Multimodal Emotion Recognition using Deep Learning Architectures

Hiranmayi Ranganathan, Shayok Chakraborty and Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing (CUbiC)
Arizona State University
{hiranmayi.ranganathan,shayok.chakraborty,panch}@asu.edu

Abstract

Emotion analysis and recognition has become an interesting topic of research among the computer vision research community. In this paper, we first present the emoFBVP database of multimodal (face, body gesture, voice and physiological signals) recordings of actors enacting various expressions of emotions.The database consists of audio and video sequences of actors displaying three different intensities of expressions of 23 *different emotions along with facial feature tracking, skeletal tracking and the corresponding physiological data. Next, we describe four deep belief network (DBN) models and show that these models generate robust multimodal features for emotion classification in an unsupervised manner. Our experimental results show that the DBN models perform better than the state of the art methods for emotion recognition. Finally, we propose convolutional deep belief network (CDBN) models that learn salient multimodal features of expressions of emotions. Our CDBN models give better recognition accuracies when recognizing low intensity or subtle expressions of emotions when compared to state of the art methods.*

1. Introduction

In recent years, there has been a growing interest in the development of technology to recognize an individual's emotional state. There is also an increase in the use of multimodal data (facial expressions, body expressions, vocal expressions and physiological signals) to build such technologies. Each of these modalities have very distinct statistical properties and fusing these modalities helps us learn useful representations of the data. Emotion recognition is a process that uses low level signal cues to predict high level emotion labels. Literature has shown various techniques for generating robust multimodal features [1]-[4] for emotion recognition tasks. The high dimensionality of the data, the non-linear interactions across the modalities along with the fact that the way an emotion is expressed varies across people complicate the process of generating emotion specific features [5],[6]. Deep architectures and learning techniques have shown to overcome these limitations by capturing complex non-linear feature interactions in multimodal data [7].

In this paper, as our first contribution, we present the *emoFBVP* database of multimodal recordings-facial expressions, body gestures, vocal expressions and physiological signals, of actors enacting various expressions of emotions. The database consists of audio and video sequences of actors enacting 23 different emotions in three varying intensities of expressions along with facial feature tracking, skeletal tracking and the corresponding physiological data. This is one of the first emotion databases that has recordings of varying intensities of expressions of emotions in multiple modalities recorded simultaneously. We strongly believe that the affective computing community will greatly benefit from the large collection of modalities recorded. Our second contribution investigates the use of deep learning architectures - DBNs and CDBNs for multimodal emotion recognition. We describe four deep belief network (DBN) models and show that they generate robust multimodal features for emotion classification in an unsupervised manner. This is done to validate the use of our *emoFBVP* database for multimodal emotion recognition studies. The DBN models used are extensions of the models proposed by [7] for audio-visual emotion classification. Finally, we propose convolutional deep belief network (CDBN) models that learn salient multimodal features of low intensity expressions of emotions.

2. Related Work

Previous research has shown that deep architectures effectively generate robust features by exploiting the complex non-linear interactions in the data [8]. Deep architectures and learning techniques are very popular in the speech and language processing community [9]-[11]. Ngiam *et al*. [12] report impressive results on audio-visual speech classification. They use sparse Restricted Boltzmann Machines (RBMs) for cross-modal learning, shared representation learning and multimodal fusion on CUAVE and AVLetters dataset. Srivastava *et al*. [13] applied multimodal deep

belief networks to learn joint representations that outperformed SVMs. They used multimodal deep Boltzmann machines to learn a generative model of images and text for image retrieval tasks. Kahou *et al*. used an ensemble of deep learning models to perform emotion recognition from video clips [14]. This was the winning submission to the Emotion Recognition in the Wild Challenge [15]. Deep learning has also been applied in many visual recognition studies [16]-[20]. Our research is motivated by the above recent approaches in multimodal deep learning. In this paper, we focus on applying deep architectures for multimodal emotion recognition using face, body, voice and physiological signal modalities. We apply extensions of known DBN models for multimodal emotion recognition using the *emoFBVP* database and investigate recognition accuracies to validate the utility of the database for emotion recognition tasks. To the best of our knowledge, the use of DBNs for multimodal emotion recognition of data comprising of all the above mentioned modalities (facial expressions, body gestures, vocal expressions and physiological signals) has not been explored by the affective research community.

Recent developments in deep learning techniques exploit the use of single layer building blocks called as Restricted Boltzmann Machines (RBMs) [21] to build DBNs in an unsupervised manner. DBNs are constructed by greedy layer-wise training of stacked RBMs to learn hierarchical representations from the multimodal data [22]. RBMs are undirected graphical models that use binary latent variables to represent the input. Like [7], we also use Gaussian RBMs for training the first layer of the network. The visible units of the first layer are real-valued. The deeper layers are trained using Bernoulli-Bernoulli RBMs that employ visible and hidden units that are binary valued. The joint probability distribution for a Gaussian RBM with visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ is given as follows:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z}\, exp\left(-E(\mathbf{v}, \mathbf{h})\right), \tag{1}$$

The corresponding energy function with $\mathbf{q} \in \mathbb{R}^D$ and $\mathbf{r} \in \mathbb{R}^K$ as biases of visible and hidden units and $\mathbf{W} \in \mathbb{R}^{D \times K}$ as weights between visible units and hidden units is given as:

$$E(\mathbf{v}, \mathbf{h}) = \frac{1}{2\sigma^2} \sum_i v_i^2 - \frac{1}{\sigma^2} \left(\sum_i q_i v_i + \sum_j r_j h_j + \sum_{i,j} v_i W_{i,j} h_j \right), \tag{2}$$

These parameters are learned using a technique called contrastive divergence, explained in [23]. σ is a hyperparameter and Z is a normalization constant. The conditional probability distributions of the Gaussian RBM are as follows:

$$P(h_j = 1 \mid \mathbf{v}) = \textit{sigmoid}\left(\frac{1}{\sigma^2} \left(\sum_i W_{i,j} v_i + r_j \right) \right), \tag{3}$$

$$P(v_i \mid \mathbf{h}) = \mathcal{N}\left(v_i; \sum_j W_{i,j} h_j + q_i, \sigma^2 \right). \tag{4}$$

We include a regularization penalty as in [16] given as:

$$\lambda \sum_{j=1}^{k} \left| p - \frac{1}{m} \sum_{l=1}^{m} E\left[h_j^{(l)} \mid \mathbf{v}^{(l)} \right] \right|^2 \tag{5}$$

Here, *E*[.] is the conditional expectation given the data, λ is a regularization parameter, and *p* is a constant that specifies the activation of the hidden unit.

Convolutional deep belief networks (CDBNs) [24] are similar to DBNs and can be trained in a greedy layer-wise fashion. Lee *et al* [24] used CDBNs to show good performance in many visual recognition tasks. Convolutional restricted Boltzmann machines (CRBMs) [24]-[26] are building blocks for (CDBNs). In a CRBM [22], the weights between the hidden units and visible units are shared among all locations in the hidden layer. The CRBM consists of two layers: an input (visible) layer **V** and a hidden layer **H**. The hidden units are binary-valued, and the visible units are binary-valued or real-valued. Please refer to Lee *et al.* [24] for the expression for the energy function, conditional and joint probabilities. In this paper, we use CRBMs with probabilistic max-pooling as building blocks for convolutional deep belief networks. For training the CRBMs, we use contrastive divergence [23] to approximate the gradient of the log-likelyhood term effectively. Like in [16], we add a sparsity penalty term as well. Post training, we stack the CRBMs to form a CDBN.

The rest of the paper is organized as follows. Section 3 describes the *emoFBVP* database and its salient properties. The descriptions of experimental set up for deep learning, feature extraction techniques and baseline models are in Section 4. Section 5 introduces our D*emoDBN* models and investigates their usage for multimodal emotion recognition in an unsupervised context. Section 6 describes our CDBN models and investigates their usability to recognize subtle or low intensities of expressions of emotions. Finally, we share our conclusions and future work in Section 7.

3. *emoFBVP* Database

To study human emotional experience and expression in more detail and to develop benchmark methods for automatic emotion recognition, researchers are in need of rich sets of data. We have recorded responses of actors to affective emotion labels using four different modalities - facial expressions, body expressions, vocal expressions and physiological signals. Along with the multimodal recordings, we provide facial feature tracking and skeletal tracking data. The recordings of all the data are rated through an evaluation form completed by the actors immediately after each excerpt of acting emotions. The recordings of this database are synchronized to enable study of simultaneous emotional responses using all the modalities.

Ten participants (who are professional actors) were involved in data capture, and every participant displayed 23

different emotions. Recordings of each emotion were done six times: three in a standing position and three in a seated position when the body gestures and facial expressions were tracked and recorded along with vocal expressions, physiological data and activity respectively. Therefore, the database provides six examples of each of the 23 emotions in varying intensities of expression. The two sessions of recordings (standing and seated) are independent of each other. This makes it possible to use our database for unimodal (using only face, body, physiological signals or activity), bimodal (face and voice, body and voice, etc.) and multimodal emotion recognition studies. Our database provides information about the affective communication skills of every participant. It also provides evaluation details about the confidence of expression of emotion, intensity of expression of emotion and level of ease of expression of emotion using facial expressions, body gestures and vocal expressions. Our database, therefore, provides both valuable expression data and metadata that will contribute to the ongoing development of emotion recognition algorithms.

3.1. Properties of *emoFBVP* Database

The *emoFBVP* database consists of responses of participants to affectively stimulating emotion labels. Different modalities of measurement require different equipments. We set up apparatus to record face videos, facial feature tracking, body gesture videos, skeletal tracking, vocal expressions and physiological signals simultaneously. The sensor equipments used to facilitate the recordings of the modalities include the Microsoft Kinect Sensor, the Zephyr BioHarness and wrist-worn accelerometers.

Recruitment: Participants were recruited after a city-wide call for people who have completed basic coursework in acting/non-verbal communication. They were requested to provide their formal consent to participate by signing a consent form that gave a detailed description of the purpose and data capture procedure of the study.

Participant information and assessment: Participants were asked to provide their age range, gender and ethnic background. They answered questions to help assess their affective communication skills. In particular, each participant rated his/her overall skill in expressing emotions, their affective communication skill using facial expressions, body gestures, vocal expressions and how emotionally expressive they were (on a scale of 1 to 5, where 5 was very effective). This information is provided in the database as part of metadata.

Data capture: Participants were instructed to perform/express emotions using facial expressions, body gestures and vocal expressions. Their acted responses (using face, body, voice and physiological modalities) were recorded for 23 emotion labels: Happy, Sad, Anger, Disgust, Fear, Surprise, Boredom, Interest, Agreement, Disagreement, Neutral, Pride, Shame, Triumphant, Defeat, Sympathy, Antipathy, Admiration, Concentration, Anxiety, Frustration, Content and Contempt during two sessions. During the first session, participants expressed each of the 23 emotions in three varying intensities of expression in a standing position; their body gestures were recorded and their skeletal representation was tracked. During the second session, participants expressed each of the 23 emotions in three varying intensities of expression in a seated position; their facial expressions were recorded and facial features were tracked. Physiological signals, vocal expressions and acceleration were recorded continuously during both sessions. After recording their responses to each emotion label, they filled out an evaluation form. Here, they provided details about their level of comfort in acting/expressing emotions in each modality. They were asked to rate their confidence level with expressing each emotion, their intensity while expressing each emotion and their ease of expressing each emotion using facial expressions, body gestures and vocal expressions on a scale of 1 (low) to 5 (high). Participants were given about 2 minutes between expressing different emotions. They used this time to complete the evaluation form and think about their responses to the next emotion label. Further, participants were requested to share their comments and feedback about the data capture process. This information is also made available in the database as part of metadata.

Ground truth: The database was labeled by three evaluators to help improve the authenticity of expression of emotions. This is available as metadata along with the database.

This database is comprised of 1380 samples of audio sequences, video sequences of face and body and physiological data corresponding to various expressions of emotions. The salient properties of the database are summarized below.

Face and Voice: We obtained facial expression video sequences and facial tracking data from the Microsoft Kinect for Windows sensor. All video sequences were recorded at the rate of 30 fps with a video resolution of 640×480 pixels. The sequences are of variable length lasting between 600 and 2000 frames. We used Brekel Kinect Pro Face software to record 3D face tracking data obtained from the Kinect sensor. The face tracking data consists of 3D head position and rotation information, 3D coordinates for 11 animation units and 3D coordinates for 11 shape units for each frame of the video. Figure 1 shows a snapshot of a subject portraying emotion, "Surprise" along with a 3D face-mesh corresponding to the emotion. The animation and shape units tracked are shown as yellow dots over the subject's face and an indicator shows their presence or absence. The voice data was recorded using Microsoft Kinect for Windows sensor. The Kinect sensor includes a four-element linear microphone array that captures audio data at 24-bit

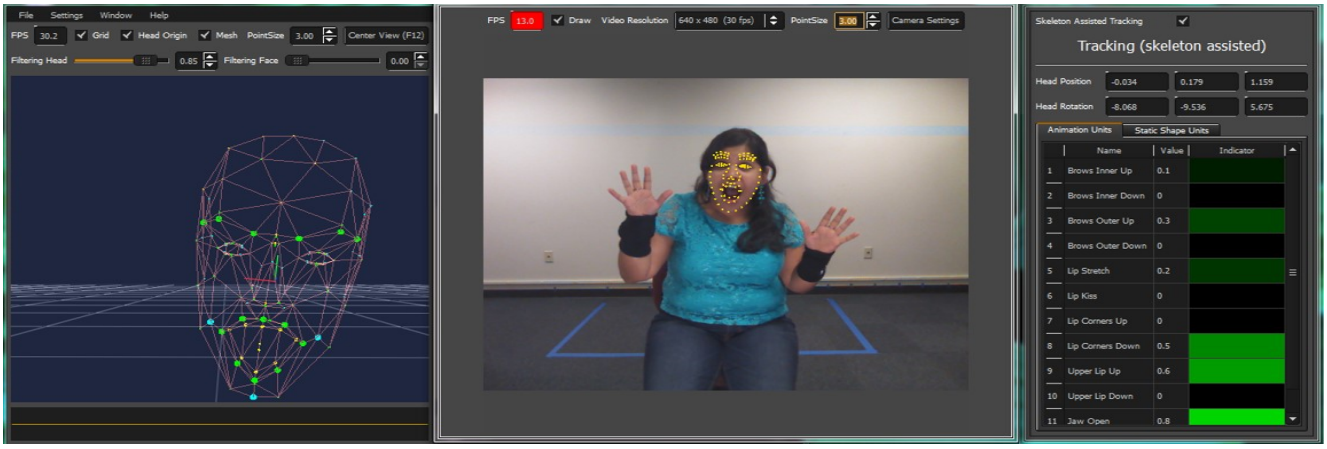

Figure 1. Snapshot of a subject portraying emotion, "Surprise". Left: 3D face-mesh corresponding to the emotion. Middle: Brekel Kinect Pro Face tracking animation and shape units shown as yellow dots over the subject's face. Right: Tracking indicator showing presence or absence of animation units at each instant.

resolution. This allows accuracy across a wide dynamic range of voice data. The sensor enables high quality audio capture with focus on audio coming from a particular direction with beamforming. The audio sequences are provided in standard .wav format and are synchronized with the face and body sequences.

Body: We obtained body expression video sequences and skeletal tracking data from the Microsoft Kinect for Windows sensor. All video sequences were recorded at the rate of 30 fps with a video resolution of 640×480 pixels. The sequences are of variable length and synchronized with the face and audio data. We used the Brekel Kinect Pro for Body software to record the skeletal tracking data. The skeletal tracking data provides 3D coordinates of twenty joints of the users body along with 3D coordinates for hand, foot and head rotations for each frame of the video sequence.

Physiological data: We obtained physiological data using the Zephyr BioHarness. This provides measurements of the users Heart Rate, ECG R-R interval, Breathing Rate, Posture, Activity level and Peak Acceleration. This data is available in .csv and synchronized using time stamps with the face, body and voice data.

Availability of Database: The *emoFBVP* database is freely available for download to the academic research community, and is easily accessible through a web-interface [1].

4. Experimental Setup

Here, we first explain the need for developing and training DBNs on the *emoFBVP* database for multimodal emotion recognition. Sub-section 4.1 describes the extracted emotion-specific features and the baseline models used for comparison.

One of the best ways to validate the authenticity of a new emotion database is to apply known methods of feature extraction and investigate the performance of state of the art models on the collected data. In order to show the usefulness of our *emoFBVP* database, we apply extensions of known deep learning techniques for feature learning and investigate the performance accuracies for emotion recognition in unimodal, bimodal and multimodal scenarios. Our database also provides facial feature tracking and skeletal tracking data. We investigate the advantages of adding these features to the deep models using feature selection methods. Kim *et al.* [7] developed a suite of deep belief network models that showed improvements in emotion classification performance over baselines that do not use deep learning. They perform rigorous experiments to show that deep learning techniques can be used for multimodal emotion recognition. We build extensions of these models, train them on our multimodal data, perform similar experiments and investigate the usefulness of the *emoFBVP* database for emotion recognition.

4.1. Multimodal Features and Baseline Models

The *emoFBVP* database allows the study of unimodal, bimodal and multimodal emotion recognition studies. In this paper, we consider primary expressions of emotions: Angry, Happy, Sad, Disgust, Fear, Surprise and Neutral for multimodal emotion recognition. Using information from the metadata available along with the database, we split the data into three sets: (1) Ideal data (all three evaluators agree on the emotion label), (2) Non-ideal data (majority of the evaluators agree on the emotion label) and (3) Combined data (combination of ideal and non-ideal data).

The *emoFBVP* database gives us facial feature track-

[1] http://emofbvp.org/

ing data and skeletal tracking data for all expressions of emotions. We compute prosodic and spectral features like pitch, energy and mel-frequency filter banks from the voice data of the database [27]. We compute the mean, variance, lower and upper quantiles and quantile range of the audio features (prosodic and spectral features from vocal expressions), video features (facial tracking features from facial expressions and skeletal tracking features from body gestures) and physiological signals. These features are used to assess the utility of adding feature extraction techniques to the features learnt from deep learning while performing multimodal emotion recognition. All of these features are normalized to avoid person dependency [28]. We use 180 features extracted from vocal expressions, 540 features extracted from facial expressions, 540 features extracted from body gestures and 120 features extracted from physiological signal data, giving a total of 1380 features.

We need baseline models for comparison and validation of results obtained from using deep learning techniques. For this, these models should not use features generated via deep learning. We follow [7] and use two SVM models with radial basis function (RBF) kernels. The SVMs are trained using one-versus-all approach. The first SVM baseline model uses Information Gain (IG) for supervised feature selection and the second SVM baseline model uses Principal Feature Analysis (PFA) for unsupervised feature selection. We apply IG to each emotion class; this gives us seven sets of emotion specific features. The baseline models are optimized using leave-one-person-out cross validation method. The parameters of the baseline models are: the number of selected features using IG and PFA (chosen over $\{60, 120, 180\}$ for each ideal, non-ideal and combined data types), the value of the box constraint for the soft margin in the SVM ($C = 1$) and the scaling factor in the RBF kernels ($sigma = 8$).

5. D*emoDBN* Models

In this section, we introduce our D*emoDBN* models and study their usage for multimodal emotion recognition in an unsupervised manner. Sub-section 5.1 gives classification accuracies for emotion classification using our D*emoDBN* models on the *emoFBVP* database in bimodal and multimodal scenarios. In order to show generalizability, we also deploy our unimodal and multimodal DBN models to perform emotion recognition on popular and standard emotion databases.

An illustration of the proposed D*emoDBN* models is given in Figure 2. We extend the DBN models proposed in [7] to include video data from body gestures and physiological signal data of expressions of emotions. D*emoFBVP*, (Figure 2(a)), is a basic two-layer DBN that learns features from vocal expressions, facial expressions, body gestures and physiological signals individually in the first hidden layer. All of these features are concatenated and fed as input to the second hidden layer. f+D*emoFBVP*, (Figure 2(b)), is a two-layer DBN that uses supervised feature selection using IG selection methods prior to DBN pre-training. D*emoFBVP*+f, (Figure 2(c)), is also a two-layer DBN that uses supervised feature selection using IG post DBN pre-training. 3D*emoFBVP*, (Figure 2(d)), is a three-layer DBN that stacks another RBM layer over the second layer. In summary, the four D*emoDBN* models are:

1. D*emoFBVP*: basic two-layer DBN model.
2. f+D*emoFBVP*: two-layer DBN with feature selection prior to the training of D*emoFBVP*.
3. D*emoFBVP*+f: two-layer DBN with feature selection added post training of D*emoFBVP*.
4. 3D*emoFBVP*: three-layer DBN model.

The hyperparameters are selected using cross validation over the training data for ideal, non-ideal and combined data types. We use leave-one-speaker-out cross validation for selecting the sparseness parameters. The number of hidden nodes of the two-layer DBN and weight regularization parameters are kept constant. The sparsity parameters of the bias of vocal expressions, facial expressions, body gestures and physiological signal data are selected as $\{0.1, 0.2\}$, $\{0.02, 0.1\}$, $\{0.02, 0.1\}$ and $\{0.2, 0.3\}$ respectively. The sparsity parameters of λ for all multimodal features are fixed at 6. The number of hidden units in the first layer of the D*emoDBN* models (D*emoFBVP*, D*emoFBVP*+f, 3D*emoFBVP*) is 2000. This results from concatenating 700 units from facial expressions, 700 units from body gestures, 400 units from vocal expressions and 200 units from physiological signal features. We fix the number of second layer hidden units for these D*emoDBN* models as 200. For f+D*emoFBVP*, we perform feature selection using IG and select 100 features from vocal expressions, 200 features each from facial expressions and body gestures and 80 features from physiological signals. We then pre-train the RBM layer with 100 nodes for vocal expressions, 150 nodes each for facial video and body gestures and 50 nodes for physiological signals features. Here, the sparsity parameters are chosen over $\{0.1, 0.6\}$ for all the RBMs and the λ sparsity parameter is kept at 6. The features learned by the RBM are concatenated and used to pre-train the first layer of the f+D*emoFBVP* model.

We propose D*emoFV* (face and voice), D*emoBV* (body and voice), D*emoFBV* (face, body and voice) DBN models similar to D*emoFBVP* models and their variants: f+D*emoFV*, D*emoFV*+f, 3D*emoFV*, f+D*emoBV*, D*emoBV*+f, 3D*emoBV*, f+D*emoFBV*, D*emoFBV*+f and 3D*emoFBV* models. We use the same SVM models as the baseline models to classify the output of all the D*emoDBN* models.

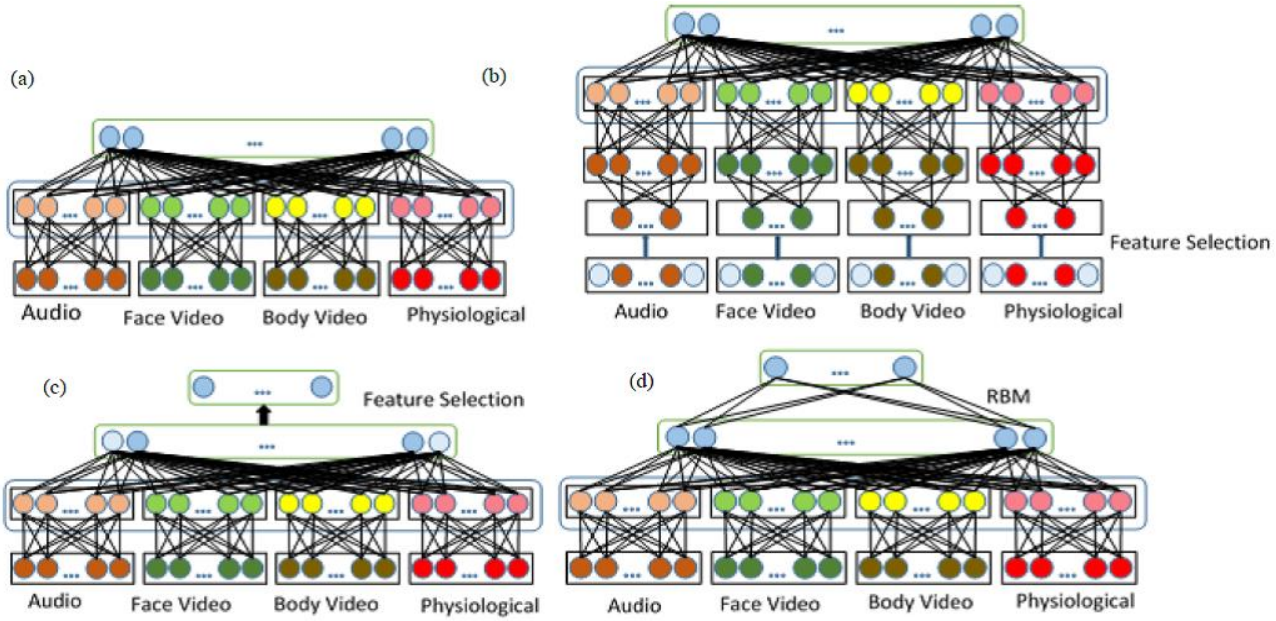

Figure 2. Proposed multimodal D*emoDBN* models: (a) D*emoFBVP*, (b) f+D*emoFBVP*, (c) D*emoFBVP*+f, and (d)3*DemoFBVP*

.

5.1. Results for D*emoDBN* Models

Tables 1-4 give classification accuracies for emotion classification using D*emoDBN* models on *emoFBVP* database. Our results demonstrate that the proposed DBN models generate robust multimodal features for emotion classification and that we can successfully apply D*emoDBN* models for emotion recognition using facial expressions, body gestures, vocal expressions and physiological signal modalities. These results also validate the multimodal data collected in our *emoFBVP* database and show that the database can be used for unimodal, bimodal and multimodal emotion recognition studies.

We find that the three-layer D*emoDBN* models achieve maximum percentage classification accuracies for ideal and non-ideal data. For the combined data type, the D*emoDBN*+f models achieve the maximum accuracies. The baseline models using IG methods always perform better than the baseline models using PFA methods for ideal, non-ideal and combined data types.

Results for D*emoFV* models: *Ideal data*: The D*emoFV* models achieve classification accuracies ranging from 82.98% to 86.56%. The difference between the performance accuracies of the baseline model using IG and the baseline model using PFA methods is 3.99%. The difference between classification accuracies of 3D*emoFV* and baseline model using PFA is 4.23%.

Non-ideal data: Here, the baseline models using IG and baseline models using PFA methods show similar performance. The 3D*emoFV* model shows a small improvement in performance over the D*emoFV*+f (0.59%) and f+D*emoFV* (1.2%) models.

Combined data: The D*emoFV*+f model achieves a maximum accuracy of 78.32% for combined data.

Results for D*emoBV* models: *Ideal data*: The D*emoBV* models achieve classification accuracies ranging from 80.78% to 84.99%. The difference between the performance accuracies of baseline models using IG methods and baseline models using PFA methods is 3.97%. The difference between classification accuracies of 3D*emoBV* and the baseline model using PFA is 4.74%.

Non-ideal data: The baseline models using IG and baseline models using PFA methods show similar performance. The 3D*emoBV* model shows a small improvement in performance over the D*emoBV*+f (0.75%) and f+D*emoBV* (1.22%) models.

Combined data: The D*emoBV*+f model achieves a maximum accuracy of 76.32% for combined data.

Results for D*emoFBV* models: *Ideal data*: The D*emoFBV* models achieve classification accuracies ranging from 83.10% to 86.68%. The difference between the performance accuracies of the baseline model using IG and the baseline model using PFA is 3.99%. The difference between classification accuracies of 3D*emoFBV* and the baseline model using PFA is 4.25%.

Non-ideal data: The two baseline models show similar performance. The 3D*emoFBV* model shows a small improvement in performance over the D*emoFBV*+f (0.61%) and

Data Type	Base line IG	Base line PFA	D*emo FV*	D*emo FV+f*	*f*+D*emo FV*	3D*emo FV*
Ideal	86.32	82.33	82.98	84.92	84.56	**86.56**
Non-Ideal	64.78	64.95	65.52	65.82	65.21	**66.41**
Comb-ined	75.64	75.83	77.25	**78.32**	77.78	77.62

Table 1. Classification accuracy (%) for D*emoFV* models

Data Type	Base line IG	Base line PFA	D*emo BV*	D*emo BV+f*	*f*+D*emo BV*	3D*emo BV*
Ideal	84.22	80.25	80.78	82.88	82.46	**84.99**
Non-Ideal	62.66	62.86	63.67	63.89	63.42	**64.64**
Comb-ined	73.64	73.83	75.25	**76.32**	75.78	75.62

Table 2. Classification accuracy (%) for D*emoBV* models

Data Type	Base line IG	Base line PFA	D*emo FBV*	D*emo FBV+f*	*f*+D*emo FBV*	3D*emo FBV*
Ideal	86.42	82.43	83.10	84.99	84.68	**86.68**
Non-Ideal	64.89	65.75	68.66	68.93	68.34	**69.54**
Comb-ined	75.77	75.90	77.38	**78.45**	77.89	77.78

Table 3. Classification accuracy (%) for D*emoFBV* models

Data Type	Base line IG	Base line PFA	D*emo FBVP*	D*emo FBVP+f*	*f*+D*emo FBVP*	3D*emo FBVP*
Ideal	89.41	85.33	86.20	87.82	87.52	**90.10**
Non-Ideal	68.89	68.71	71.14	71.84	71.22	**73.11**
Comb-ined	79.82	79.90	82.28	**83.10**	82.54	82.40

Table 4. Classification accuracy (%) for D*emoFBVP* models

f+D*emoFBV* (1.2%) models.
Combined data: The D*emoFBV+f* model achieves a maximum accuracy of 78.45% for combined data.
Results for D*emoFBVP* models: *Ideal data*: The D*emoFBVP* models achieve classification accuracies ranging from 86.20% to 90.10%. The difference between the performance accuracies of the baseline models is 4.08%. The difference between classification accuracies of 3D*emoFBVP* and PFA baseline model is 4.77%.
Non-ideal data: The two baseline models show similar performance. The 3D*emoFBVP* model gives a small improvement in performance over the D*emoFBVP+f* (1.27%) and *f*+D*emoFBVP* (1.89%) models.
Combined data: The D*emoFBVP+f* model achieves a maximum accuracy of 83.10% for combined data.

Our results show that we can successfully employ D*emoDBN* models for the task of multimodal emotion recognition. The proposed D*emoDBN* models successfully retain complex non-linear feature relationships that exist between the different modalities for ideal, non-ideal and combined data types (as shown by the performance accuracies achieved). Our results highlight the importance of feature learning using deep architectures over unsupervised feature selection for bimodal and multimodal emotion classification using the *emoFBVP* database of facial expressions, body gestures, vocal expressions and physiological signals. With this study, we validate the use of the *emoFBVP* database for emotion recognition studies and believe that the affective computing community will benefit from the collection of modalities recorded.

5.2. Results on Standard Emotion Corpora

We compare our models to the SVM baseline we explained in earlier sections for each modality. Tables 5-8 give emotion recognition accuracies while using unimodal (facial, vocal, physiological expressions of emotions) and multimodal DBN models (multimodal expressions of emotions).

Database	SVM Baseline	D*emoF*	3D*emoF*
Cohn Kanade	95.4 %	95.9 %	**96.3** %

Table 5. Emotion recognition using facial expressions

Database	SVM Baseline	D*emoV*	3D*emoV*
Mind Reading	90.62%	92.1 %	**92.87** %

Table 6. Emotion recognition using vocal expressions

Database	SVM Baseline	D*emoP*	3D*emoP*
DEAP	78.6%	78.8 %	**79.2** %

Table 7. Emotion recognition using physiological data

Database	SVM Baseline	D*emoFBVP*	3D*emoFBVP*
MAHNOB-HCI	52.4%	53.1 %	**54.8** %

Table 8. Emotion recognition using multimodal data

To depict generalizability, we use the Cohn Kanade, MindReading, DEAP and MAHNOB-HCI databases to evaluate respective performances. These databases are very popular and are standard datasets used by the affective research community for emotion recognition. We observe that our deep models perform better than the SVM baselines in both unimodal and multimodal scenarios.

6. Convolutional Deep Belief Model (CDBN)

In this section, we describe our multimodal CDBN model and investigate their usability to recognize subtle or low intensities of expressions of emotions. Convolutional RBMs are an extension of regular RBMs [24]. These are inspired by convolutional neural nets and rely on convolution and weight sharing. When convolutional RBMs are stacked together, they form convolutional deep belief networks [22]. Convolutional DBNs are solely generative models that are trained in a greedy layer-wise manner. Here, the input is fed into the networks and the features learned by the last layer are fed to a Support Vector Machine (SVM). In CRBMs, the network's visible layer is a matrix, instead of a vector. This enables the network to understand the spatial proximity of the pixels, leading to more robust feature learning (when compared to regular RBMs).

6.1. Results for CDBN Models

We used primary expressions of emotion of the lowest intensity from the *emoFBVP*, Cohn-Kanade, Mind Reading, DEAP and MAHNOB-HCI databases.

SVM Baseline	D*emoFBVP*	CD*emoFBVP*	CD*emoFBVP* +ROI
75.67	76.54	81.41	**83.18**

Table 9. Emotion recognition using *emoFBVP* database

SVM Baseline	D*emoF*	CD*emoF*	CD*emoF* +ROI
95.4	95.9	96.8	**97.3**

Table 10. Emotion recognition using Cohn Kanade database

SVM Baseline	D*emoV*	CD*emoV*
90.62	92.1	**93.4**

Table 11. Emotion recognition using mind reading database

SVM Baseline	D*emoP*	CD*emoP*
78.6	78.8	**79.5**

Table 12. Emotion recognition using DEAP database

SVM Baseline	D*emoFBVP*	CD*emoFBVP*	CD*emoFBVP* +ROI
52.4	53.1	57.9	**58.5**

Table 13. Emotion recognition using MAHNOB-HCI database

We applied the CD*emoFBVP* model (a model very similar to D*emoFBVP* but formed by stacking convolutional RBMs) to learn the multimodal deep features. We also extracted regions of interest (ROI) in the face (around the eyes, eyebrows and mouth area) and body images (head, hands and legs) and fed them to the deep CD*emoFBVP*+ROI model. Tables 9-13 show percentage emotion recognition accuracies on various emotion corpora. Tables 9, 10 and 13 compare performances of DBN, CDBN and CDBN+ROI models with the SVM baselines. Tables 11 and 12 compare performances of DBN and CDBN models with SVM baseline models on voice and physiological signal data (there is no ROI in voice and physiological data). Again, to depict generalizability, we show results on standard emotion datasets. We notice that our CDBN+ROI models outperform our CDBN models which in turn perform better than the DBN models and SVM baselines.

7. Conclusions and Future Work

We made three major contributions in this work. We presented the *emoFBVP* database of multimodal recordings of actors enacting various expressions of emotions. This is one of the first emotion datasets that has recordings of varying intensities of expressions of emotions in multiple modalities recorded simultaneously. We strongly believe that the affective computing community will greatly benefit from the large collection of modalities recorded. Next, we described four deep belief network D*emoDBN* models and showed that these models generate robust multimodal features for emotion classification in an unsupervised manner. Our experimental results showed that our D*emoDBN* models perform better than the state of the art methods for emotion recognition using popular emotion corpora. This validated the use of our *emoFBVP* database for multimodal emotion recognition studies. Thirdly, we showed that convolutional deep belief network (CDBN) models along with region of interest extraction learn salient multimodal features for recognition of low intensity/subtle expressions of emotions.

One of the main goals for the future is to build a real-time multimodal emotion recognition system using deep architectures. In a real-time scenario, data from one or more modalities may be absent. We hope to develop models that will continue to perform and successfully recognize emotions even when one or more modalities are absent. Our preliminary experiments revealed that the first deep layer learns to identify edges and simple shapes, the second layer identifies more complex shapes and objects (like eyes, nose, mouth etc.) and the third layer learns which shapes and objects can be used to define a facial expression. Analyzing the multimodal features learned using the deep models to better understand affect, is an interesting direction for future research.

References

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. *ACM International Conference on Multimodal Interfaces*, 2004.

[2] M. Pantic, G. Caridakis, E. Andre, J. Kim, K. Karpouzis and S. Kollias. Multimodal emotion recognition from low-level cues. *Emotion Oriented Systems*, 2011.

[3] T. Vogt and E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *IEEE International Conference on Multimedia and Expo*, 2005.

[4] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll and B. Radig. Low level fusion of audio and video feature for multi-modal emotion recognition. *International Conference on Computer Vision Theory and Applications*, 2008.

[5] G. Taylor, G. Hinton and S. Roweis. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 2007.

[6] C. Anagnostopoulos, T.Iliou and I.Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Articial Intelligence Review*, 2012.

[7] Y. Kim, H. Lee and EM. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. *Acoustics, Speech and Signal Processing*, 2013.

[8] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2009.

[9] N. Morgan. Deep and wide: Multiple layers in automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[10] G. Sivaram and H. Hermansky. Sparse multilayer perceptron for phoneme recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[11] A. Mohamed, G. Dahl and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Ng. Multimodal deep learning. *International Conference on Machine Learning*, 2011.

[13] N. Srivastava and RR. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems.*, 2012.

[14] SE. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Glehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, RC. Ferrari and M. Mirza. Combining modality specific deep neural networks for emotion recognition in video. *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

[15] A. Dhall, R. Goecke , J. Joshi, M. Wagner and T. Gedeon. Emotion recognition in the wild challenge 2013. *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

[16] H. Lee, C. Ekanadham and A. Ng. Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems*, 2008.

[17] Y. Tang and C. Eliasmith. Deep networks for robust visual recognition. *International Conference on Machine Learning*, 2010.

[18] K. Sohn, D. Jung, H. Lee and A. Hero. Efficient learning of sparse, distributed, convolutional feature representations for object recognition. *IEEE International Conference on Computer Vision*, 2011.

[19] H. Lee, R. Grosse, R. Ranganath, and A. Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 2011.

[20] A. Krizhevsky, I. Sutskever and G. Hinton. Imagenet classication with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012.

[21] P.Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1986.

[22] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 2006.

[23] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.

[24] H. Lee, R. Grosse, R. Ranganath, and A. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *International Conference on Machine Learning*, 2009.

[25] G. Desjardins and Y. Bengio. Empirical evaluation of convolutional RBMs for vision. *Technical report, Universit de Montral*, 2008.

[26] M. Norouzi, M. Ranjbar and G. Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[27] C. Busso, S. Lee and S. Narayanan. Using neutral speech models for emotional speech analysis. *Proceedings of Interspeech*, 2007.

[28] E. Mower, M. Mataric, and S. Narayanan. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech and Language Processing*, 2011.