Expression Classification in Children Using Mean Supervised Deep Boltzmann Machine

Shruti Nagpal[1], Maneet Singh[1], Mayank Vatsa[1], Richa Singh[1], Afzel Noore[2]
[1]IIIT Delhi, India, [2]Texas A&M University, Kingsville, USA
{shrutin, maneets, mayank, rsingh}@iiitd.ac.in, afzel.noore@tamuk.edu

Abstract

Automated facial expression classification has widespread application in multiple domains such as human computer interaction, health and entertainment, biometrics, and security. There are six basic facial expressions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise, apart from a neutral state. Most of the research in expression classification has focused on adult face images, with no dedicated research on automating expression classification for children. To the best of our knowledge, this is the first research which presents a deep learning based expression classification approach for children. A novel supervised deep learning formulation, termed as Mean Supervised Deep Boltzmann Machine (msDBM) is proposed which classifies an input face image into one of the seven expression classes. The proposed approach has been evaluated on two child face datasets - Radboud Faces and CAFE, along with experiments on the adult face images of the Radboud Faces dataset. Experimental results and analysis reinforces the challenging nature of the task at hand, and the effectiveness of the proposed msDBM model.

1. Introduction

Facial expressions are caused by the movement of facial muscles, and often convey the emotional state of a person. Based on the commonality in the change of facial muscles, there exist six forms of universally accepted emotions since 1992: *anger, disgust, fear, happiness, sadness,* and *surprise* [10]. Expressions are one of the earliest forms of communication, and recent literature has focused on automating the task of expression classification [2, 13, 29]. However, since each individual has a unique way of expression, the classification task suffers from the inherent challenge of low inter-class and large intra-class variations. Automated facial expression classification [5,31] has widespread application in different domains such as human computer interaction, affective computing, health, and entertainment. It can also be used for monitoring patients incapable of other modes of communication, and for building automated psychological profiles. In literature, it has been studied that the expressions of each individual are unique [8], and thus can be utilized as ancillary information for biometric recognition as well. Moreover, due to the increased intra-class variations caused by varying expressions, expression normalization has also aided in improving the performance of existing face recognition models [33,51].

Figure 1: Images depicting the variations observed for *Anger* and *Happy* in adults and children. All images are taken from the Internet.

While facial expression classification has garnered significant attention over the past few years, most of the research has focused on expression classification in adults with limited attention to child expression classification. Children are known to be more expressive in case of positive emotions, while projecting ambiguous expressions in case of negative emotions [15, 26]. Figure 1 presents sample face images for two expressions of *angry* and *happy*, for adults and children. Owing to the expressive nature of children, large variations can be observed between samples of the same class, thereby rendering the problem of automated expression classification further challenging in their case.

With the advancement in technology and its increased applicability in our routine life, there exist several scenarios

which demand an automated expression recognition system, particularly for children. Such a system has utility in day-care centres or crèches to alert the attendants or supervisors in case of an unusual expression. In case a child is hurt, in pain, or is hungry, the child will emote an expression such as sadness, which can be detected through an automated system. Similarly, e-learning systems for classrooms or distance learning platforms can also benefit from an automated expression recognition system, where expressions can be utilized to assess the understanding of students. Automated expression classification can also be used in multiple scenarios related to mental illness. Certain illnesses induce a difference in the pattern of emotions experienced by a person, which can be recognized via the expressions being presented by the individual. Major depressive disorder or seasonal affective disorder results in consistently feeling sad for an elongated period, while the bipolar disorder leads to constant sudden variation in emotions, which can be observed through an individual's expressions. We believe that an automated child expression classification algorithm can be used to detect such conditions at an early stage. Early detection could result in early intervention and assistance for children from a young age. Moreover, such a system could also be used by autistic kids to understand their peers and not get isolated due to their inability to assess expressions. Wide scale applicability of an automated facial expression recognition system in day to day life and medical conditions makes it a necessity in the coming times, thus demanding dedicated research attention.

1.1. Research Contributions

In this research, we address the unexplored and demanding task of automated facial expression classification for children. Inspired from the feature learning capabilities of deep learning models and their ability to encode hierarchical feature representations, the *Mean Supervised Deep Boltzmann Machine (msDBM)* is proposed. It incorporates supervision in the otherwise unsupervised DBM architecture, and models the inter-class and intra-class variations during the feature learning process. To the best of our knowledge, this is the first work where an automated pipeline is proposed for face expression recognition in children, with dedicated focus on modeling the inter and intra-class variations. Experiments are performed on two child expression datasets: Radboud Faces dataset [20] and Child Affective Facial Expression (CAFE) dataset [25]. Both the datasets contain child face images imitating the six basic expressions of *anger, disgust, fear, happy, sad*, and *surprise*, along with a *neutral* state. The experimental results and analysis showcase the challenging nature of the given problem and the effectiveness of the proposed model.

2. Automated Expression Classification in Adults

In literature, researchers have focused primarily on performing automated expression classification in adult face images. Ekman *et al.* [11] developed a Facial Action Coding System (FACS) to understand and define facial movements in images. FACS has been widely used by researchers in order to perform expression classification [14, 22, 40, 46, 49]. Subspace learning techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on whole images have also been used for expression recognition [3, 6, 36].

Gu *et al.* [12] proposed a method for developing person-independent expression classification, inspired by the human ability to perceive emotions. Gabor filters are applied on local patches to obtain features, followed by PCA and LDA. These features are then provided to independent local classifiers, outputs of which are concatenated to obtain a global feature. This is followed by PCA and LDA on the global feature for classification. A non-negative matrix factorization based supervised approach was proposed by Zhi *et al.* [50]. A sparseness constraint is introduced in non-negative matrix factorization. Further, the neighborhood of the samples is preserved by minimizing the graph preserving criterion. With the recent Emotion Recognition in the Wild challenge, many studies have focused on multi-modal expression classification where apart from the face image, researchers also utilize the voice component to predict the expression of the subject more accurately. Sikka *et al.* [37] proposed a multiple kernel learning based technique for fusing information of the two modalities. Kahou *et al.* [19] proposed EmoNets wherein face images are classified using a combination of Convolutional Neural Networks for feature extraction and a Support Vector Machine (SVM) for classification, while results on audio are predicted using Deep Belief Networks. The authors also explore relational autoencoders to learn spatio-temporal information and a shallow network trained on features around the mouth region.

More recently, deep learning based techniques have been explored for static face images [7, 18, 23, 29, 48]. Liu *et al.* [24] proposed a Boosted Deep Belief Network, which learns expression related discriminative features, which are selected to form a boosted and strong classifier. Gui *et al.* [13] devised a curriculum learning based deep learning framework to efficiently perform the given task. The authors create a curriculum from the training data, i.e. divide the data into *easy* and *hard*: the easier samples are used to train the model first, followed by the harder ones. Curriculum learning is used to fine-tune pre-trained CNN models. Ng *et al.* [30] presented a transfer learning based approach, where CNNs pre-trained on the ImageNet dataset [9] undergo a two-step supervised fine-tuning process. The first

Figure 2: Mean face images for seven expressions generated from the Radboud Faces dataset [20].

step of fine-tuning is performed specific to the task on an additional emotion recognition dataset, and the second stage fine-tuning is performed with the specific training set of the given dataset being used for evaluation. Yang *et al.* [45] proposed a de-expression learning procedure. A neutral face image is generated for the input image, and the *residue* deep learning features in the intermediate layer are used for expression classification.

Inspired by the capabilities of deep learning based models and addition of supervision in traditionally unsupervised models [38, 39, 44, 47], a novel supervised deep learning based framework is presented for the task of expression classification in children. The proposed model and its preliminaries are discussed in detail in the following section.

3. Proposed Mean Supervised Deep Boltzmann Machine

The proposed Mean Supervised Deep Boltzmann Machine (msDBM) incorporates supervision in the traditionally unsupervised Deep Boltzmann Machine (DBM) [34]. It models the inter-class and intra-class variations at the time of feature learning. The following subsection presents some preliminaries, followed by a detailed explanation of the proposed model.

3.1. Preliminaries

A DBM is a hierarchical model built using the fundamental unit of the Restricted Boltzmann Machine (RBM). RBMs are probabilistic bipartite graphs, consisting of a visible and a hidden layer. The visible layer (v) corresponds to the input data, while the hidden layer (h) corresponds to the learned representation. For binary input data, a visible layer of n dimension, and a hidden layer of m dimension, the energy function of a Restricted Boltzmann Machine is modeled as follows:

$$E(v,h) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i w_{i,j} h_j \quad (1)$$

where, a and b are the visible and hidden layer bias vectors, respectively. Using the energy function, a probability distribution is defined over the hidden and visible vectors as follows:

$$P(v,h) = \frac{1}{Z} e^{\ E(v,h)} \quad (2)$$

where, Z is a normalization constant, termed as the *partition function*, which corresponds to the sum of $e^{\ E(v,h)}$, over all possible configurations of the hidden and visible layers. Following this, the probability which a network assigns to a given visible vector is further calculated as:

$$P(v) = \frac{1}{Z} \sum_{h} e^{\ E(v,h)} \quad (3)$$

For n given vectors, a Restricted Boltzmann Machine thus optimizes the following loss function:

$$\ell_{RBM} = \sum_{i=1}^{n} P(v_i) \quad (4)$$

In case of real-valued data, Equation 1 is modified to create the Gaussian-Bernoulli RBM [17]. As can be observed, RBMs are unsupervised feature learning models, which do not utilize the class information at the time of feature learning. In literature, researchers have proposed supervised extensions to RBMs in order to learn classification oriented features. In 2008, Larochelle and Bengio [21] presented the Discriminative Restricted Boltzmann Machine (DRBM) which minimizes the joint probability of the input sample and the class label. DRBM is thus presented as a model capable of learning discriminative features, and performing classification as well. In 2017, Sankaran *et al.* [35] built upon the DRBM by proposing Class Sparsity Signature based Restricted Boltzmann Machine (cssRBM), which incorporates a $\ell_{2,1}$ norm based regularizer to the loss function of the DRBM. This enforces a signature onto features belonging to the same class, thereby reducing the intra-class variations.

3.2. Proposed msDBM

Existing techniques incorporate class information during feature learning in order to reduce the intra-class variations, or model the relationship between the input sample and the class label. In this research, we propose to incorporate supervision by modeling both inter-class, and intra-class variations during the feature learning process of a RBM. In terms of expression recognition, Figure 2 presents the mean face images of seven expressions, obtained from the Radboud Faces Dataset [20]. It can visually be observed that the mean images vary significantly from each other, and thus can be distinguished with ease. Motivated by these visual differences, this research incorporates these differences at the feature level in RBMs. In order to achieve this, the loss of a traditional RBM is modified to incorporate terms for minimizing the intra-class variations, and maximizing the inter-class variations in terms of the mean feature vectors. This is performed by utilizing the distance of the learned feature from the mean feature of a particular class. For a sample v^c, belonging to class c, the distance between its corresponding hidden representation, h^c, and the mean representation of class c is minimized as follows:

$$\ell = \ell_{RBM} + \lambda^c \|h^c - m^c\|_2^2 \quad (5)$$

where, λ^c is the intra-class regularization constant for class c, and m^c corresponds to the mean feature vector of class c, calculated as follows:

$$m^c = \mu(\mathbf{H}^c) \tag{6}$$

where, $\mathbf{H}^c$ corresponds to a matrix containing feature vectors of class c, and μ refers to the mean operator. In Equation 5, the additional term promotes the minimization of intra-class variations by forcing feature vectors of a particular class closer to the mean representation. However, it does not maximize the inter-class variations. In order to maximize the inter-class variations, the distance of the hidden representation from the mean representation of all other classes is also incorporated in the loss function. For a n class problem, this is performed as follows:

$$\ell_{msRBM} = \ell_{RBM} + \lambda^c \|h^c - \mu(\mathbf{H}^c)\|_2^2 - \nu^c \sum_{i=1, i \neq c}^{n} \|h^c - \mu(\mathbf{H}^i)\|_2^2 \tag{7}$$

where, ν^c is the class-specific inter-class regularization constant for the c^{th} class, and $\mathbf{H}^i$ refers to the learned representations of all the samples of the i^{th} class. Thus, the loss function of the proposed model consists of three terms: the first term is the standard RBM loss, which aims to learn a meaningful representation for the given input, the second and the third terms are responsible for minimizing the intra-class variations and maximizing the inter-class variations, respectively. It is important to note that since the entire loss function is minimized, the second term is added to the loss, while the third term is subtracted. The proposed model thus learns representations such that the learned features of one class are closer to each other as compared to the representations of other classes, thereby introducing discriminability during the feature learning process. We believe that incorporating such supervision at the time of feature learning facilitates learning of *discriminative* features for different classes, which further helps at the time of classification. For example, for a two-class classification problem containing images of classes *happy* and *sad*, the loss function of the proposed model for a sample belonging to class *happy* is written as follows:

$$\ell_{msRBM} = \ell_{RBM} + \lambda^{happy} \|h^{happy} - \mu(\mathbf{H}^{happy})\|_2^2 - \nu^{happy} \|h^{happy} - \mu(\mathbf{H}^{sad})\|_2^2 \tag{8}$$

where, h^{happy} is the learned representation of a sample belonging to class *happy*. Therefore, the proposed model aims to learn h^{happy} such that it is close to the mean representation of class *happy*, and different from the mean representation of class *sad*. Figure 3 presents a diagrammatic representation of the proposed msRBM for a two class problem.

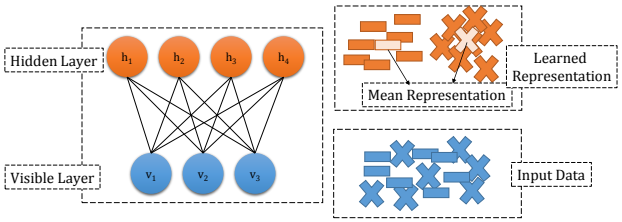

Figure 3: Representation of the proposed msRBM model for a two-class classification problem. v corresponds to the visible layer (input) and h corresponds to the hidden layer (representation). The proposed model learns features such that samples belonging to one class are closer to the mean representation of that class, but at a distance from the mean representation of the other class.

The proposed model can be extended to multiple layers to create the Mean Supervised Deep Boltzmann Machine. A DBM is created by stacking multiple RBMs, such that the input to the k^{th} layer RBM is the learned feature representation from the RBM at $(k-1)^{th}$ layer. The input to the first RBM is the training data. msDBM is created similarly, where the model learns mean supervised features at each layer in a hierarchical manner. For a l-layered msDBM, the loss function of the proposed architecture is thus written as:

$$\ell_{msDBM} = \ell_{DBM} + \sum_{k=1}^{l} \left(\lambda_k^c \|h_k^c - \mu(\mathbf{H}_k^c)\|_2^2 - \nu_k^c \sum_{i=1, i \neq c}^{n} \|h_k^c - \mu(\mathbf{H}_k^i)\|_2^2 \right) \tag{9}$$

where, h_k^c represents the feature vector of a sample v belonging to class c, at the k^{th} layer. Similarly, $\mathbf{H}_k^i$ refers to the representations of all samples of class i, at the k^{th} layer. ν_k^c refers to the inter-class regularization constant for the c^{th} class at the k^{th} layer. The above model can be trained in a greedy layer-by-layer manner [4], where only a single layer is optimized at a time, while keeping the remaining fixed. As with traditional techniques, Contrastive Divergence [16] is applied for solving the proposed model. Since the additional term is easily differentiable, its derivative is used for performing gradient descent in order to learn the optimal parameters of msDBM.

4. Expression Classification using msDBM

The proposed msDBM is used for feature extraction to perform expression classification using the pipeline demonstrated in Figure 4. Pre-processing is performed on the input images, which is then provided to the proposed msDBM model. msDBM is trained using Equation 9 for a seven class problem of *Anger, Disgust, Fear, Happy, Neutral, Sad,* and

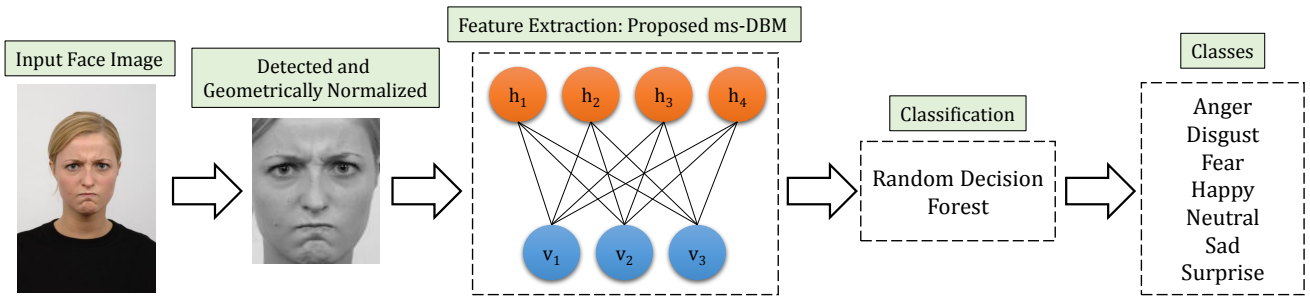

Figure 4: Pipeline used for expression classification in children using the proposed msDBM model.

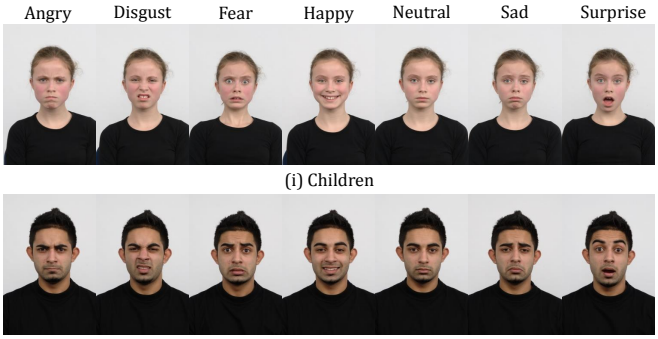

Figure 5: Sample images of two subjects from the Radboud Faces dataset [20] presenting six facial expressions, along with a neutral state.

Surprise. The learned features are given as input to a Random Decision Forest (RDF) classifier which is trained for performing seven-class classification. At the time of testing, a test image is projected on the learned msDBM model to obtain the representation of the given sample. This is followed by classification using the trained RDF model.

Face detection is performed on all images using Viola Jones face detector [43], followed by geometric normalization. The images are re-sized to a dimension of 64×64, and converted to gray-scale. A two layer msDBM model is used for feature extraction having dimensionality $[k, \frac{k}{4}, \frac{k}{4}]$, where k is the size of the input image. Since the input data consists of real values, the msRBM model is built over the Gaussian Bernoulli RBM, extended to msDBM. The model is trained for 100 epochs, and data augmentation is performed on the training and fine-tuning set by flipping across the y-axis and introducing illumination variations.

4.1. Datasets Used

There exists only two datasets containing images of children with expression variations: Radboud Faces Dataset [20], and CAFE dataset [25]. Out of these, the Radboud Faces dataset was released in 2010, containing a mix of adult and children face images having expression variations. Recently, in 2014, the CAFE dataset was released, which contains images pertaining to 154 children. The proposed model has been evaluated on these two datasets for expression classification in children and adults. Details regarding each dataset, along with the experimental protocols are explained below:

Radboud Faces Dataset (RaFD) [20] contains a total of 8,040 images corresponding to 67 subjects (57 adults and 10 children). A subset of the dataset containing six expressions: Anger, Disgust, Fear, Happy, Sad, Surprise, and a Neutral state are used in this experiment for all subjects. For each expression, three frontal images are provided in the dataset. Owing to the availability of both adult and children images, this dataset is used to report performance for child expression classification, as well as adult expression classification.

Child Affective Facial Expression (CAFE) Dataset [25] contains images pertaining to 154 subjects belonging to an age range of 2-8 years. There are a total of 1,192 images for six expressions: Anger, Disgust, Fear, Happy, Sad, Surprise, and a Neutral state. Two images are captured for each expression: with the mouth open and with mouth closed, except for the Surprise class. Since all children were not able to imitate all emotions successfully, some images were manually eliminated while creating the dataset, resulting in an imbalanced number of images across expressions.

4.2. Experimental Protocols

Child Expression Recognition: The proposed msDBM model is pre-trained with 1197 images corresponding to the adults from the Radboud Faces dataset. This pre-trained model is used for performing experiments on the Radboud Faces and CAFE dataset independently. For each dataset, images pertaining to children are divided into fine-tuning and testing set. Data pertaining to 70% subjects (children) make up the test set, while the remaining subjects form the fine-tuning partition. This is repeated five times for five times random sub-sampling cross validation. Mutual exclusion of subjects is maintained in the fine-tuning and test partitions.

Adult Expression Recognition: Face images of 70% of the subjects belonging to the Radboud Faces dataset are used

Table 1: Confusion matrix of the proposed msDBM model for expression classification on Radboud Faces dataset (children).

		Predicted						
Actual		Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
	Anger	**58.1**%	6.7%	0.0%	7.6%	15.2%	12.4%	0.0%
	Disgust	2.9%	**93.3**%	0.0%	1.9%	1.9%	0.0%	0.0%
	Fear	0.0%	0.0%	**74.3**%	1.0%	12.3%	1.0%	11.4%
	Happy	0.0%	9.5%	0.0%	**90.5**%	0.0%	0.0%	0.0%
	Neutral	0.0%	20.0%	1.0%	4.7%	**67.6**%	6.7%	0.0%
	Sad	15.2%	9.5%	12.4%	1.9%	19.1%	**41.9**%	0.0%
	Surprised	0.0%	0.0%	4.8%	0.0%	1.0%	0.0%	**94.2**%

for training the model, while the remaining 30% subjects are used as the test set. Five-times random sub-sampling cross-validation is performed for this experiment as well.

5. Results and Observations

Comparison has been performed with other deep learning feature extractors, namely Deep Boltzmann Machine (DBM) [34] and Stacked Denoising Autoencoder (SDAE) [42]. A pre-trained Convolutional Neural Network (CNN) based feature extractor, VGG-face [32], has also been used for comparison. Results are shown by using the pre-trained model directly and also by fine-tuning it for the given task. Since the proposed model incorporates supervision in the feature learning process, comparison has also been performed with Discriminative Restricted Boltzmann Machine (DRBM) [21]. In order to maintain consistency, the architecture of all feature learning models are exactly as those of the proposed msDBM model. Other than deep learning techniques, comparison has also been performed with two common techniques used in literature: Principle Component Analysis (PCA) [41] and Dense Scale Invariant Feature Transform (DSIFT) [27]. For all comparative techniques, as with msDBM, feature extraction is followed by learning a RDF classifier. A commercial API, Microsoft Cognitive Services [1] has also been used for performing comparison on the Radboud Faces dataset[1]. In order to understand the statistical significance of the results obtained by the proposed msDBM model, McNemar test [28] has been performed. The results obtained by the proposed model are compared with those obtained via other comparative algorithms. The p-values obtained via the McNemar test have also been reported and all claims are made at a confidence interval of 95%.

5.1. Child Expression Classification on the Radboud Faces Dataset

Table 2 presents the results obtained on the Radboud Faces dataset; the mean class-wise accuracy and standard deviation across five folds is reported. It can be observed that the proposed two layer msDBM achieves a classification accuracy of 75.0%, which is at least 3.3% better than other comparative unsupervised feature extractors. The improvement in performance obtained over traditional DBM can be attributed to the additional terms incorporated in the proposed msDBM algorithm. An improvement of 5.5% is also observed in comparison to the existing supervised RBM model (Discriminative RBM). Table 2 also presents the p-values of the McNemar test. It can be observed that the results obtained via the proposed msDBM model are statistically different from other comparative techniques (except DBM) at a confidence interval of 95%.

Table 2: Mean expression classification accuracy (%) on the Radboud Faces dataset (children) for five times random sub-sampling.

Algorithm	**Accuracy (%)**	p-**Value**
DBM [34]	71.7 ± 2.0	0.068
Discriminative RBM [21]	69.5 ± 4.2	0.010
SDAE [42]	70.0 ± 2.0	0.003
VGG-Face (CNN) [32]	52.2 ± 2.1	< 0.001
Fine-Tuned VGG-Face (CNN) [32]	67.4 ± 4.7	< 0.001
PCA [41]	68.4 ± 0.7	0.0004
DSIFT [27]	69.4 ± 4.2	0.018
Microsoft Cognitive [1]	17.7 ± 2.5	< 0.001
Proposed msDBM	$\mathbf{75.0 \pm 1.5}$	-

Comparison has also been drawn with a commercial API, Microsoft Cognitive Services [1], where the API gives a classification accuracy of only 17.7%. All face images were detected by the software, and there was no failure to process. It is important to note that the API has shown to perform well on adult faces, however, its performance on child face images further reinstates the need for dedicated attention and need for specialized approaches for the given problem. It is also interesting to note that the performance of CNN-based feature extractor, VGG-Face, is 52.2%,when a model pre-trained on large-scale face images is used. When the model is fine-tuned with child expression faces, the performance improves to 67.4%. This suggests that deep learning based based models need to be trained

[1]The license agreement of the CAFE dataset does not allow us to use the API.

Table 3: Confusion matrix of the proposed msDBM model for expression classification on the CAFE dataset.

		Predicted						
		Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Actual	Anger	**36.3**%	17.3%	6.3%	24.2%	7.7%	1.9%	6.3%
	Disgust	33.4%	**31.7**%	5.0%	15.3%	11.4%	1.8%	1.4%
	Fear	2.6%	3.0%	**35.2**%	16.1%	13.4%	1.9%	27.8%
	Happy	6.0%	3.3%	6.9%	**73.7**%	4.7%	0.3%	5.1%
	Neutral	4.5%	5.3%	8.0%	1.3%	**69.5**%	3.2%	8.2%
	Sad	17.2%	8.1%	13.1%	6.4%	27.6%	**21.8**%	5.8%
	Surprised	2.8%	1.3%	13.9%	4.4%	9.5%	0.6%	**67.5**%

Figure 6: Sample images from the Radboud Faces dataset [20] misclassified by the proposed msDBM model.

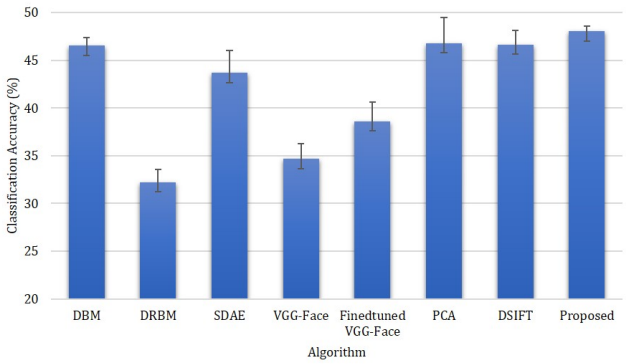

Figure 7: Bar graph representing the mean class-wise accuracies obtained on the CAFE dataset of the proposed model, along with other comparative techniques.

specifically for performing expression classification in children, as opposed to utilizing models trained on large-scale face image databases.

Upon analyzing the confusion matrix (Table 1) for the proposed msDBM algorithm, it can be observed that the proposed model performs least on the expression of *sad*, while performing over 90% for *disgust*, *happy*, and *surprise*. Figure 6 also presents some mis-classifications of the proposed model for the classes of *sad*, *disgust*, *surprise*, and *fear*. Most of these images appear to have a mix of

Table 4: Case study: Effect of combining *anger* and *disgust* into one class. Mean class-wise expression classification accuracy (%) on the CAFE dataset for five times random sub-sampling.

Algorithm	**Accuracy (%)**	p**-Value**
DBM [34]	54.6 ± 0.9	< 0.001
Discriminative RBM [21]	37.4 ± 1.5	< 0.001
SDAE [42]	53.9 ± 0.6	0.261
VGG-Face (CNN) [32]	41.3 ± 1.1	< 0.001
Fine-Tuned VGG-Face (CNN) [32]	44.2 ± 1.5	< 0.001
PCA [41]	44.7 ± 0.8	0.024
DSIFT [27]	51.7 ± 1.3	0.831
Proposed msDBM	$\mathbf{56.0 \pm 0.5}$	-

two expression, as opposed to just one. For example, most of the *sad* mis-classifications are because of the subtle expression changes brought about by the subject, and similarity with the *neutral* expression, resulting in a large number of images being mis-classified as *neutral*. Similarly, several imitations of the *fear* expression demonstrate similarity with the *surprise* class; an effect of which can also be observed from the confusion matrix, where most of the mis-classifications of *surprise* are into *fear*.

5.2. Child Expression Classification on the CAFE Dataset

Figure 7 presents the mean class-wise expression classification accuracies obtained for all the models, across five times random sub-sampling. It is important to note that the CAFE dataset contains images of children between the age of 2-8 years, while the Radboud Faces dataset contains images of teenagers. The expression variations observed in CAFE dataset are thus more challenging than those of Radboud Faces dataset. The proposed two layer msDBM achieves an average classification accuracy of 48.0%, which displays an improvement over other comparative models by at most 14% (Figure 7). Table 3 presents the confusion matrix obtained using the proposed msDBM model. It is observed that the model achieves a large mis-classification

percentage of the facial expression $disgust$ into $anger$, and vice versa. Moreover, consistent with earlier findings on the Radboud Faces dataset, a large mis-classification percentage of *surprise* into *fear* and vice versa is also observed.

Effect of Combining Expressions: Upon visually inspecting the mis-classifications, we observe very less inter-class variations between *anger* and *disgust*, and *surprise* and *fear*[2]. For example, Figure 8 presents the mean images corresponding to *anger* and *disgust* classes from the CAFE dataset. It can be observed that these images appear visually similar, especially for the upper part of the face. Moreover, the human evaluation reported in the original publication [26] was also analyzed, and a large overlap between the expressions of *anger* and *disgust* was found in the human responses. Building upon the above findings, as a case study, experiments are performed by combining the expressions of *anger* and *disgust* into one class. From Table 4, it can be observed that all models yield improved performance upon combining the two classes. The proposed msDBM achieves a mean class-wise classification accuracy of 56.0%, outperforming other comparative models.

It is interesting to observe that the proposed model yields an improvement of over 18% over the Discriminative RBM model, thereby motivating the inclusion of the inter-class and intra-class terms in the proposed formulation. The enhanced performance observed across all models upon combining *anger* and *disgust* into one class suggests low inter-class variations across the combined classes, thereby reinforcing the challenging nature of child expression classification. A standard deviation of less than 1% is observed for all experiments, which suggests that the proposed model is robust to the training-testing partitions, and not biased towards one particular data distribution. The low child expression classification performance across models presents a need for dedicated and focused research in this area.

5.3. Adult Expression Classification on the Radboud Faces Dataset

The effectiveness of the proposed model has also been evaluated on the adult face images of the Radboud Faces dataset. The proposed msDBM model achieves a mean classification accuracy of 85.9% over random five times sub-sampling cross validation. Table 5 reports the classification accuracy of the proposed model, along with other comparative techniques. Improvement is observed from other techniques, including the CNN-based feature extractor (VGG-Face) pre-trained on adult face images. Upon analyzing the results, we observe that the proposed model achieves an accuracy of almost 100% across all five folds for the *happy* expression, whereas the least accuracy is obtained for *sad* expression. The p-values obtained via the McNemar test further demonstrate statistical difference of the proposed model in comparison with other algorithms, for a confidence interval of 95%.

[2]The samples cannot be shown in the paper due to restrictions from the license agreement.

Figure 8: Mean images obtained from the CAFE dataset [25] for *anger* and *disgust* expressions.

Table 5: Mean class-wise expression classification accuracy (%) on the Radboud Faces dataset (adults) for five times random sub-sampling.

Algorithm	**Accuracy (%)**	p**-Value**
DBM [34]	83.9 ± 4.2	0.003
Discriminative RBM [21]	83.8 ± 5.7	0.005
SDAE [42]	82.6 ± 2.8	< 0.001
VGG-Face (CNN) [32]	69.2 ± 3.9	< 0.001
Fine-tuned VGG-Face (CNN) [32]	80.3 ± 1.9	< 0.001
PCA [41]	83.6 ± 3.9	0.009
DSIFT [27]	80.4 ± 2.8	< 0.001
Proposed msDBM	$\mathbf{85.9 \pm 2.2}$	-

6. Conclusion

This research addresses the important and challenging yet unexplored problem of expression classification of child face images. A novel supervised deep learning model, termed as Mean Supervised Deep Boltzmann Machine (msDBM) is proposed for the given task. The model learns discriminative representations by minimizing the intra-class variations and maximizing the inter-class variations with respect to the mean feature vectors. The performance of the proposed model is evaluated on two challenging datasets: Radboud Faces and CAFE, for child and adult expression classification. Experimental results motivate the utility of the proposed msDBM model for automated facial expression classification. While the proposed model achieves state-of-the-art performance on both the datasets, however, its performance suggests that automated expression classification on children (especially in the range of 2-8 years) requires further research and dedicated attention.

7. Acknowledgement

This research is partially supported through the Infosys Center for Artificial Intelligence, IIIT-Delhi. S. Nagpal is supported via the TCS PhD fellowship.

References

[1] Microsoft cognitive services. https://www.microsoft.com/cognitive-services/en-us/emotion-api/.

[2] Dawood Adel AL CHANTI and Alice Caplier. Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. *IEEE Transactions on Affective Computing*, 2018.

[3] Keith Anderson and Peter W McOwan. A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(1):96–105, 2006.

[4] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Conference on Neural Information Processing Systems*, pages 153–160, 2006.

[5] Vinay Bettadapura. Face expression recognition and analysis: The state of the art. *CoRR*, abs/1203.6722, 2012.

[6] Andrew J Calder, A.Mike Burton, Paul Miller, Andrew W Young, and Shigeru Akamatsu. A principal component analysis of facial expressions. *Vision Research*, 41(9):1179 – 1208, 2001.

[7] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *International Journal of Computer Vision*, 2019.

[8] Jeffrey F. Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *IEEE International Conference on Multimodal Interfaces*, pages 491–496, 2002.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200, 1992.

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

[12] Wenfei Gu, Cheng Xiang, Y.V. Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80 – 91, 2012.

[13] Liangke Gui, Tadas Baltrušaitis, and Louis-Philippe Morency. Curriculum learning for facial expression recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 505–511, 2017.

[14] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, 6(1):1–12, 2015.

[15] Catherine Herba and Mary Phillips. Annotation: Development of facial expression recognition from childhood to adolescence: behavioural and neurological perspectives. *Journal of Child Psychology and Psychiatry*, 45(7):1185–1198, 2004.

[16] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

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

[19] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron Courville, Pascal Vincent, Roland Memisevic, Christopher Pal, and Yoshua Bengio. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[20] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. Presentation and validation of the Radboud faces database. *Cognition and Emotion*, 24(8):1377–1388, 2010.

[21] Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted boltzmann machines. In *ACM International Conference on Machine Learning*, pages 536–543, 2008.

[22] James J Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. Automated facial expression recognition based on facs action units. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395, 1998.

[23] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*, pages 143–157, 2014.

[24] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[25] Vanessa LoBue. *The Child Affective Facial Expression CAFE set*. Databrary, 2014.

[26] Vanessa LoBue and Cat Thrasher. The child affective facial expression (CAFE) set: validity and reliability from untrained adults. *Frontiers in Psychology*, 5, 2015.

[27] David G Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[28] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

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

[30] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on

small datasets using transfer learning. In *ACM on International Conference on Multimodal Interaction*, pages 443–449, 2015.

[31] Maja Pantic and Leon JM Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[32] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, pages 41.1–41.12, 2015.

[33] Mahesh Ramachandran, Shaohua Kevin Zhou, Divya Jhalani, and Rama Chellappa. A method for converting a smiling face to a neutral face with applications to face recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii/977–ii/980, 2005.

[34] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[35] Anush Sankaran, Gaurav Goswami, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Class sparsity signature based restricted boltzmann machine. *Pattern Recognition*, 61:674 – 685, 2017.

[36] Caifeng Shan, Shaogang Gong, and P. W. McOwan. A comprehensive empirical study on linear subspace methods for facial expression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 153–153, 2006.

[37] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ACM on International Conference on Multimodal Interaction*, pages 517–524, 2013.

[38] Maneet Singh, Shruti Nagpal, Mayank Vatsa, and Richa Singh. Are you eligible? predicting adulthood from face images via class specific mean autoencoder. *Pattern Recognition Letters*, 119:121 – 130, 2019.

[39] Maneet Singh, Shruti Nagpal, Mayank Vatsa, Richa Singh, and Afzel Noore. Supervised COSMOS autoencoder: Learning beyond the euclidean loss! *CoRR*, abs/1810.06221, 2018.

[40] Ying-li Tian, Takeo Kanade, and Jeffrey F. Cohn. Multimodal interface for human-machine communication. chapter Recognizing Action Units for Facial Expression Analysis, pages 32–66. World Scientific Publishing Co., Inc., 2002.

[41] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[42] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

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

[44] Shuyang Wang, Zhengming Ding, and Yun Fu. Feature selection guided auto-encoder. In *AAAI Conference on Artificial Intelligence*, 2017.

[45] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[46] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[47] Kun Zeng, Jun Yu, Ruxin Wang, Cuihua Li, and Dacheng Tao. Coupled deep autoencoder for single image super-resolution. *IEEE Transactions on Cybernetics*, 47(1):27–37, 2017.

[48] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018.

[49] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[50] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and W Bastiaan Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):38–52, 2011.

[51] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.