J Multimodal User Interfaces (2016) 10:173–189
DOI 10.1007/s12193-015-0209-0

ORIGINAL PAPER

Hierarchical committee of deep convolutional neural networks for robust facial expression recognition

Bo-Kyeong Kim[1] · **Jihyeon Roh**[1] · **Suh-Yeon Dong**[1] · **Soo-Young Lee**[1]

Received: 18 November 2015 / Accepted: 16 December 2015 / Published online: 16 January 2016

Abstract This paper describes our approach towards robust facial expression recognition (FER) for the third Emotion Recognition in the Wild (EmotiW2015) challenge. We train multiple deep convolutional neural networks (deep CNNs) as committee members and combine their decisions. To improve this committee of deep CNNs, we present two strategies: (1) in order to obtain diverse decisions from deep CNNs, we vary network architecture, input normalization, and random weight initialization in training these deep models, and (2) in order to form a better committee in structural and decisional aspects, we construct a hierarchical architecture of the committee with exponentially-weighted decision fusion. In solving a seven-class problem of static FER in the wild for the EmotiW2015, we achieve a test accuracy of 61.6 %. Moreover, on other public FER databases, our hierarchical committee of deep CNNs yields superior performance, outperforming or competing with state-of-the-art results for these databases.

Keywords Hierarchical committee · Exponentially-weighted decision fusion · Deep convolutional neural network · Facial expression recognition

✉ Bo-Kyeong Kim
bokyeong1015@gmail.com

Jihyeon Roh
rohleejh@kaist.ac.kr

Suh-Yeon Dong
suhyeon.dong@gmail.com

Soo-Young Lee
sylee@kaist.ac.kr

[1] Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

1 Introduction

Recognizing human emotion has attracted considerable attention because of its many potential applications, such as human-robot interaction, clinical monitoring, call-center systems, and marketing. In addition to analyzing people's speech and gesture, one of the major parts in emotion analysis is facial expression recognition (FER). For real-world applications of FER, it is crucial to handle visual information captured under uncontrolled illumination, varying poses, and various subjects, i.e., "in the wild" environments.

With this engineering demand for robust FER in the wild, many grand challenges have been held [10–12,16,42,54]. One representative challenge is Emotion Recognition in the Wild (EmotiW) [10–12]. The EmotiW challenges organized in 2013 and 2014 [10,11] mainly focused on audio-video emotion recognition (acted facial expression recognition in the wild, AFEW), whereas the third EmotiW2015 [12] provided two sub-competitions: AFEW and image-based static facial expression recognition in the wild (SFEW). We believed that research outcomes from SFEW could be fundamental but core techniques for studying AFEW, leading to our participation in the SFEW sub-competition of EmotiW2015 [26].

This paper is an extension of our previous work [26] for EmotiW2015. In [26], a pattern recognition framework to improve committee machines of deep convolutional neural networks (deep CNNs) was proposed and tested on the SFEW2.0 database. Here, we discuss our proposed method and experimental results more deeply. Furthermore, we report new results evaluating our method on three other FER databases: the Facial Expression Recognition 2013 database (FER-2013) [16], the Toronto Face Database (TFD) [50], and the GENKI-4K database [57].

As a base system for robust FER, we use a standard committee of deep CNNs, also known as the multi-column deep neural network (MCDNN) [6]. It yields an ensemble of outputs from multiple deep CNNs with a simple averaging decision rule in a single structure level. Because the MCDNN has already proved its outstanding performance in many visual classifications, we expect that its excellence in recognition could also be demonstrated in the FER problem. More importantly, we explore two simple yet effective ways to improve the MCDNN: training individual models that are *more diverse* and forming a better committee in both *decisional* and *structural* aspects. The former is achieved by varying the network architecture of deep CNNs in addition to applying the commonly-used strategies in the MCDNN (i.e., different input normalization and different random weight initialization). The latter is achieved with a better ensemble rule based on an exponentially-weighted decision fusion and with a hierarchical committee architecture having several structural levels in combining decisions. The overall system for robust FER is shown in Fig. 1. On the SFEW2.0 database, our method yields a test accuracy of 61.6 %, significantly higher than the SFEW baseline of 39.1 %. On other public databases, we also obtain good FER accuracies, outperforming or competing with previous state-of-the-art results for these databases.

The remainder of this paper is structured as follows. Section 2 briefly reviews related works on committee machines and motivating studies about neural network ensembles. Section 3 introduces our proposed approach regarding the hierarchical committee of deep CNNs. Section 4 describes the databases used in this study and the procedure of face registration. Section 5 describes the experimental method and results on the SFEW2.0 database, and Sect. 6 presents those on other public FER databases. In Sect. 7, the paper is concluded with an outlook on future work.

2 Related work

Over the past few decades, committee machines have been actively studied, with names such as classifier ensembles, ensemble-based systems, multiple classifier systems, and mixture of experts [37]. They make a team consisting of several member classifiers and combine individual decisions from their members in order to obtain a final decision for an input data sample. Committee machines have been applied extensively in various research fields including vision [35,48], speech [41,59], text [3], and bio-data [4], because they yield better performance than a single classifier, mainly for the following reason [13]: good generalization while reducing the risk of selecting poorly-performing classifiers and expanding the possible hypothesis space. For designing these good committees to outperform their individual members, a key requirement is known as diversity [2,45]. The term *diversity* refers to uncorrelated, independent, or different errors and decisions from various member classifiers. Therefore, it could provide a chance to correct misclassifications from some members by yielding complementary information for input data.

With recent advances in deep learning and parallel computing, forming a committee of multiple deep neural networks was presented in [7,8], has attained impressive successes [5,6,28], and now becomes a widely used approach [1,49,60]. In past iterations of the EmotiW challenge, ensembles of deep neural networks have also been applied successfully. The EmotiW2013 winner [23] learned several modality specific deep models, including a deep CNN for video and a deep belief network for audio, and aggregated their outputs using random-search-based weighting. The EmotiW2014 winner [34] used the deep CNN activations along with other features as the input of partial least squares classifiers and conducted the decision level fusion. Moreover, many top-performing teams in the 2015 challenge [14,26,62,63] explored the tactic of combining activations from deep models. In this study, we investigate the MCDNN [6], consisting of several deep CNNs whose output posterior probabilities are combined by the simple algebraic mean operation. The MCDNN achieves diverse errors by using different input preprocessing and random weight initialization for training deep CNNs. Because of its superior performance

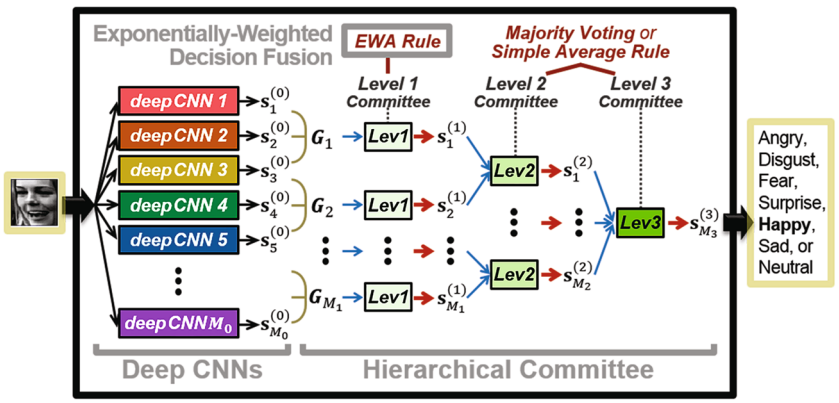

Fig. 1 Overall system for facial expression recognition

in many computer vision problems such as hand-written digit classification, traffic-sign recognition, and object recognition, we choose this MCDNN as a base system for solving the FER problem.

To improve the standard MCDNN, we propose intuitional methods motivated by pioneer studies on committee machines. These studies did not use deep neural networks and deep learning frameworks, but we expect that their main philosophies and methodologies could be valuably applied in the MCDNN to achieve a better committee. As one of the earliest works on neural network ensembles pointed out [18], forming various networks that have different local minima in optimizing objective functions is important to achieve a good ensemble. This is because these different minima eventually contribute to different generalizations on input data, leading to the diversity of decisions. In order to create these various neural networks for an ensemble, four strategies could be considered [15,44]: varying the architecture (e.g., changing the number of hidden units and layers), varying the algorithm employed, varying the initial random weights, and varying the normalization of data. Regarding this aspect towards diversity, we construct multiple deep CNNs having different architectures as well as applying the tactics used in the MCDNN.

Regarding a decisional aspect in formation of a committee, how to combine outputs from individual members has been extensively investigated [27,29,37]. The ensemble rules could be categorized into two groups, trainable and non-trainable rules, depending on whether the combination weights are trained or not. The trainable (or dynamic) rules include the stacked generalization [58], in which a meta classifier learns how the individual classifiers make errors, and the mixture-of-experts [21], in which a gating network provides instance-specific weights to select the most proper classifiers for each instance. The non-trainable (or static) rules could be applied to either class labels or class-related continuous scores. The ensemble rules using class labels include the behavioral knowledge space [19], which records the frequency of possible labeling combinations of all classifiers, and the majority voting. The rules using class-related continuous scores include the decision templates [30], which store the degree of supports for each class as the average of decision profiles, and the algebraic combiners such as the average, the median, and the product rules. In this study, we first consider three widely-used rules to combine outputs of deep CNNs: the majority voting rule, the median rule, and the simple average rule. The majority voting rule selects a class with the largest number of predicted labels, while the median and the simple average rule decide a class with the highest median and mean of posterior class probabilities. A limitation of these rules is that the individuals have equal rights for participation so that any reliability or importance on each of their decisions is not considered. Thus, we introduce an effective combination rule based on exponential weighting to give more weights to well-performing individuals.

Regarding a structural aspect in formation of a committee, hierarchical or multi-level architectures were explored. In [32,48], hierarchical ensembles efficiently combined the outputs of different classifiers trained on heterogeneous features. In [22,53], hierarchical architectures for neural network ensembles were studied to divide a difficult problem into easier ones, i.e., the divide-and-conquer strategy, with a statistical framework. In this study, we build a hierarchical committee which can make decisions that are more reliable. As structural levels in the committee become higher, the consensus of multiple sub-groups could be formed, thus enhancing the reliability of decisions. Moreover, we empirically show that placing some members to be overlapped in the hierarchy could be beneficial for the divide-and-conquer strategy by increasing the diversity of decisions.

3 Proposed approach

3.1 Deep CNNs as individual members

A deep CNN consists of several feature extraction stages (with alternating convolutional and pooling layers), followed by a recognition stage (with fully-connected layers) [28,31]. Because of its excellent classification ability as well as hierarchical feature development mimicking the human visual system, we select the deep CNN for the base member of a committee as in a standard MCDNN.

To build diverse deep CNNs in standard MCDNNs, being trained with "different training data sets" is the main focus, rather than using "different classifiers". The effect of "different training data" is achieved by several preprocessing methods on the original data such as deformation and normalization. For the effect of "different classifiers", the MCDNNs apply multiple random seeds for weight initialization, but the identical network architectures are used for all individual members. We believe that *various network architectures* also largely contribute to obtaining different classifiers and thus to increasing diversity of decisions in forming a committee. Therefore, we apply various architectures for deep CNNs (i.e., varying the size of receptive fields in convolutional layers and the number of neurons in fully-connected layers) as well as differently preprocessed data and different weight initialization.

Notably, for training deep CNNs in the experiment on the SFEW2.0 data, we investigate several strategies for making use of external databases (FER-2013 and TFD) to overcome the small size of the SFEW2.0 (see Sect. 5.1.3 for the examined strategies). We finally select one type of *trans-*

fer learning scheme [36], i.e., the models pre-trained using large databases (in source domain) are set as initialization and fine-tuned using a relatively small database (in target domain). Based on the success of applying this transfer learning for the SFEW2.0 data, we also use pre-trained models from the FER-2013 data for the experiments on the TFD and the GENKI-4K data.

3.2 Hierarchical committee

3.2.1 Exponentially-weighted decision fusion

A straightforward way to regard the importance of members' decisions is to compute a weighted mean of class scores, assigning the weights as validation accuracies. We denote this as the simple weighted average (SWA) rule. However, when the committee members yield similar accuracies and thus almost equal weights are used, the SWA rule does not differ from the simple average rule. Our exponentially-weighted decision fusion is motivated by considering the aforementioned case. In determining the weights, we adopt an *exponential* function which influences on the differences between numbers (e.g., "$3^1 - 2^1 = 1$" < "$3^2 - 2^2 = 5$"). We expect that this characteristic of exponent could give higher weights to the members with (even slightly) better accuracies.

Let us denote our method as the exponentially-weighted average (EWA) rule, and continue our discussion with mathematical notations. Suppose a member model $m (= 1, \ldots, M)$ with its validation accuracy of z_m provides a posteriori class probability vector $\mathbf{s}_m$ for an input pattern. Then, the final ensemble of M models' decisions in our EWA rule becomes

$$\mathbf{s}_{final} = \frac{\sum_{m=1}^{M} (z_m)^q \mathbf{s}_m}{\sum_{m=1}^{M} (z_m)^q} = \sum_{m=1}^{M} d_m \mathbf{s}_m \tag{1}$$

where a decision weight d_m reflects the normalized significance of the model m's decision ($0 \leq d_m \leq 1$) and an exponent q is a hyper-parameter to determine how much the qualified members are emphasized ($q > 1$) or de-emphasized ($q \leq 1$). Finally, a class with the highest value in the exponentially-weighted class probabilities is chosen. In our experiments, the value of q is selected to provide the maximum validation performance after the fusion.

3.2.2 Hierarchical committee architecture

In forming a committee, we construct a hierarchical architecture with the following procedure according to the two expected merits.

1. Organize the M_0 individual members into the first level sub-groups, $\{\boldsymbol{G}_1,\ldots,\boldsymbol{G}_{M_1}\}$, having some overlapping members. After that, make a decision for each group according to the first level decision fusion rule.
2. Collect all sub-groups' decisions in the l^{th} level, $\{\mathbf{s}_{m_l}^{(l)}, m_l{=}1,\ldots,M_l\}$, where $l\,(=1,\ldots,L-1)$ and m_l are indices for the level and the sub-group, respectively. Then, re-organize them into the $(l+1)^{\text{th}}$ level groups, and make a decision for each group according to the $(l+1)^{\text{th}}$ level decision fusion rule.
3. Repeat step 2 until reaching a final decision at the last L^{th} level.

The first merit is that, decisions that are *more reliable* could come from the strong consensus of multiple sub-groups in higher structural levels. Second, the *increased diversity* of errors could be obtained by setting some members to be overlapped in certain sub-groups. Then, depending on other members in these groups, the overlapping members differently contribute to the next level decision. Because the former is quite intuitive, let us explain the latter with a simple example. Suppose a two-class problem and five member classifiers (a, b, c, d, e) who claim the class label for an input sample as (1, 1, 1, 2, 2), respectively. When they are divided into two sub-groups, $\boldsymbol{G}_1$: a, b, c and $\boldsymbol{G}_2$: c, d, e, with an overlapping member c and the majority voting rule is applied, member c differently contributes to the final decision. More specifically, without grouping, there is no doubt of the selection of class 1 by 3 votes from a, b, and c among the five members. However, with grouping, both classes obtain an equal number of mid-level decisions (the class 1 from $\boldsymbol{G}_1$ and the class 2 from $\boldsymbol{G}_2$) due to the different impacts of member c's claim on both sub-groups, so the final decision depends on the mean class probabilities. We expect that these groups with overlapping members could lead to more various decisions in the low structural levels, finally serving as diverse errors in the last level.

4 Databases and face registration

4.1 Databases

The SFEW2.0 database [12] was created by extracting frames from emotional movie clips in the AFEW data corpus [9]. The task is to assign seven expression labels (angry, disgust, fear, happy, sad, surprise, and neutral) to these frames in close-to-real world conditions. For the training, validation, and test set, the SFEW2.0 contains 958, 436, and 372 images, respectively.

The FER-2013 database, released for ICMLW2013's subchallenge [16], was created using the Google image search API. Because of realistic facial expressions collected from the Internet, large variations reflecting real-world conditions exist in the FER-2013 DB. From this dataset, 28,698 training faces (after removing 11 non-number-filled images from the original training data) and 3589 private testing faces are used for our experiment.

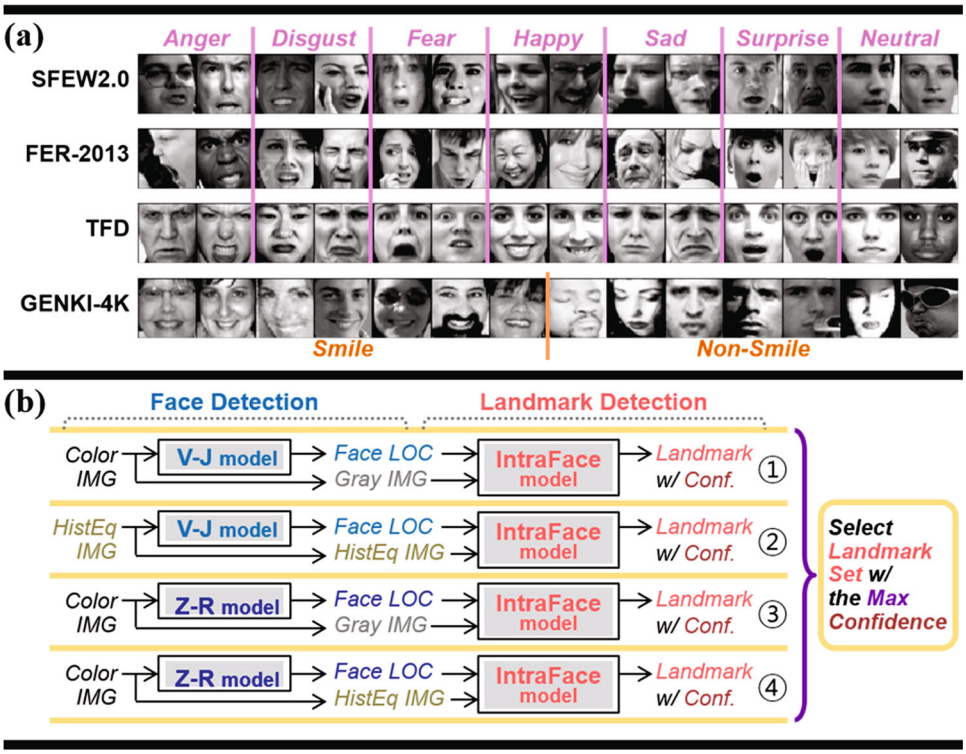

Fig. 2 **a** Examples of face images for each database and **b** our multi-pipeline-based face registration. In **a**, faces of the SFEW2.0 and the GENKI-4K are obtained from **b**

The TFD [50] was constructed by merging together 30 pre-existing face datasets. The faces in TFD are strictly aligned and almost all of them are fully-frontal. From the TFD, 4178 labeled faces are used for our experiment. Notice that both of the FER-2013 and the TFD consist of 48 × 48 gray-scale faces labeled with the identical seven expression categories used in the SFEW2.0 data.

The GENKI-4K [57] database was constructed for studying smile detection under real-world environments. It contains 4,000 images (2162 for smiling and 1,838 for non-smiling), downloaded from the web. Similar to the SFEW2.0 and FER-2013, there are huge variations in lighting conditions, head poses, subjects, and resolutions in this GENKI-4K. Figure 2a depicts examples of face images for each of four databases.

4.2 Face registration

For raw images in the SFEW2.0 and the GENKI-4K, face registration is required to extract face-related information. For face registration, we conduct a conventional 2-D alignment using face/landmark detection. First, a set of landmarks is localized from the detected face location. Then, face alignment is done by removing in-plane rotation and cropping the face region of interest based on the eye coordinates. Finally, the aligned face crop is resized into 48 × 48.

To improve the robustness of face/landmark detection, multiple detection pipelines are designed to produce different landmark estimations, and the best estimation among them is finally chosen for the 2-D alignment. We use the Viola-Jones (V-J) model [56] and the Zhu-Ramanan (Z-R) model [64] for face detection along with the IntraFace model [61] for landmark detection. As shown in Fig. 2b, we consider four single pipelines on the basis of the following observations: (a) some faces, failed by the V-J, could be detected by the Z-R and vice versa, (b) depending on face locations from the V-J and Z-R, the landmark estimation of the IntraFace changes, and (c) the V-J and IntraFace sometimes yield complementary outputs when histogram-equalized images are used as input. Among four possible landmark sets from those pipelines, the landmark set with the highest confidence provided from the IntraFace is eventually selected for alignment.

Our four-pipeline-based alignment performs better than the single-pipeline ones, implying that complementary detection results are obtained from four single pipelines (see our previous work [26] for detailed results of alignment-success rate). Therefore, combining single pipelines for face/landmark detection can boost performance of face registration in real-world conditions.

5 Experiments on the SFEW2.0 database

5.1 Designing and training deep CNNs as individual members

With the aim of obtaining diversity of decisions from individual members, we design 216 deep CNNs using different

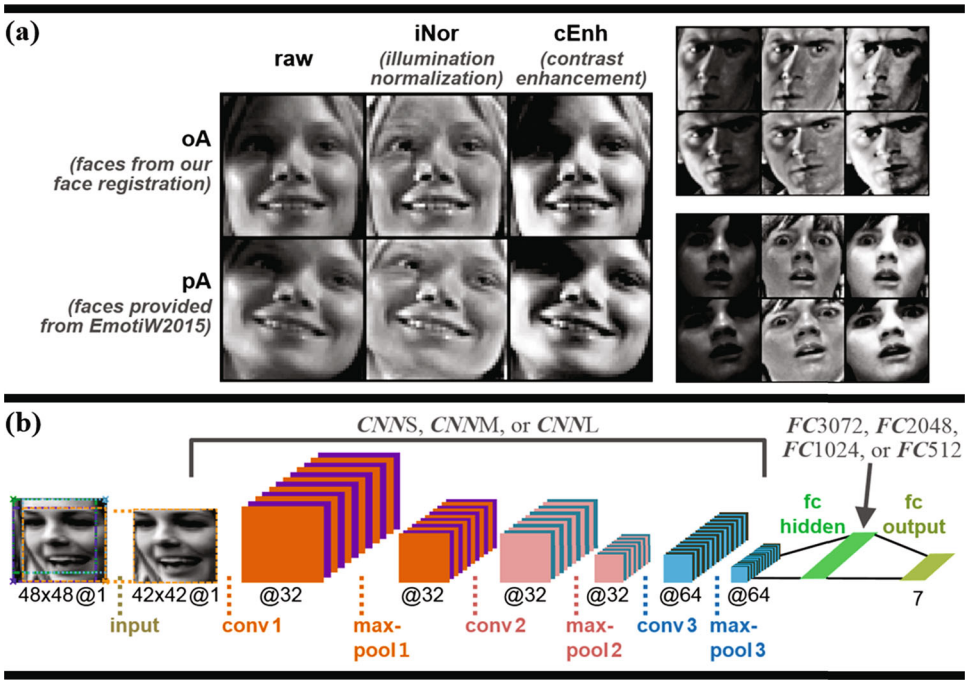

Fig. 3 **a** Examples of different input preprocessing and **b** our deep CNN architecture

input preprocessing methods, network architectures, and random weight initializations. Then, we train these deep models after evaluating several strategies to make use of external data (FER-2013 and TFD) along with the SFEW2.0 data. A single deep CNN is denoted as (2), and detailed explanations of designing and training individual deep CNNs are presented.

$$\mathbf{PREP}\alpha_1,\ \alpha_2 - \{\mathbf{CNN}\beta - \mathbf{FC}\gamma\}_{\mathbf{R}\delta} \tag{2}$$

PREPα_1, α_2: input preprocessing methods,

$$\alpha_1 = \begin{cases} raw & \text{Raw} \\ iNor & \text{Illumination Normalization} \\ cEnh & \text{Contrast Enhancement} \end{cases}$$

$$\alpha_2 = \begin{cases} oA & \text{Our Aligned Faces} \\ pA & \text{Provided Aligned Faces} \end{cases}$$

CNNβ: size of CNN receptive field,

$$\beta = \begin{cases} S & \text{Small} \\ M & \text{Medium} \\ L & \text{Large} \end{cases}$$

FCγ: number of neurons in a fully-connected hidden layer,
$\gamma = \{512, 1024, 2048, 3072\}$

Rδ: random seed number for weight initialization,
$\delta = \{1, 2, 3\}$

5.1.1 Input preprocessing methods

We consider various normalization techniques on differently aligned faces. Each raw image is rescaled from 0 to 1 by a min-max normalization. To reduce illumination variation in images, as used in [23], the isotropic diffusion based normalization [17] from the INface toolbox [47] is applied with the default parameter setting. Moreover, to enhance contrast for each image, as used in [6], we apply the histogram equalization implemented in MATLAB.

For different input deformations (e.g., translation and rotation), we use the aligned faces provided from EmotiW2015 (pA) as well as our aligned faces (oA) in training deep CNNs. In addition to these deformation effects, using pA could provide complementary information when faces are erroneously aligned by our face registration. Examples of input preprocessing methods are shown in Fig. 3a.

5.1.2 Network architectures of deep CNNs

As the baseline architecture, we refer to Tang's deep CNN [52], the winning model of ICMLW2013's facial expression recognition challenge [16]. It consists of one input-transform and three convolution+pooling stages, followed by fully-connected hidden and output layers. In the input-transform stage for image mirroring and translating, data are augmented by extracting 42×42 patches from the 48×48 faces. The subsequent layers correspond to a configuration of {**CNN***M* - **FC***3072*} in our notation (see Table 1), except for average-pooling in the second and third stages of Tang's model. For more specific settings, see Tang's implementation [51].

On the basis of Tang's architecture, we design diverse deep CNNs by changing the sizes of filters for various receptive fields and by changing the number of neurons in a fully-connected hidden layer as denoted in Table 1. The **CNN***M* has a medium-size receptive field (with $5 \times 5, 4 \times 4$, and 5×5 filters for each convolutional (conv) layer, respectively), the **CNN***L* has a relatively large receptive field (with 7×7 filters for all conv layers), and the **CNN***S* has a relatively small

Table 1 Configuration of deep CNNs

Layer[1]	**CNN***S*		**CNN***M*		**CNN***L*	
	maps[2]	kernel[3]	maps	kernel	maps	kernel
input	42 × 42 @1	–	42 × 42 @1	–	42 × 42 @1	–
conv 1	42 × 42 @32	**3** × **3**, (1, 1)	42 × 42 @32	**5** × **5**, (1, 2)	42 × 42 @32	**7** × **7**, (1, 3)
max-pool 1	21 × 21 @32	2 × 2, (2, 0)	21 × 21 @32	3 × 3, (2, 1*)	21 × 21 @32	2 × 2, (2, 0)
conv 2	19 × 19 @32	**3** × **3**, (1, 0)	20 × 20 @32	**4** × **4**, (1, 1)	19 × 19 @32	**7** × **7**, (1, 2)
max-pool 2	10 × 10 @32	2 × 2, (2, 1*)	10 × 10 @32	3 × 3, (2, 1*)	10 × 10 @32	2 × 2, (2, 1*)
conv 3	10 × 10 @64	**3** × **3**, (1, 1)	10 × 10 @64	**5** × **5**, (1, 2)	10 × 10 @64	**7** × **7**, (1, 3)
max-pool 3	5 × 5 @64	2 × 2, (2, 0)	5 × 5 @64	3 × 3, (2, 1*)	5 × 5 @64	2 × 2, (2, 0)
fc hidden	**FC***512*: 512 neurons with a dropout probability of 0.5,					
	FC*1024*: 1024 neurons with a dropout probability of 0.5,					
	FC*2048*: 2048 neurons with a dropout probability of 0.5, or					
	FC*3072*: 3072 neurons with a dropout probability of 0.8					
fc output	7 neurons (one per class)					

[1]conv, max-pool, and fc: convolutional layer, max-pooling layer, and fully-connected layer, respectively.
[2]maps: *the size @the number of output maps*.
[3]kernel: *the size of kernels* (*stride*, *pad*) where "stride" refers to kernel spacing size, "pad without an asterisk (*)" refers to zero-padding to all four spatial directions (top, bottom, left and right directions) of input maps, and "pad with an asterisk" refers to zero-padding to the top and left

one (with 3 × 3 filters for all conv layers). For all **CNN** types, the strides and pads are properly set to ensure the same sizes of output maps (5 × 5 @64) in the max-pool 3 layer. Moreover, for each **CNN** type, four kinds of fully-connected hidden layer (**FC**) are used. In this **FC** layer, the dropout [46] is applied to reduce over-fitting in training deep models. Notice that from Tang's model, we modify the pooling layers in the second and third stages from average-pooling to max-pooling, because it provides better classification results in our preliminary experiments. Figure 3b illustrates our deep CNN architecture. For the nonlinearity, the rectified linear unit activation is applied to all conv and penultimate layers, and the softmax activation is applied to the output layer.

5.1.3 Usage of external data in training deep CNNs

The size of the SFEW2.0 data to train deep CNNs is quite small. Inspired by [23], we also decide to use two external databases along with the SFEW2.0 data for training models: FER-2013 and TFD. Then, we explore how to use them together with the SFEW data for training models. The following three strategies are considered:

i. Random initialization ⇒ In learning, using data as follows:
   {FER-2013 DB + TFD} for training
   {*SFEW Train* + *SFEW Valid*} for validation
ii. Random initialization ⇒ In learning, using data as follows:
   {FER-2013 DB + TFD + *SFEW Train*} for training
   {*SFEW Valid*} for validation
iii. Initialization from a pre-trained model constructed by using
   {FER-2013 DB Train + TFD} for training
   {FER-2013 DB Test} for validation
   ⇒ In learning, using data as follows:
   {*SFEW Train*} for training
   {*SFEW Valid*} for validation

In [23], strategies i and ii were discussed to train a deep CNN, which yielded per-frame predictions for video-based emotion recognition. In the experiment of [23], the strategy i was finally selected on the basis of its better validation performance. In addition to strategies i and ii, we also investigate one type of transfer learning scheme as denoted in strategy iii.

To determine the most proper usage of external data for the EmotiW2015 challenge, we evaluate the performance of several models trained differently with the aforementioned three strategies. Specifically, the 12 deep CNNs of $\mathbf{PREP}_{raw,oA}-\{\mathbf{CNN}\beta-\mathbf{FC}\gamma\}_{\mathbf{R}1}$ for $\forall\beta, \forall\gamma$ having various architectures are used. For all examined network architectures, strategy iii always shows higher validation performance than the other two strategies (see our previous work [26] for detailed results of learning curves and validation accuracies). This indicates that the transfer learning scheme is effective because it represents more similar and suitable feature distributions between training and validation data. Therefore, we eventually decide to use strategy iii. First, 108 deep CNNs are pre-trained using two external databases (FER-2013 and TFD). Then, these deep models are fine-tuned using the SFEW2.0 data, resulting in the final 216 deep CNNs: 108 models fine-tuned using oA, plus 108 using pA.

Note that on the last two submissions of test labels for the EmotiW2015 challenge, in addition to the 216 deep CNNs trained with strategy iii, we also incorporate the 24 models of strategies i and ii to form a committee. The reason behind adding these 24 models is as follows. Since these models are

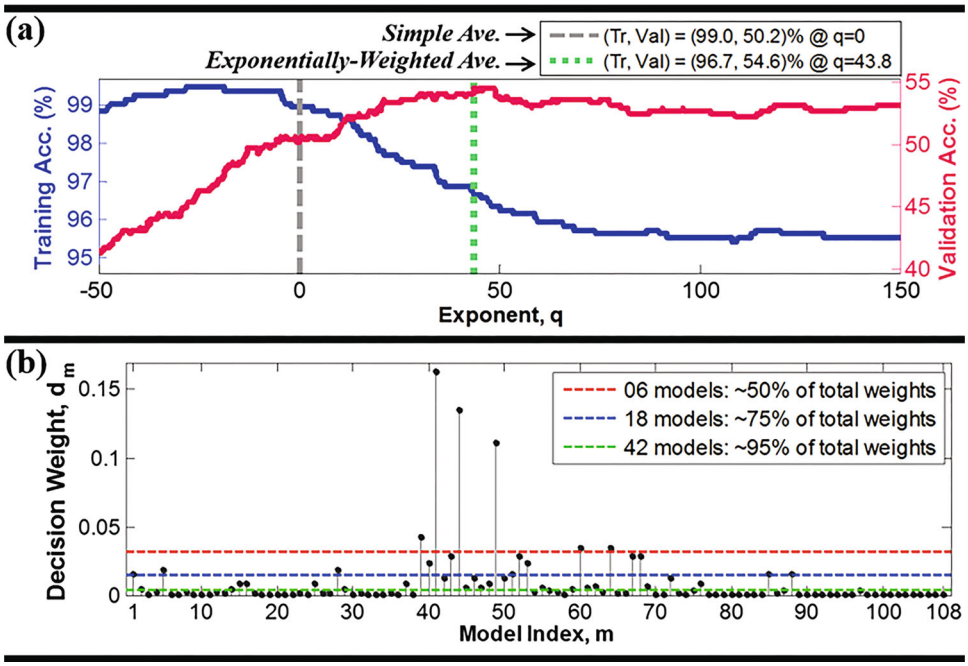

Fig. 4 **a** Training and validation accuracies (%) during the scanning procedure of an exponent q in the exponentially-weighted decision fusion and **b** corresponding decision weights for the selected q

trained with different strategies (more specifically, different composition of training data), their local minima in optimization of the training objective could differ greatly from those of the 216 models trained with transfer learning. We expect that these different local minima could result in different generalizations on input data, and thus be beneficial to handling of various facial expressions.

For training deep CNNs, we use the MatConvNet toolbox [55] on NVIDIA GeForce GTX 690 GPUs. Each deep CNN is trained using the stochastic gradient descent with a batch size of 200 and momentum of 0.9. Except for the last fully-connected layer with a weight decay of 0.002, a weight decay of 0.0001 is applied to all other layers. To avoid over-fitting, the data augmentation is conducted as follows. The training data are augmented by 10 times, by using five crops of size 42×42 (one from resizing an original 48×48 face and four from extracting its four corners) and their horizontal flopping. At the test phase, to maintain consistency with the training, 10 patches extracted from each face are fed to the model and the corresponding 10 predictions are averaged to produce a final prediction.

5.2 Hierarchical committee of deep CNNs

5.2.1 Exponentially-weighted decision fusion

Before moving on to examining the hierarchical committee architecture, our EWA rule is evaluated in conventional single-level committees, by comparing it to the widely-used decision fusion rules: the majority voting rule, the median rule, the simple average rule, and the SWA rule. We consider three types of single-level committees consisting of 108 deep CNNs trained using our aligned faces (oA), 108 models using provided alignments (pA), and 216 models using both alignments of oA and pA.

Here, the exponent q in our EWA rule is found by a simple uniform search: scanned over $[50 : 0.1 : 150]$ and selected to provide the maximum performance on validation data after the fusion. Figure 4a depicts the scanning procedure in the committee of 108 models using oA. We confirm that this searching method requires little additional computation, while it finds a proper q that can improve the generalization on validation data. Figure 4b shows the corresponding decision weights d_m for the selected q in Fig. 4a. We see that the decision weights of the EWA rule can reveal the importance or contribution of each individual on a final ensemble. The horizontal dotted line in Fig. 4b indicates the threshold above which a certain portion of the total weights is covered. The 6, 18, and 42 models are in charge of about 50, 75 and 95 % of the total, respectively. This information could be used for model selection and pruning.

5.2.2 Hierarchical committee

There are several ways to build hierarchical committees by considering "decisional" aspects (regarding the decision fusion rule for each hierarchical level) and "structural" aspects (regarding the number of hierarchical levels and the structure for recombination of higher-level decisions). In this section, we first present how to set the decision fusion rule for each level in our hierarchy, and then discuss how to design the hierarchical structure in our experiments.

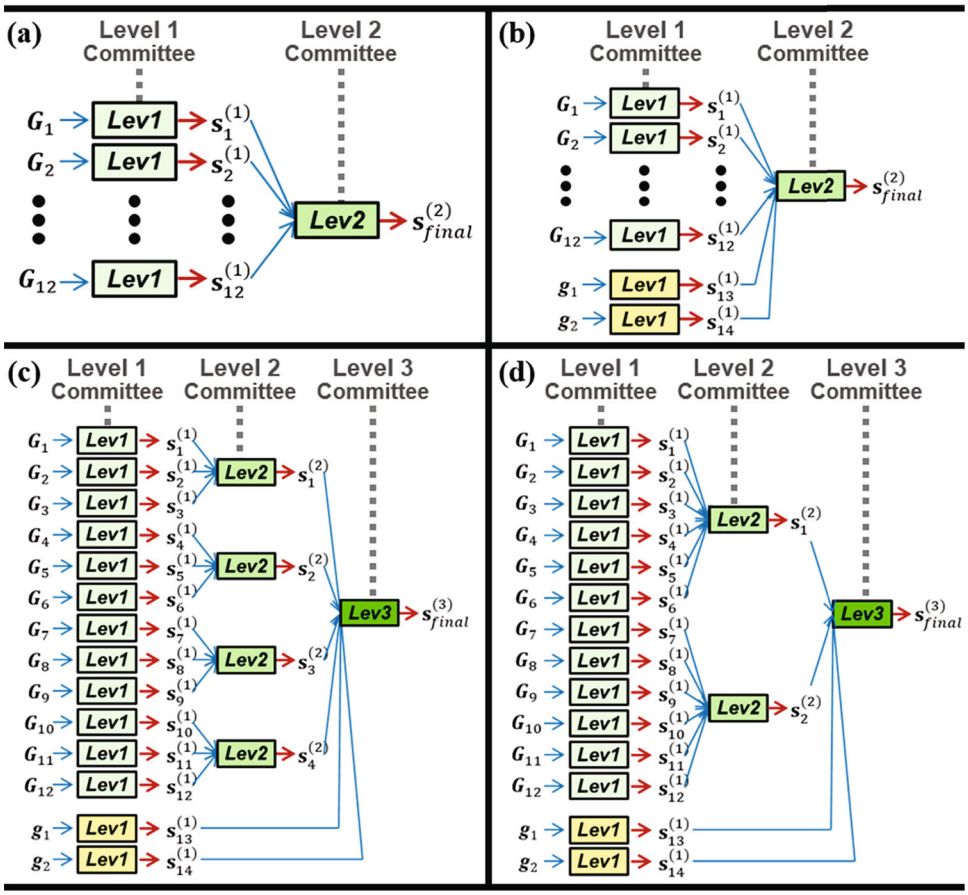

Fig. 5 Diagrams of two- and three-level hierarchical committees

At an earlier phase of the experiments for the EmotiW challenge, we submitted the predicted test labels obtained from adopting the EWA rule for all structural levels in the hierarchies. From these submissions, we observed some unexpected results; validation accuracies were higher than our previous submissions, but testing accuracies dropped. However, we learned an empirical lesson: applying the EWA rule for all levels of a hierarchy may have a negative impact on generalization for test data, because the exponent selection was optimized for the validation performance. Therefore, we decide to use the EWA rule only for the first level. For decision fusions in higher levels, the majority voting and simple average rules are used for a better generalization.

For the first structural level in our hierarchy, we organize 216 deep CNNs (obtained with the training strategy iii) into 12 sub-groups, having some overlapping members:

$\boldsymbol{G}_1$: **PREP***raw,oA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_2$: **PREP***iNor,oA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_3$: **PREP***cEnh,oA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_4$: **PREP***raw,pA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_5$: **PREP***iNor,pA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_6$: **PREP***cEnh,pA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\beta, \forall\gamma, \forall\delta$
$\boldsymbol{G}_7$: **PREP**α_1, *oA* $-$ $\{$**CNN***S* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$
$\boldsymbol{G}_8$: **PREP**α_1, *oA* $-$ $\{$**CNN***M* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$
$\boldsymbol{G}_9$: **PREP**α_1, *oA* $-$ $\{$**CNN***L* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$
$\boldsymbol{G}_{10}$: **PREP**α_1, *pA* $-$ $\{$**CNN***S* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$
$\boldsymbol{G}_{11}$: **PREP**α_1, *pA* $-$ $\{$**CNN***M* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$
$\boldsymbol{G}_{12}$: **PREP**α_1, *pA* $-$ $\{$**CNN***L* $-$ **FC**$\gamma\}_{\mathbf{R}\delta}$ for $\forall\alpha_1, \forall\gamma, \forall\delta$

Each sub-group consists of 36 deep CNNs. The three **per-PREP group**s ($\boldsymbol{G}_1$–$\boldsymbol{G}_3$) and three **per-CNN group**s ($\boldsymbol{G}_7$–$\boldsymbol{G}_9$) are formed from 108 models trained using our aligned faces (oA), and similarly the three **per-PREP** ($\boldsymbol{G}_4$–$\boldsymbol{G}_6$) and three **per-CNN** ($\boldsymbol{G}_{10}$–$\boldsymbol{G}_{12}$) **group**s are from 108 models using provided alignments (pA). By combining the decisions from these 12 groups, we first consider a simple two-level hierarchical structure as illustrated in Fig. 5a. With a fixed EWA rule of the first level, we vary decision fusion rules of the second level as the majority voting or simple average rules.

Moreover, as mentioned in Sect. 5.1.3, for handling more various face expressions and pursuing more diverse errors, we additionally examine 24 models from the training strategies i and ii. These models are also formed into the first-level groups, having 12 models each:

$\boldsymbol{g}_1$: **PREP***raw,oA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}I}$ for $\forall\beta, \forall\gamma$
$\boldsymbol{g}_2$: **PREP***raw,oA* $-$ $\{$**CNN**β $-$ **FC**$\gamma\}_{\mathbf{R}I}$ for $\forall\beta, \forall\gamma$

By adding these two groups' decisions to the first level, we build another two-level hierarchy that combines decisions from 14 groups, as shown in Fig. 5b. For clarity in the sub-

sequent discussion, we name each hierarchy as each index with a parenthesis in Fig. 5.

After analyzing classification performance of the two-level hierarchies (see the top part of Table 4), we decide to reduce the influence of decisions from $\boldsymbol{G}_1$–$\boldsymbol{G}_{12}$ on the final prediction by forming the three-level hierarchical committee as shown in Fig. 5c, d. Note that these three levels on the side of $\boldsymbol{G}_1$–$\boldsymbol{G}_{12}$ not only reduce the number of decisions [from 12 in hierarchy (b) to four in (c) or to two in (d)] but also make compact and reliable decisions passing through multiple levels. Here, two types of decision fusions in the second and third levels are considered, as demonstrated in the bottom part of Table 4.

5.3 Experimental results for the SFEW2.0 database

5.3.1 Classification performance of individual deep CNNs

Among the 216 deep CNNs trained with strategy iii (a transfer learning scheme), the best single model was $\mathbf{PREP}iNor,oA-\{\mathbf{CNN}L-\mathbf{FC}3072\}_{\mathbf{R}1}$, yielding the highest validation accuracy of 52.5 %. This model became the first submission for the EmotiW2015 challenge, achieving a test classification rate of 57.3 %. For validation accuracies of all individual deep CNNs, see our previous work [26].

We further analyzed general tendencies in the performances of deep CNNs. In the aspect of face alignments, the 108 models trained using oA yielded a mean validation accuracy of 47.0 %, while those using pA did 44.7 %. This indicates that, oA are more similarly aligned with each other compared to the provided alignments pA, leading to the superior mean accuracy with oA. In addition, to examine the trends according to preprocessing types and CNN architectures, we computed the mean accuracy of 36 models for each preprocessing (**per-PREP group**) and for each CNN (**per-CNN group**), as shown in Table 2. The illumination normalization was superior to other preprocessing types, and the **CNN***M* with the medium size of receptive field performed better than other CNN architectures.

5.3.2 Exponentially-weighted decision fusion

Classification accuracies of single-level committees using our EWA rule and widely-used decision fusion rules are shown in Table 3. Regardless of the committee types, our EWA rule outperformed all other rules. Moreover, as we expected, the SWA rule did not differ much from the simple average rule because individual models produced similar validation accuracies. Compared to the validation performance of the single best model, the three equal-weighting rules and the SWA rule that used the even distribution of weights yielded worse accuracies. However, interestingly, our EWA rule that eventually applied the sparse distribution of weights (as shown in Fig. 4b) achieved better validation accuracies. This supports that giving higher weights to few well-performing models could boost the ensemble performance.

Table 2 Mean and standard deviation of validation accuracies (%) of individual deep CNNs for each input normalization type and for each CNN receptive field size

Aligned faces from our method ($\alpha_2 = oA$)		Mean ± Std Acc. of 36 Deep CNNs
per-PREP group	$\boldsymbol{G}_1$ ($\alpha_1 = raw$)	46.9 ± 1.7
	$\boldsymbol{G}_2$ ($\alpha_1 = iNor$)	49.2 ± 1.6
	$\boldsymbol{G}_3$ ($\alpha_1 = cEnh$)	45.1 ± 2.4
per-CNN group	$\boldsymbol{G}_7$ ($\beta = S$)	46.3 ± 1.9
	$\boldsymbol{G}_8$ ($\beta = M$)	48.4 ± 1.7
	$\boldsymbol{G}_9$ ($\beta = L$)	46.5 ± 3.2
Aligned faces provided from EmotiW2015 ($\alpha_2 = pA$)		Mean ± Std Acc. of 36 Deep CNNs
per-PREP group	$\boldsymbol{G}_4$ ($\alpha_1 = raw$)	44.5 ± 2.0
	$\boldsymbol{G}_5$ ($\alpha_1 = iNor$)	46.1 ± 1.9
	$\boldsymbol{G}_6$ ($\alpha_1 = cEnh$)	43.5 ± 2.4
per-CNN group	$\boldsymbol{G}_{10}$ ($\beta = S$)	43.3 ± 2.0
	$\boldsymbol{G}_{11}$ ($\beta = M$)	46.3 ± 1.8
	$\boldsymbol{G}_{12}$ ($\beta = L$)	44.5 ± 2.2

Meanwhile, it is worth noting that even though the committee of 216 models with the EWA provided the highest classification rate on the validation data, its test rate was not the best. The best test accuracy of 60.5 % was achieved by the committee of 108 models using oA. This may imply that, at some level yielding a maximum validation performance, adding other models to the committee did not work properly with the EWA. Rather, it seemed to harm the generalization on test data because too many models participated to improve the final validation accuracy of the committee.

5.3.3 Hierarchical committee

The top part of Table 4 denotes classification performance of the two-level hierarchies. Notice that each hierarchy corresponds to each index with a parenthesis in Fig. 5. For both (a) and (b), the majority voting in the second level performed better than the simple average rule. However, the hierarchy (b) with a high validation accuracy of 56.2 % did not yield a better test accuracy compared to previous submissions. We suggested the following reason: The two groups, $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, added to the first level were expected to produce more diverse decisions on the basis of different training strategies, but their impacts on the second level were quite small because of competition with the 12 decisions from $\boldsymbol{G}_1$-$\boldsymbol{G}_{12}$.

Table 3 Validation (and testing, if available) accuracy (%) of single-level committees with various decision fusion rules

Decision fusion rule	Single-level committee		
	108 deep CNNs from oA	108 deep CNNs from pA	216 deep CNNs from both
Majority vote	51.6	48.2	50.7
Median	50.7	47.3	49.8
Simple ave.	50.2 (58.3)	47.8	49.8
SWA	50.7	47.8	49.8
EWA	**54.6 (60.5)**	52.2 (56.7)	**56.4 (60.0)**

Table 4 Validation (and testing, if available) accuracy (%) of hierarchical committees when using the EWA rule as the first-level decision fusion

Two-level hierarchical committee	Decision fusion in the second level	
	Simple ave.	Majority vote
(a)	53.4	56.2
(b)	53.9	56.2 (60.2)
Three-level hierarchical committee	Decision fusion in the second and third level	
	Simple ave. and majority vote	Majority vote and majority vote
(c)	**53.9 (61.6)**	56.2
(d)	52.5	**52.8 (61.6)**

The bottom part of Table 4 denotes the accuracies obtained from the three-level hierarchies. For both (c) and (d), applying majority voting to both the second and third levels showed better validation accuracies than using the simple average rule for the second level with the majority voting for the third level. More importantly, in the test performance, these two types of three-level hierarchies did not differ from each other, but were superior to the two-level hierarchies. This indicates that, when forming a hierarchical committee with higher levels, the structural consideration is more important than the decision fusion method to provide sufficient diversity in decisions.

5.3.4 Summary of our submissions for the EmotiW2015 challenge

In Table 5, we summarized classification results on the SFEW2.0 database, used for our submissions to the EmotiW2015 challenge. The test accuracy of 61.6 % obtained from three-level hierarchical committees of deep CNNs was the highest performance among our submissions. Note that we submitted ten sets of predicted test labels, but only eight accuracies were shown in this paper. The other two accuracies, which were not reported in this paper, used the following strategies: training support vector machines (SVMs) by using hidden activations of fully-connected layers in deep CNNs, and incorporating the class-probabilities from these SVMs with those from the deep CNNs to form the committees. These strategies using SVMs together with deep CNNs yielded the test accuracies of 58.6 and 60.0 % for a single-level committee and a hierarchical committee, respectively. Because these two did not improve classification performance and we'd like to clarify our core method, we presented eight recognition rates in detail.

The confusion matrices of testing performance for the best single model, the single-level committees with the simple average rule and with the EWA rule, and the three-level hierarchical committee are shown in Fig. 6. The correct recognitions of "anger", "sad" and "neutral" classes improved the overall accuracy in the three-level hierarchy. However, the "disgust" class was not correctly classified at all in the single-level and hierarchical committees. This may be because our models were more trained on other six classes than the "disgust" class, regarding the number of per-class samples. Hence, learning algorithms to handle the class-imbalance problems in training deep models could be used to boost performance of the "disgust" class having a small number of samples.

Notice that we suffered from over-fitting to validation data for some submissions, i.e., higher validation accuracies but lower testing accuracies. Eventually, for the EmotiW2015 challenge, we designed the hierarchical committees mainly based on intuition to give a better generalization for testing data. Our intuition towards the diversity of decisions in three-level hierarchical committees was right, resulting in the improvement of test performance. However, this process to

Table 5 Summary of classification accuracies (%) on the SFEW2.0 database

Description		Train	Valid	Test
–	Baseline [12]: {PHOG-LPQ} + nonlinear SVM	–	35.9	39.1
1	The best single deep CNN	97.1	52.5	57.3
2	A single-level committee with the simple average rule, 108 deep CNNs trained using oA	99.0	50.2	58.3
3	A single-level committee with the EWA rule, 108 deep CNNs trained using oA	96.7	54.6	60.5
4	A single-level committee with the EWA rule, 108 deep CNNs trained using pA	99.3	52.2	56.7
5	A single-level committee with the EWA rule, 216 deep CNNs trained using oA+pA	97.6	56.4	60.0
6	A two-level hierarchical committee, 240 deep CNNs trained using oA+pA, the architecture of Fig 5. (b) with the (EWA, majority vote) rules in the (first, second) level	98.0	56.2	60.2
7	A three-level hierarchical committee, 240 deep CNNs trained using oA+pA, the architecture of Fig 5. (c) with the (EWA, simple ave., majority vote) rules in the (first, second, third) levels	97.3	53.9	**61.6**
8	A three-level hierarchical committee, 240 deep CNNs using oA+pA, the architecture of Fig 5. (d) with the (EWA, majority vote, majority vote) rules in the (first, second, third) levels	96.6	52.8	**61.6**

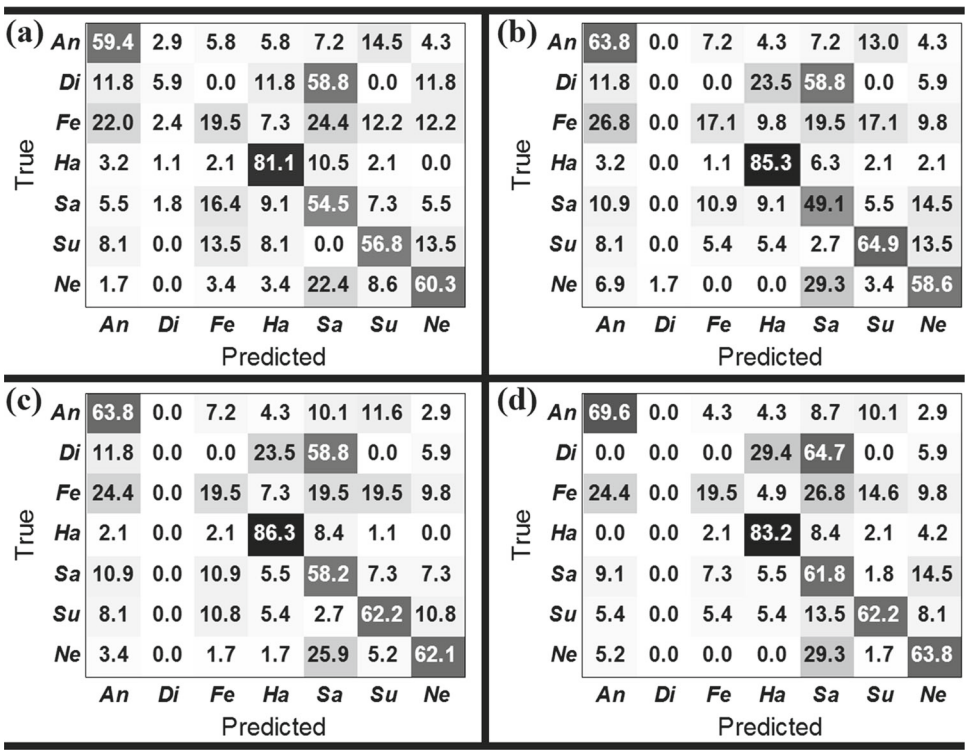

Fig. 6 Confusion matrices of testing accuracies (%) for the SFEW2.0 database. **a** the best single deep CNN of submission No. 1 in Table 5, **b** a single-level committee with the simple average rule of No. 2, **c** a single-level committee with the EWA rule of No. 3, and **d** a three-level hierarchical committee of No. 7. The terms of Au, Di, Fe, Ha, Sa, Su, and Ne are short for angry, disgust, fear, happy, sad, surprise, and neutral, respectively

determine the structure of hierarchical committees should be studied in both academic and engineering manners in our future works.

6 Experiments on other FER databases

We also test our method on other public FER databases (the FER-2013, the TFD, and the GENKI-4K) to confirm its generalization capability. We first describe a simplified experimental setting used for this section, and then present classification performance for each database.

6.1 Designing and training deep CNNs

The deep models in the experiment on the SFEW2.0 database cannot be used here, because testing data of the FER-2013 and the TFD (as well as training data of these two databases) are used in the training session of Sect. 5.1.3. Hence, we newly train deep CNNs as follows. First, we divide the 28,698 training faces of the FER-2013 database into the 25,110 samples (about seven-eighths of the whole training set) for training deep CNNs and the 3,588 samples (the remaining one-eighth) for validation. The validation data are used to

find the stopping epochs in learning deep models and to apply the EWA rule. Then, with this data division of the FER-2013 database, we train the following 36 deep CNNs which have the medium size of CNN receptive field.

$$\mathbf{PREP}\alpha - \{\mathbf{CNN}M - \mathbf{FC}\gamma\}_{\mathbf{R}\delta} \quad (3)$$

PREPα: input preprocessing methods,

$$\alpha = \begin{cases} raw & \text{Raw} \\ iNor & \text{Illumination Normalization} \\ cEnh & \text{Contrast Enhancement} \end{cases}$$

FCγ: number of neurons in a fully-connected hidden layer,
$\gamma = \{512, 1024, 2048, 3072\}$
Rδ: random seed number for weight initialization,
$\delta = \{1, 2, 3\}$

For the TFD and the GENKI-4K database, we adopt the "transfer learning" scheme as used in the SFEW2.0 database, i.e., the above 36 deep CNNs pre-trained using the FER-2013 database are set as initialization, and fine-tuned using each of the TFD and the GENKI-4K. Notice that, unlike the seven output neurons (corresponding to class labels) in deep CNNs for the SFEW2.0, the FER-2013, and the TFD, there are only two output neurons (smiling vs. non-smiling) for the GENKI-4K. Therefore, before fine-tuning using the GENKI-4K database, the weights between the fully-connected hidden layer and the output layer having two output neurons are initialized as random numbers, while other weights from the input layer to the fully-connected hidden layer are initialized as the pre-trained weights from the FER-2013.

6.2 Hierarchical committee of deep CNNs

For the first structural level in our hierarchy, we organize 36 deep CNNs into 7 sub-groups, having some overlapping members:

G_1: $\mathbf{PREP}raw - \{\mathbf{CNN}M - \mathbf{FC}\gamma\}_{\mathbf{R}\delta}$ for $\forall\gamma, \forall\delta$
G_2: $\mathbf{PREP}iNor - \{\mathbf{CNN}M - \mathbf{FC}\gamma\}_{\mathbf{R}\delta}$ for $\forall\gamma, \forall\delta$
G_3: $\mathbf{PREP}cEnh - \{\mathbf{CNN}M - \mathbf{FC}\gamma\}_{\mathbf{R}\delta}$ for $\forall\gamma, \forall\delta$
G_4: $\mathbf{PREP}\alpha - \{\mathbf{CNN}M - \mathbf{FC}512\}_{\mathbf{R}\delta}$ for $\forall\alpha, \forall\delta$
G_5: $\mathbf{PREP}\alpha - \{\mathbf{CNN}M - \mathbf{FC}1024\}_{\mathbf{R}\delta}$ for $\forall\alpha, \forall\delta$
G_6: $\mathbf{PREP}\alpha - \{\mathbf{CNN}M - \mathbf{FC}2048\}_{\mathbf{R}\delta}$ for $\forall\alpha, \forall\delta$
G_7: $\mathbf{PREP}\alpha - \{\mathbf{CNN}M - \mathbf{FC}3072\}_{\mathbf{R}\delta}$ for $\forall\alpha, \forall\delta$

Each of three **per-PREP group**s (G_1-G_3) consists of 12 deep CNNs, while each of four **per-FC group**s (G_4-G_7) contains 9 deep CNNs. By combining the decisions from these 7 groups, we form two-level hierarchical committees. Similar to the experiment on the SFEW2.0 database, we fix the decision fusion rule in the first level as our EWA rule, and vary the rule in the second level as the majority voting, the median rule, or the simple average rule.

Table 6 Test accuracies (%) of individual deep CNNs for the FER-2013 database

FER-2013		**CNN***M*			
		FC*512*	**FC***1024*	**FC***2048*	**FC***3072*
PREP *raw*	**R***1*	**69.69**	**70.41**	**69.91**	**70.58***
	R*2*	68.99	69.32	69.16	69.88
	R*3*	67.85	68.29	68.90	69.63
PREP *iNor*	**R***1*	67.82	69.16	69.21	68.32
	R*2*	68.77	68.71	68.82	67.73
	R*3*	68.93	69.04	68.74	68.46
PREP *cEnh*	**R***1*	68.90	69.04	69.63	68.71
	R*2*	68.49	68.60	69.80	69.21
	R*3*	68.24	68.96	69.52	70.05

The highest accuracy for a given architecture (each column) is written in bold. The asterisk (*) denotes the single best model

6.3 Experimental results for other FER databases

6.3.1 The FER-2013 database

The private test set of the FER-2013 database was used for evaluation. The test accuracies of individual deep CNNs are shown in Table 6. The best single model was $\mathbf{PREP}raw - \{\mathbf{CNN}M - \mathbf{FC}3072\}_{\mathbf{R}1}$, yielding a test accuracy of 70.58 %. For two-level hierarchical committees, we fixed the first-level decision fusion rule as the EWA rule and changed the second-level rule as denoted in Table 7. All examined two-level hierarchies showed better accuracies than the best single deep CNN. Moreover, the simple average rule in the second level achieved the highest test performance of 72.72 %, outperforming state-of-the-art results of the FER-2013 database.

6.3.2 The TFD database

We strictly followed the official evaluation protocol of the TFD using 5-fold cross validation. For individual deep CNNs, the test accuracies averaged over 5 folds are reported in Table 8. The best single model was $\mathbf{PREP}raw - \{\mathbf{CNN}M - \mathbf{FC}3072\}_{\mathbf{R}1}$, yielding the mean accuracy of 86.72 %. Table 9 shows classification performance of two-level hierarchies with applying the EWA rule to the first level. Here, the median rule in the second level showed the best mean performance of 87.71 %, competing with state-of-the-art results for the TFD.

6.3.3 The GENKI-4K database

We conducted 4-fold cross validation for evaluation of the GENKI-4K database as used in [24,33,43]. For individual models, the test recognition rates averaged over 4 folds are shown in Table 10. The deep CNN of $\mathbf{PREP}raw - \{\mathbf{CNN}M -$

Table 7 Test performance comparison on the FER-2013. Our two-level hierarchical committees use the EWA rule for the first-level decision fusion

Our two-level hierarchical committee	Decision fusion in the second level	Acc. (%)
	Majority voting rule	72.28
	Median Rule	72.39
	Simple average rule	**72.72**
The winning method of [16]: a deep CNN trained using the L2-SVM loss function [52]		71.16
The first runner-up of [16]		69.27
The second runner-up of [16]		68.82
Multiple kernel learning based on SIFT descriptors [20]		67.48

Table 8 Mean and standard deviation of test accuracies (%) of individual deep CNNs from 5-fold cross validation using the TFD database

TFD		**CNN***M*			
		FC*512*	**FC***1024*	**FC***2048*	**FC***3072*
PREP *raw*	**R***1*	84.33 ± 2.03	84.50 ± 1.69	85.29 ± 2.00	**86.72*** ± 2.11
	R*2*	**86.22** ± 1.89	**86.00** ± 2.40	85.06 ± 1.70	85.50 ± 2.20
	R*3*	85.72 ± 1.77	85.03 ± 1.84	**85.82** ± 1.80	85.91 ± 1.96
PREP *iNor*	**R***1*	84.91 ± 1.62	85.34 ± 1.45	85.28 ± 1.35	85.89 ± 1.53
	R*2*	85.68 ± 0.69	85.09 ± 1.49	85.28 ± 1.29	85.41 ± 1.56
	R*3*	85.49 ± 1.34	85.22 ± 1.28	85.25 ± 1.36	85.76 ± 1.40
PREP *cEnh*	**R***1*	85.39 ± 2.53	85.13 ± 1.65	84.98 ± 1.76	85.82 ± 1.53
	R*2*	84.99 ± 2.19	85.41 ± 2.06	84.89 ± 1.60	85.70 ± 2.05
	R*3*	84.91 ± 1.57	85.32 ± 1.84	85.08 ± 1.78	85.72 ± 2.48

The highest mean accuracy for a given architecture (each column) is written in bold. The asterisk (*) denotes the single best model

Table 9 Test performance comparison on the TFD database

Our Two-Level Hierarchical Committee	Decision Fusion in the Second Level	Acc. (%)
	Majority voting rule	87.58 ± 2.36
	Median rule	**87.71** ± 2.21
	Simple average rule	87.63 ± 2.53
Zero-bias deep CNN with dropout and data augmentation [25]		**89.8** ± 1.8
Spatial-pyramid-pooled OMP-1 features + A deep net trained on noisy labels with bootstrapping [38]		86.8
Disentangling Boltzmann machine (disBM) [39]		85.43 ± 2.54
Contractive convolutional network (CCNET) + Contractive discriminative analysis (CDA) + SVM [40]		85.0 ± 0.47

Our two-level hierarchical committees use the EWA rule for the first-level decision fusion

FC*2048*$\}_{\mathbf{R}3}$ achieved the highest mean accuracy of 95.25 %. In Table 11, we showed the test accuracies of two-level hierarchical committees. Although the hierarchy with using the simple-average rule in the second level performed better than the best single model, the performance difference between them was not significant. It may imply that the classification rate of the single deep CNN was already saturated near a maximum accuracy. Nevertheless, our results outperformed state-of-the-art accuracies for the GENKI-4K database. We believed that superior accuracies of our methods in the GENKI-4K came from the transfer learning scheme in training deep models as well as the hierarchical committee.

Table 10 Mean and standard deviation of test accuracies (%) of individual deep CNNs from 4-fold cross validation using the GENKI-4K database

GENKI-4K		**CNN***M*			
		FC*512*	**FC***1024*	**FC***2048*	**FC***3072*
PREP *raw*	**R***1*	94.95 ± 0.75	**95.10** ± 0.59	95.00 ± 0.82	**95.03** ± 0.52
	R*2*	94.88 ± 1.06	94.80 ± 0.78	94.75 ± 0.71	94.72 ± 0.71
	R*3*	94.85 ± 0.78	95.00 ± 0.67	**95.25*** ± 0.76	94.73 ± 0.73
PREP *iNor*	**R***1*	94.65 ± 0.81	94.48 ± 0.41	94.88 ± 0.59	94.55 ± 1.22
	R*2*	94.68 ± 1.28	94.95 ± 0.72	94.38 ± 1.18	94.60 ± 0.45
	R*3*	94.73 ± 0.74	94.45 ± 0.60	94.50 ± 0.66	94.65 ± 0.69
PREP *cEnh*	**R***1*	**94.97** ± 0.97	94.35 ± 0.78	94.97 ± 0.67	94.48 ± 0.48
	R*2*	94.30 ± 0.42	94.58 ± 0.89	94.58 ± 0.65	94.67 ± 1.04
	R*3*	94.28 ± 0.77	95.07 ± 0.72	94.60 ± 0.68	94.95 ± 0.77

The highest mean accuracy for a given architecture (each column) is written in bold. The asterisk (*) denotes the single best model

Table 11 Test performance comparison on the GENKI-4K database. Our two-level hierarchical committees use the EWA rule for the first-level decision fusion

Our Two-Level Hierarchical Committee	Decision Fusion in the Second Level	Acc. (%)
	Majority voting rule	95.35 ± 0.86
	Median rule	95.35 ± 0.86
	Simple average rule	**95.38** ± 0.88
LBP-like descriptors + Local FDA + SVM [24]		93.2 ± 0.92
HOG descriptors + Semi-supervised NMF + SVM [33]		92.26 ± 0.81
A deep CNN at 48 × 48 pixel resolution [24]		91.5
Feature based on pixel comparison + Adaboost classifier [43]		89.70 ± 0.45

7 Conclusion

In this paper, we present a framework based on committee machines of deep CNNs and its application to robust FER. To generate diverse errors for a better committee, we first constructed multiple deep CNNs as individual committee members. Here, deep models were trained by applying various network architectures, several strategies to use external data, and different input preprocessing and random initialization. With these individuals, we formed hierarchical committees that adopted the valid-accuracy-based exponentially-weighted average rule. This exponentially weighted decision fusion was superior to other commonly used ensemble methods by increasing the generalization capability. Furthermore, the hierarchical structure indeed made more reliable decisions with the consensus of various sub-groups.

Our proposed approach was demonstrated on the SFEW2.0 competition data released for the EmotiW2015 challenge. To sum up our submissions, the test accuracy of the best single deep CNN was 57.3 %, whereas the single-level committees of 108 models trained using our aligned faces yielded 58.3 % with the simple average rule and 60.5 % with the exponentially weighted decision fusion. Furthermore, based on three-level hierarchical committees of total 240 deep CNNs, we achieved 61.6 % test accuracy, greatly outperforming the SFEW2.0 baseline of 39.1 %. On other public FER databases, we also obtain impressive performance that are similar to or higher than state-of-the-art results for these databases. We believe that the superiority of our committee machines could further be drawn in other pattern recognition problems as well as robust FER.

In our future works, we will design various objective functions in training individual deep CNNs in order to obtain more diverse decisions. Moreover, methods of determining the structure of hierarchical committees will be intensively studied in both academic and engineering manners.

Acknowledgments This work was supported by the Industrial Strategic Technology Development Program (10044009, Development of a self-improving bidirectional sustainable HRI technology for 95 % of successful responses with understanding users complex emotion and transactional intent through continuous interactions) funded by the Ministry of Knowledge Economy (MKE, Korea).

References

1. Agostinelli F, Anderson MR, Lee H (2013) Adaptive multi-column deep neural networks with application to robust image denoising. In: Advances in Neural Information Processing Systems, pp 1493–1501

2. Aksela M, Laaksonen J (2006) Using diversity of errors for selecting members of a committee classifier. Patt Recog 39(4):608–623
3. Bell D, JwW Guan, Bi Y et al (2005) On combining classifier mass functions for text categorization. Know Data Eng IEEE Trans 17(10):1307–1319
4. Boulesteix AL, Porzelius C, Daumer M (2008) Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics 24(15):1698–1706
5. Cireşan D, Meier U, Masci J, Schmidhuber J (2012a) Multi-column deep neural network for traffic sign classification. Neural Networks 32:333–338
6. Cireşan D, Meier U, Schmidhuber J (2012b) Multi-column deep neural networks for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, pp 3642–3649
7. Cireşan DC, Meier U, Gambardella LM, Schmidhuber J (2010) Deep, big, simple neural nets for handwritten digit recognition. Neural Comput 22(12):3207–3220
8. Cireşan DC, Meier U, Gambardella LM, Schmidhuber J (2011) Convolutional neural network committees for handwritten character classification. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on, IEEE, pp 1135–1139
9. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. MultiMedia IEEE 19(3):34–41
10. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2013) Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, pp 509–516
11. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, pp 461–466
12. Dhall A, Murthy OVR, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp 423–426
13. Dietterich TG (2000) Ensemble methods in machine learning. In: Multiple classifier systems, Springer, pp 1–15
14. Ebrahimi Kahou S, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp 467–474
15. Giacinto G, Roli F (2001) Design of effective neural network ensembles for image classification purposes. Image Vision Comput 19(9):699–707
16. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH et al (2015) Challenges in representation learning: A report on three machine learning contests. Neural Networks 64:59–63
17. Gross R, Brajovic V (2003) An image preprocessing algorithm for illumination invariant face recognition. In: Audio-and Video-Based Biometric Person Authentication, Springer, pp 10–18
18. Hansen LK, Salamon P (1990) Neural network ensembles. Patt Anal Mach Intell IEEE Trans 12(10):993–1001
19. Huang Y, Suen C (1993) The behavior-knowledge space method for combination of multiple classifiers. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp 347–347
20. Ionescu RT, Popescu M, Grozea C (2013) Local learning to improve bag of visual words model for facial expression recognition. In: Workshop on Challenges in Representation Learning, ICML
21. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3(1):79–87
22. Jordan MI, Jacobs RA (1994) Hierarchical mixtures of experts and the em algorithm. Neural Comput 6(2):181–214
23. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülçehre Ç, Memisevic R, Vincent P, Courville A, Bengio Y, Ferrari RC, et al. (2013) Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, pp 543–550
24. Kahou SE, Froumenty P, Pal C (2014) Facial expression analysis based on high dimensional binary features. In: Computer Vision-ECCV 2014 Workshops, Springer, pp 135–147
25. Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition? arXiv preprint arXiv:1510.02969
26. Kim BK, Lee H, Roh J, Lee SY (2015) Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp 427–434
27. Kittler J, Hatef M, Duin RP, Matas J (1998) On combining classifiers. Patt Anal Mach Intell IEEE Trans 20(3):226–239
28. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
29. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, USA
30. Kuncheva LI, Bezdek JC, Duin RP (2001) Decision templates for multiple classifier fusion: an experimental comparison. Patt Recogn 34(2):299–314
31. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Procee IEEE 86(11):2278–2324
32. Liu M, Zhang D, Yap PT, Shen D (2012) Hierarchical ensemble of multi-level classifiers for diagnosis of alzheimer's disease. In: Machine Learning in Medical Imaging, Springer, pp 27–35
33. Liu M, Li S, Shan S, Chen X (2013) Enhancing expression recognition in the wild with unlabeled reference data. In: Computer Vision-ACCV 2012, Springer, pp 577–588
34. Liu M, Wang R, Li S, Shan S, Huang Z, Chen X (2014) Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, pp 494–501
35. Pajares G, Guijarro M, Ribeiro A (2010) A hopfield neural network for combining classifiers applied to textured images. Neural Networks 23(1):144–153
36. Pan SJ, Yang Q (2010) A survey on transfer learning. Knowl Data Eng IEEE Trans 22(10):1345–1359
37. Polikar R (2006) Ensemble based systems in decision making. Circ Syst Magaz IEEE 6(3):21–45
38. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A (2014a) Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596
39. Reed S, Sohn K, Zhang Y, Lee H (2014b) Learning to disentangle factors of variation with manifold interaction. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp 1431–1439
40. Rifai S, Bengio Y, Courville A, Vincent P, Mirza M (2012) Disentangling factors of variation for facial expression recognition. In: Computer Vision-ECCV 2012, Springer, pp 808–822
41. Rodríguez-Liñares L, García-Mateo C, Alba-Castro JL (2003) On combining classifiers for speaker authentication. Patt Recogn 36(2):347–359
42. Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M (2011) Avec 2011-the first international audio/visual emotion challenge. In: Affective Computing and Intelligent Interaction, Springer, pp 415–424
43. Shan C (2012) Smile detection by boosting pixel differences. Image Process IEEE Trans 21(1):431–436

44. Sharkey AJC (1996) On combining artificial neural nets. Conn Sci 8(3–4):299–314
45. Shipp CA, Kuncheva LI (2002) Relationships between combination methods and measures of diversity in combining classifiers. Inform Fusion 3(2):135–148
46. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
47. Štruc V, Pavešic N (2011) Photometric normalization techniques for illumination invariance. Advances in Face Image Analysis: Techniques and Technologies pp 279–300
48. Su Y, Shan S, Chen X, Gao W (2009) Hierarchical ensemble of global and local classifiers for face recognition. Image Process IEEE Trans 18(8):1885–1896
49. Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, pp 1891–1898
50. Susskind JM, Anderson AK, Hinton GE (2010) The toronto face database. Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech Rep
51. Tang Y (2013a) deep-learning-faces. https://code.google.com/p/deep-learning-faces/
52. Tang Y (2013b) Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239
53. Titsias MK, Likas A (2002) Mixture of experts classification using a hierarchical mixture model. Neural Comput 14(9):2221–2244
54. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, pp 921–926
55. Vedaldi A, Lenc K (2014) Matconvnet-convolutional neural networks for matlab. arXiv preprint arXiv:1412.4564
56. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
57. Whitehill J, Littlewort G, Fasel I, Bartlett M, Movellan J (2009) Toward practical smile detection. Patt Anal Mach Intell IEEE Trans 31(11):2106–2111
58. Wolpert DH (1992) Stacked generalization. Neural Networks 5(2):241–259
59. Wu CH, Liang WB (2011) Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. Affect Comp IEEE Trans 2(1):10–21
60. Wu D, Shao L (2014) Deep dynamic neural networks for gesture segmentation and recognition. In: Computer Vision-ECCV 2014 Workshops, Springer, pp 552–571
61. Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, pp 532–539
62. Yao A, Shao J, Ma N, Chen Y (2015) Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, pp 451–458
63. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM Int Confer Multi Inter ACM, pp 435–442
64. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, pp 2879–2886