtained from only the one type of models.

Scenarios with information fusion 2. In the last two scenarios, S5 and S6, the models are trained using the merged set of $X+Z_A$, and we expect them to learn fused knowledge of non-aligned and aligned facial states. The evaluation procedure in S5 is identical to that in S2, without information fusion for the AF. In contrast, S6 applies information fusion of combining both facial states to the testing phase for the AF as well as to the training phase of the models.

4.2. Deep models and their ensemble

To achieve "diverse" FER decisions, which are necessary for a good ensemble, we design nine discriminative DCNs. These deep models are trained using three different methods for *input normalization* as well as using three different *weight initialization* by changing the random seed numbers. For input normalization, the pixel values are rescaled by applying the min-max normalization (denoted as *Raw*), the illumination variation among faces is reduced by applying the illumination normalization (*iNor*) based on the isotropic diffusion [44], or the global contrast of faces is increased by applying the contrast enhancement (*cEnh*) based on histogram equalization. Examples of input normalization are shown in Figure 7.

As illustrated in Figure 4, each DCN consists of three stages of convolutional and max-pooling layers, followed by two fully-connected layers. The convolutional layers use 32, 32, and 64 filters of size 5x5, 4x4, and 5x5, respectively. In the max-pooling layers, overlapping-pooling is applied with the kernels of size 3x3 and stride 2, and the size of

resulting maps becomes halved. The fully-connected hidden and output layers contain 1024 and 7 neurons, respectively, where each output neuron corresponds to each expression class of FER. For nonlinearity, Rectified linear unit (ReLU) activation is used in all convolutional and penultimate layers, while softmax activation is used in the output layer. Following the notation rule in [32], our model can be denoted as 1x42x42 - 32C5 - MP3 - 32C4 - MP3 - 64C5 - MP3 - 1024N - 7N and we shall use this rule in the rest of this paper for brevity. The DCN model is learned using the augmented training data obtained from label-preserving translation and reflection. At the evaluation phase, to maintain a consistency with the training phase, ten patches extracted from each face image are fed to the model, and the corresponding ten predictions are averaged. For other training details, refer to Appendix A.

Depending on the training dataset as described in Sect. 3.3, we organize our nine DCNs into a unit of ensemble:

$$\mathbf{DCN}^k = \left\{ \mathbf{DCN}_{n,r}^k \mid n \in N, r \in R \right\} \quad (1)$$

where $k \in K = \{X, Z_A, X + Z_A\}$ denotes the type of training dataset, $n \in N = \{Raw, iNor, cEnh\}$ denotes the input normalization method, and $r \in R = \{1, 2, 3\}$ denotes the random seed number for weight initialization. To combine the nine decisions in the ensemble, we apply two widely used fusion rules, the majority vote rule and the average rule. In order to select a final class, the majority vote directly uses the predicted labels to compute the largest number of votes, while the average rule uses the posterior class probabilities to compute the highest mean class score.

4.3. Experimental result

Overall performance. The 7-class FER test accuracies of DCNs and their ensembles under the six experimental scenarios are reported in Table 2. We first observe that both of the best single model and the best ensemble are achieved under the S6 scenario. Second, the scenarios S4 and S6 show better ensemble performance than other scenarios. Furthermore, the ensemble accuracies of S2 and S3 are
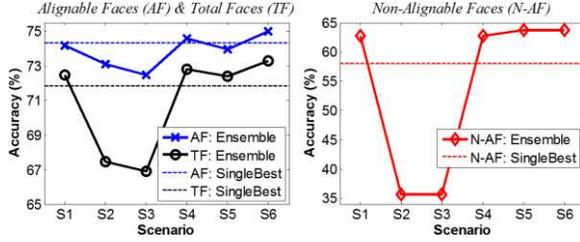
Figure 5: Classification accuracy (%) of the ensemble in each scenario for alignable faces, non-alignable faces, and total faces. The horizontal lines show performance of the single best DCN model in S6.

lower than those of other scenarios. The aforementioned observations will be examined deeply in the followed analysis with Figure 5. Notice that there is no much difference or clear trend in the ensemble accuracies depending on the majority vote and the average rule. It indicates that what to combine for ensemble has more influence on the final performance rather than how to combine decisions.

It is worth noting that S6 not only yields higher ensemble accuracy than S4 but also uses the less number of models. Specifically, the ensemble in S6 includes 9 DCNs (trained using $X + Z_A$), whereas that in S4 does 18 DCNs (9 trained using X and 9 using Z_A). It shows the strengths of the scenario S6, *i.e.*, the *efficiency* in the evaluation time with less models as well as the superiority in FER performance. For a later usage in Sect. 5.3, the best ensemble using the average rule in S6 is denoted as **ES**$_6$.

Accuracies for alignable and non-alignable faces. The six different scenarios perform differently given AF and N-AF. Here, we identify the causes for performance differences. Figure 5 shows the ensemble accuracies using the average rule in each scenario, separately computed for AF and N-AF. We also plot overall test accuracies for total faces (TF) as reported in the last column in Table 2. Accuracy for AF is computed as the proportion of correctly estimated AF over the whole AF, and the same for N-AF.

For both AF and N-AF, the ensemble performances of S4 and S6 are improved over the single best DCN. Also, the ensemble of S6 achieves the highest accuracies for both. These results suggest the following.

■ Using the merged training dataset of $X+Z_A$ is beneficial in that the *fused knowledge* about non-aligned and aligned facial states can be learned. It is verified by the improved accuracies for both AF and N-AF in S6.
■ For AF, applying the late information fusion to combine $P(y|\mathbf{x}_A)$ estimated from non-aligned states and $P(y|\mathbf{z}_A)$ from aligned states is beneficial. It is supported by the improved accuracies for AF in S4 and S6.

In addition, as shown in the right graph of Figure 5, much

lower accuracies for N-AF in S2 and S3 are reported, resulting in poor ensemble performances. It is because a non-alignable face in non-aligned states ($\mathbf{x}_{NA}$) for testing is fed to the DCNs trained using AF in aligned states (Z_A), resulting in an inconsistency of data characteristics between training and testing.

Notably, S1 in which face alignment is not applied is a common setting for the FER-2013 database. Compared to S1, S4 and S6 show better ensemble accuracies for AF. This suggests that even if some faces are non-alignable, it is better to apply face alignment to other alignable faces and then to conduct the improved information fusion.

5. Mapping from Non-Aligned State to Aligned State

This section describes our mapping function from non-aligned states to aligned states. When compared to auto-encoding functions, we show the effectiveness of our mapping. Then, we present experimental results including qualitative analysis and FER accuracy using the individual mappings as well as the final ensemble performance.

5.1. Deep models to learn our desired mapping

The basic idea driving this section is that a mapping from non-aligned facial states to aligned ones can represent similarity transformations of face alignment. First, we examine whether this mapping can be *learned* using deep neural networks (known as universal approximators). Specifically, given an alignable face in its non-aligned state $\mathbf{x}_A$, a deep model is trained to generate the corresponding aligned state $\mathbf{z}_A$ by minimizing the L2 Euclidean objective function. Therefore, X_A (as inputs) and Z_A (as target outputs) containing only AF are used for training the models, as depicted in Figure 6.

Preliminary experiments compared deep architectures in learning mapping, *e.g.*, multi-layer perceptrons (MLPs) and DCNs (see Appendix B for more details). The final DCN consists of two stages of convolutional and max-pooling layers followed by two fully-connected layers: 1x42x42 - 64C5 - MP3 - 64C5 - MP3 - 1000N - 1764N. This alignment-mapping network (AMN) is denoted as **AMN**$_n$ where $n \in N = \{Raw, iNor, cEnh\}$ is the input normalization. For training details, see Appendix C.

Notice that there are no ground-truth outputs for X_{NA}, *i.e.*, the aligned states of N-AF. We expect that the mapping output for X_{NA} (usually with extreme pose angles and occlusions) can be in the closely-aligned state if **AMN**$_n$ satisfies the following two requirements. First, **AMN**$_n$ can rotate, crop, and resize faces in the non-aligned states by learning important features for face alignment. Second,
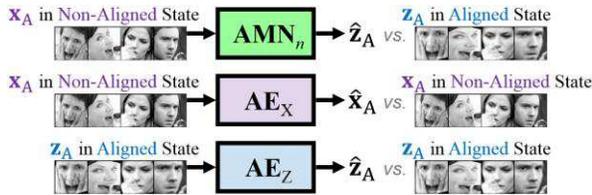
Figure 6: Training procedures for AMNs and AEs to learn the examined mappings.
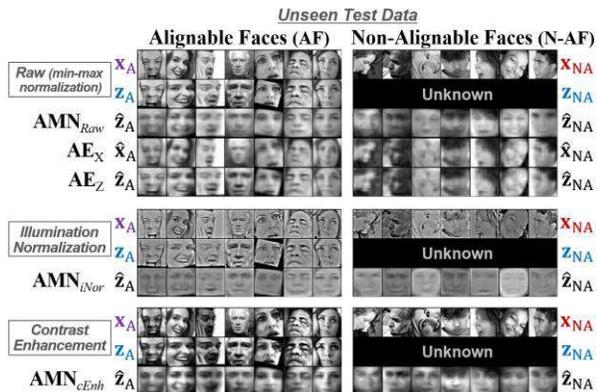
Figure 7: Examples of inputs, target outputs, and estimated outputs yielded from AMNs and AEs.

AMN_n can be generalized successfully, thus working well not only on AF but also on unobserved N-AF. Sect. 5.4 provides qualitative analysis to confirm these properties using the mapping output for X_A and for X_{NA}.

5.2. Comparison with auto-encoding functions

A popular method for feature extraction is using neural networks that learn auto-encoding functions. These networks are conventionally trained to produce outputs which are identical to inputs, and the activations of hidden neurons are used as features. Here, whether AMN_n provides better features than auto-encoding networks (AEs) in terms of classification is examined. As shown in Figure 6, we compare our AMN_n, which learns the mapping from X_A to Z_A, with two AEs. One network learns an auto-encoding function from X_A to X_A using raw face images, denoted as AE_X. The other network similarly learns an auto-encoding function from Z_A to Z_A, denoted as AE_Z. For a fair comparison, we use the same architectures and training schemes for AMN_n, AE_X, and AE_Z.

For qualitative analysis, outputs of the examined networks are observed. Then, in terms of FER performance, we investigate the effectiveness of AMN_n as follows. The 1000-D hidden activations of penultimate layer in AMN_n are extracted. Then, we use these features to train a 3-layer MLP classifier of 1000N - 1000N - 1000N - 7N. For training details, see Appendix C. This procedure is repeated

Model	L2 Euclidean Loss	Accuracy (%)
AE_X	6.77	55.45
AE_Z	4.57	59.52
AMN_{Raw}	18.09	62.94
AMN_{iNor}	16.74	59.99
AMN_{cEnh}	30.08	**63.36**

Ensemble Type	Accuracy (%)
ES_6	73.31
ES_6+AE_X	73.29
ES_6+AE_Z	73.31
ES_6+AMN_{Raw}	73.47
$ES_6+AMN_{Raw}+AMN_{iNor}+AMN_{cEnh}$	**73.73**

Existing Method		Acc. (%)
Zhang *et al.* [45]	A DCN using multi-task loss with using **external databases**	**75.10**
Kim *et al.* [18]	An ensemble of 36 DCNs in a hierarchical committee	72.72
	A DCN using cross-entropy loss	70.58
Devries *et al.* [46]	A DCN using multi task loss	67.21
Tang [42]	A DCN using L2-SVM loss	71.16
	A DCN using cross-entropy loss	70.1
Ionescu *et al.* [47]	Multiple kernel learning based on SIFT descriptors	67.48

Table 3: Performance comparison of AMNs and AEs (upper), ensemble accuracies obtained by adding AMNs or AEs to discriminative DCNs in ES_6 (middle), and existing results on the FER-2013 database (bottom).

for each of AMN_n, AE_X, and AE_Z, for comparison.

5.3. Improving the ensemble for FER

To improve the best ensemble from the scenario S6 in Sect. 4.3, *i.e.*, ES_6, we additionally use decision-level information obtained using AMN_n. Specifically, the features extracted from AMN_{Raw}, AMN_{iNor}, and AMN_{cEnh} are fed to the 3-layer MLP classifiers in Sect. 5.2, yielding estimated posterior class probabilities. After that, we compute the average of these class probabilities and the output probability from ES_6. This improved ensemble is denoted as $ES_6+AMN_{Raw}+AMN_{iNor}+AMN_{cEnh}$, used for our final FER system.

The same procedure is conducted with AE_X and AE_Z, resulting in ES_6+AE_X and ES_6+AE_Z, respectively. To ensure that the performance improvement is not just coming from adding other information, ES_6+AMN_{Raw} is also compared with the aforementioned ensembles.

5.4. Experimental result

Figure 7 shows examples of the inputs, the corresponding target outputs, and the estimated outputs of examined networks. Outputs of AE_X and AE_Z are identical to their input images. Notice that outputs of AE_Z for unseen N-AF

are in non-aligned states. In contrast, **AMN**$_{Raw}$, **AMN**$_{iNor}$, and **AMN**$_{cEnh}$ provide *closely-aligned* outputs for both the AF and the N-AF. These results support that our mapping represents geometric transformations of face alignment. Interestingly, outputs of **AMN**$_{cEnh}$ are less blurred and clearly show facial expressions while preserving identity information. It could be linked to higher accuracy using **AMN**$_{cEnh}$ in Table 3, by giving better features for FER.

In the upper part of Table 3, we report the L2 Euclidean loss and the classification accuracy of extracted hidden features for each network. Lower L2 loss values are achieved by **AE**$_X$ and **AE**$_Z$, indicating that they are trained well for their purpose of auto-encoding functions. However, the classification accuracies using **AMN**$_n$ are higher than those of the AEs. It demonstrates that our AMNs provide informative features for FER and correct unsuitable knowledge particularly by transforming N-AF to be closely-aligned.

In the middle part of Table 3, ensemble performances using the examined networks are shown. Combining **ES**$_6$ and AE decreases (**ES**$_6$+**AE**$_X$) or does not improve (**ES**$_6$+**AE**$_Z$) the accuracy of **ES**$_6$. In contrast, **ES**$_6$+**AMN**$_{Raw}$ and **ES**$_6$+**AMN**$_{Raw}$+**AMN**$_{iNor}$+**AMN**$_{cEnh}$ achieve higher accuracies than only using **ES**$_6$. It indicates that decision information from our AMNs can be complementary to that of discriminative DCNs in **ES**$_6$, thus improving FER.

In the bottom part of Table 3, the existing results on the FER-2013 database are shown for performance comparison. Without using external databases, our final system of **ES**$_6$+**AMN**$_{Raw}$+**AMN**$_{iNor}$+**AMN**$_{cEnh}$ yields the best FER accuracy. Note that the deep model in [45] has used much bigger architecture than ours as well as huge external data for its multi-task loss. Compared to the ensemble in [18], our ensemble includes fewer models having the similar architecture, but we achieve better results. Compared to other single DCNs trained only using the FER-2013, our single best DCN in S6 yields a higher accuracy of 71.86 %.

6. Conclusion

Towards automatic facial expression recognition (FER), we present a framework based on an ensemble of deep convolutional neural networks. We aim at overcoming a specific challenge to FER researchers: there are non-alignable faces in real-world conditions. To start, alignment results on a challenging FER database are analyzed. Then, we evaluate possible scenarios of information fusion with discriminative deep models. Here, an efficient and effective means to combine information of aligned and non-aligned facial states is proposed. In order to better deal with non-alignable faces, we introduce alignment-mapping networks which learn the operations in face alignment. Using these networks as well as the discriminative deep models, the final ensemble achieves excellent performance on the examined database collected in the wild. We believe that the proposed approach can be applied not only to FER but also to other face analysis research using face alignment under unconstrained conditions.

Acknowledgement

Appendix

A. Training details for discriminative DCNs in Sect. 4.2. We use the MatConvNet toolbox [48] on NVIDIA GeForce GTX 690 GPUs. To train the DCNs, stochastic gradient descent is used to minimize the cross-entropy objective function with a mini-batch size of 200 samples. The dropout probability of the penultimate layer is set to 0.5, and weight decay of 0.0001 and momentum of 0.9 are also applied. The initial learning rate is set to 0.01 and reduced by a factor of 2 at every 25 epoch where the number of total epochs is 100. Moreover, the training data were augmented by 10 times, through using 5 crops of size 42x42 (1 from resizing an original 48x48 face and 4 from extracting its 4 corners) and their horizontal flopping.

B. Preliminary experiments for architecture selection of AMN$_n$ **in Sect. 5.1.** To learn the mapping from the non-aligned state to the aligned state, we have explored several candidates: MLPs having 2 or 3 fully-connected layers (FC), DCNs having 1 or 2 convolutional layers (CONV) "without" max-pooling layers (MP) followed by 2 or 3 FC, and DCNs having 1 or 2 CONV "with" MP followed by 2 or 3 FC. After comparing the L2 Euclidean validation loss, we have finally selected the best DCN described in the main text. We also empirically find that "using CONV and MP" and "increasing the number of hidden neurons in FC (500→1000)" are beneficial for learning the desired mapping, while increasing the number of FC (2→3) is not.

C. Training details for AMN$_n$ **in Sect. 5.1 and for MLP classifiers in Sect. 5.2.** We apply the similar training schemes introduced in Appendix A. For AMNs, a mini-batch size of 500 samples and a constant learning rate of 0.0001 during a total 1000 epochs are used. For MLP classifiers, a mini-batch size of 200 is used, while the initial learning rate is set to 0.01 and halved at every 100 epoch where the number of total epochs is 400.

References

[1] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognit.*, 36(1):259–275, 2003.

[2] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect Comput.*, 1(1):18–37, 2010.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1113–1133, 2015.

[4] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):2106–2111, 2009.

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia Mag.*, 19(3):34–41, 2012.

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV Workshops*, 2011.

[7] P.-L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, and Y. Bengio. FER-2013 Face Database. *Technical report*, 1365, Université de Montréal, 2013. http://www.kaggle.com/c/challenges-in-representation-learning-facialexpression-recognition-challenge.

[8] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, and M. Pantic. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *The 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013.

[9] A. Dhall, OVR. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ICMI*, 2015.

[10] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG*, 2011.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, and A. Courville. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[12] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, 2012.

[13] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014

[14] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR Workshops*, 2015.

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

[16] S. E. Kahou, *et al*. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 2015. (online)

[17] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *ICMI*, 2015.

[18] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 2016. (online)

[19] J. Alabort-i-Medina and S. Zafeiriou. Bayesian active appearance models. In *CVPR*, 2014.

[20] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. P. Zafeiriou. Feature-Based Lucas–Kanade and Active Appearance Models. *IEEE Trans. Image Process.*, 24(9):2617–2632, 2015.

[21] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

[22] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

[23] J. Alabort-i-Medina and S. Zafeiriou. Unifying holistic and parts-based deformable model fitting. In *CVPR*, 2015.

[24] E. Antonakos, J. Alabort-i-Medina, and S. Zafeiriou. Active pictorial structures. In *CVPR*, 2015.

[25] X. Xiong and F. de la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[26] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[27] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, 1990.

[28] A. J. C. Sharkey. On combining artificial neural nets. *Connection Science*, 8(3-4):299–314, 1996.

[29] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9):699–707, 2001.

[30] C. A. Shipp and L. I. Kuncheva. Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3(2):135–148, 2002.

[31] L. I. Kuncheva. Diversity in multiple classifier systems. *Information Fusion*, 6(1):3–4 (Editorial), 2005.

[32] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, 2012.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[35] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, 2014.

[36] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013.

[37] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014.

[38] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.

[39] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

[40] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust Statistical Face Frontalization. In *ICCV*, 2015.

[41] A. Ruiz, J. Van de Weijer, and X. Binefa. From Emotions to Action Units with Hidden and Semi-Hidden-Task Learning. In *ICCV*, 2015.

[42] Y. Tang. Deep Learning with Linear Support Vector Machines, In *ICML Workshops*, 2013.

[43] F. de la Torre, *et al*. Intraface. In *FG*, 2015.

[44] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *The 4th International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[45] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning Social Relation Traits from Face Images. In *ICCV*, 2015.

[46] T. Devries, K. Biswaranjan, and G. W. Taylor. Multi-task learning of facial landmarks and expression. In *2014 Canadian Conference on Computer and Robot Vision*, 2014.

[47] R. T. Ionescu, M. Popescu, and C. Grozea. Local learning to improve bag of visual words model for facial expression recognition. In *ICML Workshops*, 2013.

[48] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *The 23rd ACM Conference on Multimedia*, 2015.