





Deep Emotion Recognition through Upper Body Movements and Facial Expression

Chaudhary Muhammad Aqduş Ilyas^{2,3}^a, Rita Nunes¹, Kamal Nasrollahi²^b, Matthias Rehm³^c
and Thomas B. Moeslund²^d

¹Department of Electronic Systems, Aalborg University, Aalborg, Denmark

²Visual Analysis of People Lab, Aalborg University, Aalborg, Denmark

³Human-Robot Interaction Lab, Aalborg University, Aalborg, Denmark

Keywords: Emotion Recognition, Facial Expressions, Body movements, Deep Learning, Convolutional Neural Networks.

Abstract: Despite recent significant advancements in the field of human emotion recognition, applying upper body movements along with facial expressions present severe challenges in the field of human-robot interaction. This article presents a model that learns emotions through upper body movements and corresponds with facial expressions. Once this correspondence is mapped, tasks such as emotion and gesture recognition can easily be identified using facial features and movement vectors. Our method uses a deep convolution neural network trained on benchmark datasets exhibiting various emotions and corresponding body movements. Features obtained through facial movements and body motion are fused to get emotion recognition performance. We have implemented various fusion methodologies to integrate multimodal features for non-verbal emotion identification. Our system achieves 76.8% accuracy of emotion recognition through upper body movements only, surpassing 73.1% on the FABO dataset. In addition, employing multimodal compact bilinear pooling with temporal information surpassed the state-of-the-art method with an accuracy of 94.41% on the FABO dataset. This system can lead to better human-machine interaction by enabling robots to recognize emotions and body actions and react according to their emotions, thus enriching the user experience.

1 INTRODUCTION


Human emotions play a vital role in human-human and human-machine interaction. Emotions represent the instantaneous mental states, which varies according to human behavior and communication. Researchers are emphasizing automatic recognition of human emotions as it is one of the essential parameters for natural human-machine interaction.


In human-machine interaction, the interaction would be impaired if machines cannot recognize or understand human emotions. Similar applies to human-human interaction if the other party fails to understand these body expressions.


If machines can react to our moods, that would enable smart homes or centers to adjust lighting, music, and temperature accordingly. It would also help


medical doctors and physiologists automatically identify the symptoms of hypertension, depression, and other behavioral disorders, enabling them to have early preparations for such conditions. This skill can enable sociable robotics to assist people in simple tasks such as delivering meals or vacuuming the house. Humanoid robots that provide services to people, the human-robot interaction would greatly improve if these robots could adjust their reactions to the current emotional state of a person Augello et al. (2018); Kim et al. (2011); Sorbello et al. (2014). Generally, it would enable machines to respond, not limited to direct commands but with the ability to adjust their reactions to have natural and human-like, human-machine interaction. However, these interactions are minimal and could be improved if the robot had more knowledge about the person they need to interact with Breazeal (2003).

Human recognize and demonstrate emotions through multi-modalities such as through facial expressions Nguyen et al. (2018); Barros et al. (2015); Ilyas et al. (2018b); Mano et al. (2016), body move-

^a  <https://orcid.org/0000-0002-0766-3531>

^b  <https://orcid.org/0000-0002-1953-0429>

^c  <https://orcid.org/0000-0001-6264-1688>

^d  <https://orcid.org/0000-0001-7584-5209>

ments Nguyen et al. (2018); Lang et al. (2015); Barros et al. (2015); de Gelder et al. (2015), speech recognition Bänziger et al. (2009) and physiological signals Agrafioti et al. (2011); Jerritta et al. (2011); Kim and André (2008); Martínez-Rodrigo et al. (2015); Picard et al. (2001). Existing methods for identifying these body expressions are heavily relying on audio-visual cues Bänziger et al. (2009) and wearable sensors such as ECG monitors Agrafioti et al. (2011); Jerritta et al. (2011); Kim and André (2008); Martínez-Rodrigo et al. (2015); Picard et al. (2001). Audio-visual fusion have achieved remarkable results with accuracy of approximately 99% Noroozi et al. (2018). However, these approaches have their limitations as audio-visual sensors cannot extract inner affection Ekman et al. (2013); Picard et al. (2001). For instance, a person can be happy or sad without smiling and crying and vice versa. In addition, people vary greatly in the expression of their emotions. Detection of signals through physiological sensors correlate heartbeat, blood pressure, and others signals with happiness, anger, surprise, and others. This approach is more suitable for identifying inner feelings as it provides information about heart rhythm interaction with the brain system. However, the wearable body sensors cause inconvenience, so it is not suitable for emotion detection in everyday practices.

Research has also shown that body language comprises a significant amount of the affective information Lang et al. (2015); de Gelder et al. (2015). According to Mehrabian Mehrabian et al. (1971), only 7% of human communication is conveyed through words, 38% through vocal tone, and 55% through non-verbal elements such as facial expression, body language, and gestures. Body posture, gestures, eye movement, hand and head movement, touch, or even personal space represent the body language Mehrabian et al. (1971). Many studies have proved theoretically and empirically the benefit of incorporating various modalities in the perception of human emotions compared to using a single method Picard et al. (2001); Soleymani et al. (2011). Complex human emotions can be fully-implied by integrating significant features from multiple modalities (e.g., facial features and body gestures).

In our research article, we have tried to explore the effectiveness of facial expressions and body gestures to recognize emotions. For this purpose, we have followed two approaches; in the first technique, Convolutional Neural Networks (CNNs) classify emotions without considering temporal features. In this system, single images are used, thus classifying each frame in videos in real-time. In the second approach, following Barros et al. (2015); Sun et al. (2018), we have used

the temporal information to classify the emotions, but a contrast to Barros et al. (2015); Sun et al. (2018) full images are fed to the Long Short Term Memory (LSTM) network to exploit the temporal information. Besides, this article will explore various fusion techniques like product fusion method (PFM), average fusion method (AFM), and Multimodal Compact Bilinear Pooling (MCB) Fukui et al. (2016) fusion and discuss their performances. Even though deep learning approaches show improved accuracy compared to conventional approaches, they are still more computationally demanding. Hence, this work will explore a solution with less computational requirements.

The paper organization is as follows: The following section 2 will discuss the related research. Section 3 presents the proposed model, illustrating how it deals with a frame-based and sequence-based recognition of the emotions with different fusion techniques. This section also discusses the evaluation of the parameters to decrease the computation power. Section 4 describes the experimental results and compares them with state-of-the-art methods. The conclusion and discussion of the experimental results with future work are presented in section 5.

2 RELATED RESEARCH

Emotion recognition through non-verbal modalities like facial expressions and body gestures is viewed as one of the most effective cues Ekman et al. (2013). Therefore, many researchers have explored the fusion of visual-modalities for improved affect understanding Picard et al. (2001); Gunes and Piccardi (2006); Chen et al. (2013); Agrafioti et al. (2011); Shan et al. (2007); Karpouzis et al. (2007). They illustrate that facial expressions and body gestures augment each other in understanding emotional states in activities of daily living (ADL) and social robot interactions. Researchers Gunes and Piccardi (2007) and Shan et al. (2007) have analyzed the facial features with body gestures, particularly upper limbs and head movements, for emotion recognition. Former has utilized facial action units (AU) and performed classification with Bayes Net with early and late fusion whereas later has employed Spatio-temporal features classification with SVM along with Canonical Correlation Analysis (CCA) at the decision level. In recent years, researchers have focused on deep learning approaches to solve this issue, which has achieved the best recognition rates.

Gunes and Piccardi Gunes and Piccardi (2008) presents the performance of facial expressions, body movements, and their fused representation for auto-

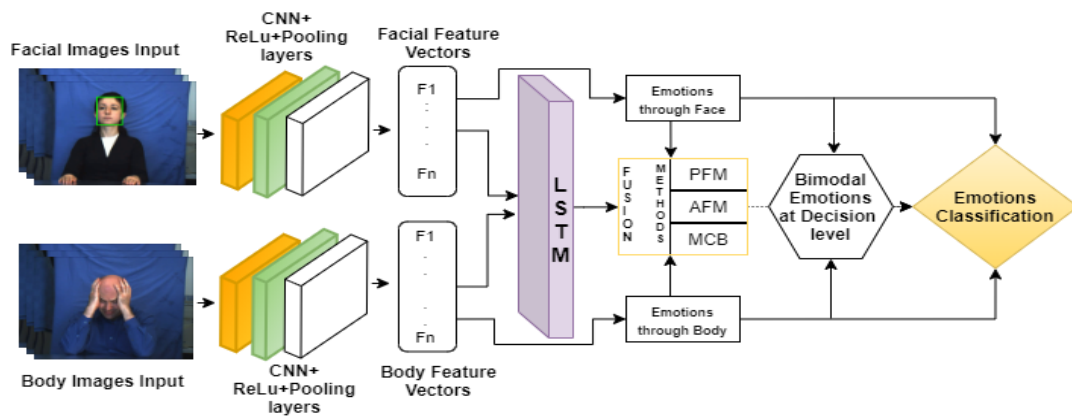


Figure 1: Bimodal Emotion Classification Model through Facial Expressions and Upper Body Movements.

matic recognition of emotions. They extract the facial and bodily features separately and compare their performance accuracy. For facial expressions, they localize face and track landmarks and then extract features. Similarly, for body motion analysis, they track hand, shoulder, and head movements and then extract a series of features with different features representation techniques. Both facial and bodily features are classified with Support Vector Machines (SVM) and Random Forests. In the last step, features are fused to recognize emotions. Studies exhibit better performance of system with feature fusion techniques Chen et al. (2013) Gunes and Piccardi (2008).

Recent work of Barros et al. (2015) and Sun et al. (2018) used Convolutional Neural Networks (CNN) to recognize emotion from both face and body movements. Both studies incorporated temporal features into their classification, which forces them to analyze an entire video before it can be classified.

2.1 Facial Expression Recognition

Facial expressions are one of the vital source to know about mood and feelings in an interpersonal communication. Therefore, researchers focus to analyze facial expressions through traditional machine leaning and advanced deep learning algorithms. Khorrani et al. (2016) used a deep learning approach to merge CNNs with RNNs and evaluate how each portion of the neural network contributed to the overall success of the emotion recognition system. For training, two different architectures were used, the first with a single CNN frame and the second with a combination of CNN and RNN. Although CNN learns valuable features from the video data from the single frame regression, it disregards temporal information. This knowledge can be implemented through the use of RNN. Results determined that the CNN+RNN model translates to more accurate predictions.

2.2 Emotion Recognition through Body Movements

Upper body movement, such as hand and head movement, conveys vital information related to emotional states. For instance, when a person displays a neutral emotion, they generally do not move their arms; however, when they are happy or sad, the body tends to be extended, and the hands move upwards closer to the head Noroozi et al. (2018). However, this information is subjective, dependant upon the personal attitude to the circumstances and cultural bias. Research presented by Piana et al. (2014) suggest real-time emotion recognition through body movements and gestures. The features are extracted from 3D motion clips containing full-body movements, recorded using two systems: a professional optical motion capture system and Microsoft Kinect. The body joints are tracked, and feature vectors of the movements are extracted and used for classification using a linear SVM classifier. The emotions tested were the six standard emotions. Human validation demonstrated that three emotions were easily recognized from body movements (happiness, sadness, and anger), while the others (surprise, disgust, and fear) were confused with each other. Because of that, a sub-problem with only four emotions (happiness, sadness, anger, and fear) was formulated. The approach showed better results when only classifying four out of the six emotions.

Glowinski et al. (2011) took a different approach, which analyzed affective behavior solely based on upper-body movements. A range of twelve different emotions was classified according to their valence and arousal. Features were extracted from two videos, one that displayed a frontal view of the subjects and another that displayed a lateral view. The trajectories of the head and hands were tracked, and low-level physical measures, i.e., position, speed, acceleration,

were extracted. Higher-level expressive and dynamic features include smoothness and continuity of movement, spatial symmetry of the hands, gesture duration were then computed, forming a 25-features vector. PCA was later applied to reduce the dimensionality of the data. Furthermore, clustering was used to classify the data into four clusters according to the categorical variables, i.e., valence (positive, negative) and arousal (high, low). The framework was tested on the GEMEP (GENeva Multimodal Emotion Portrayals) dataset Bänziger et al. (2012). The results demonstrate that gestures can be effectively used to detect human emotional expression.

Research proposed by Barros et al. (2015) model these upper body movements for emotional classification by using FABO Gunes and Piccardi (2006) dataset. They present the motion by an additional layer on the network that tracks the frame-wise difference in each sequence. This representation involves the structure and information of the gesture/motion with the aid of weighted shadows. Another study by the same authors Barros et al. (2014) extracts spatial and temporal features of gesture sequences through Deep Neural Networks (DNN) to generate a motion representation. Moreover, a Multichannel Convolutional Neural Network (MCCNN) is used to learn and extract features from the previously generated motion representation and uses such features to classify different gestures.

We have trained our system with full frames of the FABO dataset with facial and body gestures features to capture the gesture information. The networks extract the spatial and temporal information using CNN and LSTM and finally classify the emotions.

2.3 Bi-modal Emotion Recognition

Fusion of multiple modalities can achieve better recognition performance than single modality. Nevertheless, a good fusion strategy must be applied; otherwise, the fusion of modalities can hurt the accuracy of the recognition system.

Gunes and Piccardi studied this case precisely and conducted experiments where only single modalities were tested (facial expression or body gestures) and where both modalities were fused to formulate a detection Gunes and Piccardi (2006, 2007). The results revealed that the bimodal approach had better performance.

The bi-modal approach is also considered by Barros et al. (2015) to recognize emotions by taking into account facial expression and body movements. They used neural networks on their solution and achieved much higher average accuracies on fus-

ing both modalities than testing each modality separately, going from $57.84 \pm 7.7\%$ on body motion and $72.70 \pm 3.1\%$ on facial expression, to $91.30 \pm 2.7\%$ average accuracy on bimodal emotion recognition.

The research proposed by Nguyen et al. (2018) fused the audio-visual, face, and body modalities using the compact bilinear pooling (MCB) method and demonstrated the state-of-the-art results. In this article, besides general feature-level fusion techniques such as average fusion, product fusion, we have also explored the bilinear pooling technique for face and body fusion.

3 PROPOSED SYSTEM

Our model is comprised of Convolution Neural Networks (CNNs) to extract facial features and bodily features, with linear addition of Long Short Term Memory (LSTM) model to use the sequential information. Each modality (face or body) is trained with CNN in frame-based and sequence-wise to generate the emotional states. It is challenging to determine the multimodal fusion efficiency for emotion recognition accuracy, so we have proposed three different fusion techniques to identify the best approach. The network structure remains the same for all considered fusion modalities. Overview of the system is illustrated in Fig 1

3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) performs remarkably good at acquiring spatial information. Each CNN layer operates twofold; filtering through the convolution layer and max-pooling to avoid losing useful information. Generally, CNNs are composed of convolutional layers, and fully connected layers extract features. Most of the parameters are also present in the fully connected layers responsible for most of the computation power. For instance, fully connected layers of VGG16 contains 90% of all the parameters. The VGG16 is a deep convolutional network with up to sixteen layers (thirteen convolutional layers and three fully connected layers). Inception V3 reduces the parameters by the introduction of global average pooling Szegedy et al. (2016). Similarly, Xception Chollet (2016) takes advantage of the use of residual modules and depth-wise separable convolutions.

To lessen the computation cost, we have implemented CNN architecture as proposed by Arriaga et al. (2017). It is a simple architecture that achieves almost state-of-the-art performance classifying emotions. The architecture classifies emotions based on



Figure 2: Selection of the facial region on a frame from the FABO dataset and scaling into 48*48 to form image database.

facial expressions and faces according to gender. On the contrary to Arriaga et al. (2017), only relying on the FER-2013 dataset, we use the FABO dataset for training purposes. Besides, our CNN architecture contains four residual deep separable convolutional layers, where batch normalization and ReLU activation function accompany each convolutional layer. Batch normalization normalizes the activation of the previous layer at each batch. Residual modules modify the desired mapping between two subsequent layers by connecting the output of previous layers to the output of new layers. Depth-wise separable convolutions reduce further the number of needed parameters. They are composed of depth-wise convolutions and point-wise convolutions. Instead of the fully connected layers, this architecture uses Global Average Pooling that reduces each feature map to a scalar by calculating the average of all the elements in the feature map. The last convolution layer has the same number of feature maps as the number of classes. In the end, a softmax activation function is applied to produce a prediction.

3.2 Long Short Term Memory Networks (LSTMs)

LSTMs are a recurrent neural network that processes and absorb sequential information. According to Ilyas et al. (2018a):

”The LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates control the flow of information through the model by using point-wise multiplications and sigmoid functions σ , which bound the information flow between zero and one.”

To take advantage of spatial and temporal information, we have linearly combined the CNN and LSTM model, where CNN extracts the features and then sequentially feeds into the LSTM network. Such a combination works well in the case of video data, as exhibited by Yao et al. (2015); Fukui et al. (2016).

3.3 Fusion Methods

One of the issues in multimodal emotion recognition is deciding when to combine the information. There are a few different techniques to fuse the emotion recognition results of different modalities with certain advantages and disadvantages. Some of the most explored techniques are early (feature-level) fusion and late (decision-level) fusion. Recent studies explore another feature-based fusion method called bilinear pooling fusion.

Feature-level Fusion. Feature-level fusion combines the data from both modalities before classification. A single classifier is used containing features from both modalities. One of the biggest drawbacks of feature-level fusion is high-dimensional feature production resulting in more parameters and more computation power consumption. To reduce the dimensions, we have applied the compact bilinear pooling (MCB) as proposed by Lin et al. (2017). Bilinear pooling multiplies two vectors that produce tons of parameters, and it is costly. However, compact bilinear pooling reduces the dimensions with the same information level but with very few parameters.

Compact Bilinear Pooling Fusion. Lin et al. (2017) proposes a compact bilinear pooling technique for fine-grained visual recognition. In this technique, outer product \otimes is calculated by element wise multiplication of two input feature vectors $f_1 \in V^{n_1}$ and $f_2 \in V^{n_2}$ and scaling it into a matrix $[]$ to reduce dimensions. For instance $y = X[f_1 \otimes f_2]$, where X is a learned model, \otimes denotes the outer product and $[]$ represents linearizing the matrix in a vector. This technique has produced better results for multimodal emotion recognition task as mentioned by Nguyen et al. (2018), who fused audio-visual, face and body features to recognize emotions by considering correlation among them.

Decision-level Fusion. Decision level fusion does not produce high-dimensional features as each modality is trained and classified separately to fuse recognition accuracy at the end. However, this method fails to understand the correlation between input modalities.

This co-relation is more important and meaningful in human-machine interactions where body movements and facial expressions complement each other Barros et al. (2015).

4 EXPERIMENTAL RESULTS

In this section, we first describe the databases involved and their training protocols. Then we demonstrate the results.

4.1 Benchmark Datasets

We have used the bi-modal face and body FABO dataset and the FER-2013 dataset for the emotion recognition task. Details of the datasets are mentioned as follows:

FABO-dataset. The bi-modal face and body data set is presented by Gunes and Piccardi (2006). Two cameras acquire the database for monitoring face and body movements, that captures the facial data and upper body movements separately. The videos provide annotations on the stages of the affective states, therefore splitting the demonstration of each emotion into neutral, onset, apex, and offset phases. Annotation is performed for 16 subjects out of the 23 subjects for emotional classification. The face and body posture tend to shift in the onset process, and these changes reach a steady level at the apex phase. Finally, expressions and movements exhibit relaxation at the offset stage. However, these phase annotations are only done for twelve of the subjects.

Frames in the apex phase are considered for CNN training since they are the ones that reflect the emotions best. Two apex phases are assessed from the annotated videos. The dataset contains 1410 images for anger, 458 for disgust, 343 for fear, 613 for happiness, 570 for sadness, and 588 for a surprise, split into test and training. The neutral emotion is the exception, the images for this emotion were obtained from the neutral phase from each video, amounting to 786 images. The selected images display the upper body of the subjects; therefore, a facial recognition algorithm was applied to extract only the facial region within the image. The method used was a DNN face detector module included in OpenCV 3.6. The selected frames were grayscaled and resized to the FER-2013 dataset size, which is 48*48 pixels as demonstrated in the figure 2

FER-2013 Dataset. The FER-2013 database consists of approximately 36,000 images, labeled with seven emotion classes (six Ekman emotional states plus neutral expression). FER-2013 is one of the

biggest databases for FER in-the-wild environment but with a low image resolution of 48 * 48 pixels leading to problems for facial landmark detectors. The dataset contains 35887 annotated images, with 4953 anger images, 547 disgust, 5121 fear, 8989 happiness, 6077 sadness, 4002 surprise, and 6198 neutral images. Some samples of the images are shown in Fig. 3.

Experiments are performed to detect emotions from face and upper body separately and fused accuracy is also calculated. Details are provided in the section 3.3.

4.2 Network Training

The CNN architecture is trained with benchmark datasets FER-2013 and FABO datasets to extract the facial and body features and evaluate the effectiveness of each modality.

Face-CNN Model: (For Facial Emotional Recognition). Only facial features are trained to the network to evaluate facial expressions. To recognize emotions, first face is localized, tracked, and then face cropping is applied according to the network input parameters. CNN is trained with the FER-2013 dataset, with data augmentation techniques applied to train with more diverse data. It also helps to prevent overfitting and to generalize the model.

Early stopping is used to avoid overfitting. It stops the training process of the model when the error on the validation set gets higher than before. The learning rate is reduced when validation loss has stopped improving.

The CNN was trained using Adam optimizer. This optimization algorithm is an extension of the stochastic gradient descent. It has some benefits compared to other algorithms, such as less memory requirement, computationally efficient, and it is well suited for problems with extensive data and parameters Kingma and Ba (2014). The trained model that we called the face-CNN model achieved 65% accuracy in the validation set. To recognize the emotions from upper



Figure 3: FER-2013 sample images for facial emotion recognition.

body movements, we have trained the CNN model with body features of the FABO dataset.

Body-CNN Model: (For Upper Body Emotion Recognition) Only the FABO dataset is used to train the body-CNN model. This dataset is already described in Section 4.1. However, to train this model, full image frames of the FABO dataset with facial and body gestures information are used, as illustrated in Fig. 2.

CNN-body architecture is the same as CNN-face, but the images are descaled from their original 1024x768 dimensions to 128x96 and grayscaled to ease and fasten the training process. The pixel values were normalized to a range between -1 and 1. Data augmentation strategies are also applied to increase the number of training samples. Furthermore, the data was split into train and validation data with an 80/20 ratio. The model achieved a 96% accuracy in the validation set.

4.3 Bi-modal Emotion Recognition

As mentioned in section 2, the fusion of different modalities is capable of achieving more significant results than single modalities for emotion recognition. To identify which fusion technique works best in our task, we have applied MCB fusion at the feature-level, whereas product and average fusion strategies are applied at the decision level.

4.4 CNN Architecture

For our experimentation, we have used the same architecture as proposed by Arriaga et al. (2017), as described in detail in section 3.1. However, our implementation varies with Arriaga et al. (2017) as we have used different datasets for training with the addition of the LSTM model to exploit the temporal information. We aim to reduce the number of CNN parameters and computational costs and achieve better generalization. The network is composed of 4 convolutional layers and ReLu and batch normalization at each layer. As mentioned in section 3.1 instead of fully connected layers, global average pooling is applied. However, in the case of a temporal database LSTM model is installed, followed by the softmax.

4.5 Performance Analysis

4.5.1 Frame-based Emotion Recognition

The performance of each modality is tested with our trained network for emotion recognition. Various

Table 1: Results of different evaluation metrics for each frame-based emotion recognition method.

Evaluation Matrix	Facial Expressions	Upper Body Movement	Bimodal Average Fusion	Bimodal Product Fusion	Bimodal Bilinear Pooling
Precision	77.2 %	72.8 %	81.7 %	82.6 %	83.7 %
Recall	73.0 %	72.7 %	80.4 %	80.9 %	81.5 %
F1-Score	72.8 %	71.5 %	80.3 %	80.9 %	82.5 %
Accuracy	77.7 %	76.8 %	85.7 %	86.6 %	87.2 %

classification models with different evaluation metrics analyze which modality has better performed. We have analyzed the system performance with precision, recall, accuracy, and F1-score metrics, as illustrated in Table 1. Moreover, bi-modal fusion with different fusion methods is applied to identify the performance of fusion strategies.

Facial Expression Analysis. The normalized confusion matrix in Fig. 4 shows the percentage of image samples that are of a specific emotion (true label) and that are classified as corresponding to a specific emotion (predicted label). The confusion matrix shows the best classification results corresponding to the anger and neutral emotions with 94% and 90%, respectively. The worst classification result corresponds to the surprise emotion that is often mistaken with fear; however, fear rarely is misclassified as a surprise.

Upper Body Movement Analysis. Fig. 5 displays the normalized confusion matrix for the upper body movements emotion recognition. This confusion matrix shows the true positive rate; hence, the percentage of samples from each dataset that are classified correctly.

From Fig. 5 it is possible to observe that, once again, the best recognition results are attributed to anger and neutral emotions. Comparatively to the facial expression recognition, in this recognition modality, the surprise emotion has a much better recognition rate and is less often mistaken by fear. Also, the sadness emotion is quite often misclassified as a surprise.

Bi-modal Analysis. Two different decision-level fusion methods are tested, an average method and the product-method. In the average method, the average is calculated between both modalities and for each of the emotions. In the product method, the product of the probabilities of each modality is calculated for each of the emotions.

Finally, the combination of both modalities produces the results in Fig. 6 using the average fusion method, and Fig. 7 using the product fusion method.

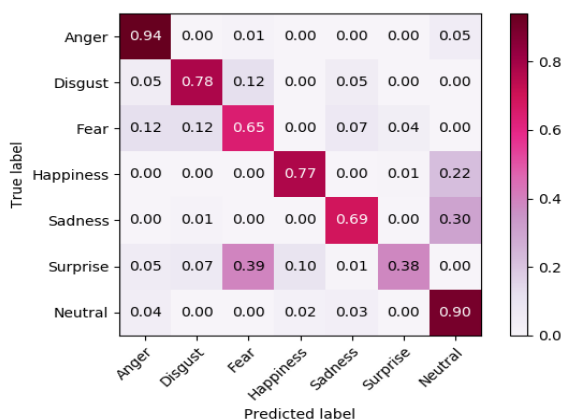


Figure 4: Normalized confusion matrix for facial expression recognition.

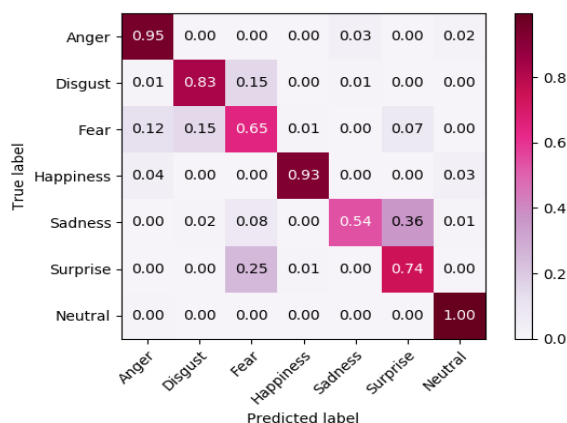


Figure 6: Normalized confusion matrix for bimodal emotion recognition using the average fusion method.

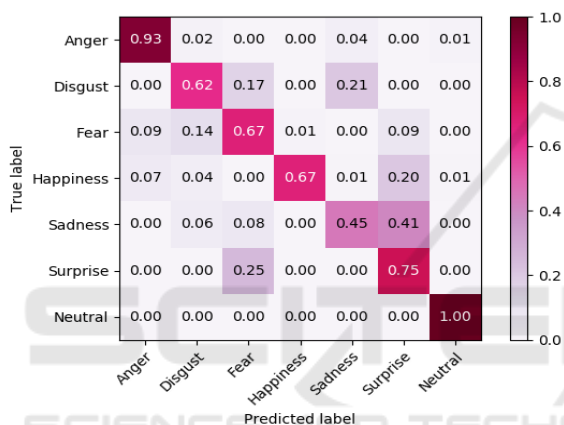


Figure 5: Normalized confusion matrix for upper body movements emotion recognition.

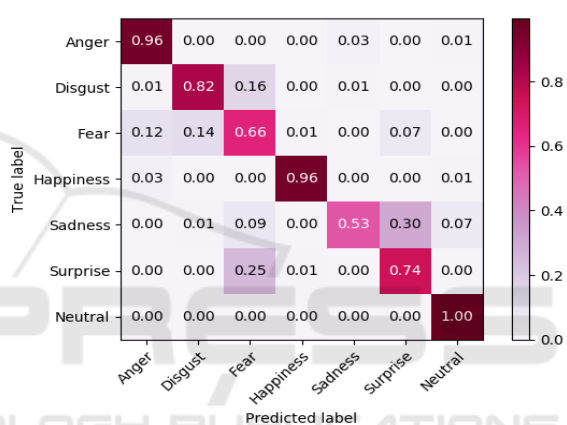


Figure 7: Normalized confusion matrix for bimodal emotion recognition using the product fusion method.

4.5.2 Sequence-based Emotion Recognition

To train the system with spatial and temporal information of the FABO dataset, we have to use the face-CNN model pre-trained on the FER-2013 dataset. As the FABO dataset poses video data with emotional annotation for 16 subjects out of 23. Each video displays the same emotion from two to four times, so we have divided each video into neutral, onset, apex, and offset maximum phase length of five seconds. We trained this network with these video data and obtained the features, as we have used the full frames of the FABO dataset containing facial and gesture features. These features are feed into LSTM in a timely manner to evaluate the sequential information for emotional classification. Details of recognition accuracy is presented in the Fig. 8.

Facial Expression Analysis. Our network achieved an average accuracy of 93.213 % when it is trained to 80 epochs. It is observed that system performance reached its maximum accuracy when epochs range

from 40 to 50; after that system, performance did not fluctuate considerably.

Upper Body Movement Analysis. It is observed that temporal information contributed to accuracy efficiency when the network is trained with full FABO dataset frames. Our network achieved an accuracy of up to 79.27 % for emotional recognition through upper body movement analysis.

Bi-modal Analysis. When the network is trained with combined facial and upper body features, our system has surpassed the state-of-art accuracy to 94.418 %. In this experiment, network parameters are less than the state-of-art method Nguyen et al. (2018), and it is robust to work in real-time scenarios.

4.6 Parameters Evaluation

Our network contains four deep-separable convolutional neural layers with ReLu and batch normalization function. We have used global average pooling and softmax for emotion classification that contribute

Table 2: Performance analysis of our system with CNN and (CNN + LSTM) models and their comparison with state-of-art methods. We have performed 3-fold cross validation after splitting data into 80/20 protocols.

Method / Modality	Facial Expressions	Upper Body Movement	Bimodal Fusion
Gunes and Piccardi (2008)	35.2 %	73.1 %	82.7 %
Chen et al. (2013)	66.5 %	66.7 %	75.0 %
Barros et al. (2015)	72.7	57.8	91.3
Barros and Wermter (2016)	87.3	74.8	93.65
Our Frame based Model	77.7	76.8	87.2
Our sequence Based Model	90.42	79.27	94.41

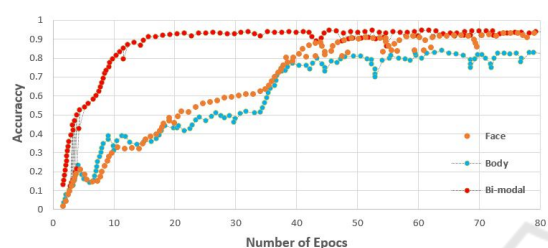


Figure 8: Performance of the combined CNN + LSTM neural network model on test data of FABO dataset.

to approximately 600,000 parameters. We trained this network with the FER-2013 database that provides an accuracy of 65% on the validation set. However, usage of the shelf CNN network that is 70 times more parametric heavy and provides 71.3% accuracy on the FER-2013 dataset. In contrast, when we have employed this network to recognize emotions from face and body, it surprised the state-of-art method when we have a bi-modal model with temporal information. Additionally, this model also showed improved accuracy in using a single modality such as the upper body movements. We acquired an accuracy of 76.8 % and 79.27 % with spatial and temporal information, respectively. Application of compact bilinear pooling (MCB) contributed to dimensionality reduction without compromising on the performance.

5 DISCUSSION AND CONCLUSION

The major problem in developing a human-machine affective system is the integration of multimodal sensory information. In this research article, we have explored the spatial-temporal technique for emotion analysis of visual modalities. We have also studied different fusion techniques with lesser computation cost. We have developed a robust architecture to identify emotions from the face and upper body move-

ments to use in real-time human-machine interaction systems.

It is illustrated through the confusion matrices that the bimodal approach shows better results than the monomodal approaches, regardless of the fusion method. In this case, the best recognition rates correspond to both fusion methods for anger, happiness, and neutral emotions. The worst recognition rate is attributed to the sadness emotion that is often misclassified as a surprise emotion.

All the evaluation metrics being considered to have greater values with this approach. Accuracy shows a significant improvement from 77,7% and 76,8% on the facial and upper body movements emotion recognition, respectively, to 85,7% and 86,6% on the fusion of both modalities.

Furthermore, the product fusion method shows slightly better results on all the evaluation metrics than the average fusion method. However, the MCB method surpassed decision level recognition accuracy. It shows that inter modalities relationship towards emotion identification is an essential factor to consider. We have demonstrated that spatial-temporal information is better classified for anger, happy and neutral emotions for further analysis. With upper body movement alone, state of the art methods found it challenging to classify the emotions accurately. However, our system has performed better with the rest of the methods, as illustrated in Table 2.

REFERENCES

- Agrafioti, F., Hatzinakos, D., and Anderson, A. K. (2011). Ecg pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115.
- Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.
- Augello, A., Dignum, F., Gentile, M., Infantino, I., Maniscalco, U., Pilato, G., and Vella, F. (2018). A social practice oriented signs detection for human-humanoid interaction. *Biologically inspired cognitive architectures*, 25:8–16.
- Bänziger, T., Grandjean, D., and Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert). *Emotion*, 9(5):691.
- Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161.
- Barros, P., Jirak, D., Weber, C., and Wermter, S. (2015). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, 72:140–151.

- Barros, P., Parisi, G. I., Jirak, D., and Wermter, S. (2014). Real-time gesture recognition using a humanoid robot with a deep neural architecture. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 646–651. IEEE.
- Barros, P. and Wermter, S. (2016). Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior*, 24(5):373–396.
- Breazeal, C. (2003). Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155.
- Chen, S., Tian, Y., Liu, Q., and Metaxas, D. N. (2013). Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, 31(2):175–185.
- Chollet, F. (2016). Xception: deep learning with depthwise separable convolutions. corr abs/1610.02357 (2016). *arXiv preprint arXiv:1610.02357*.
- de Gelder, B., De Borst, A., and Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):149–158.
- Ekman, P., Friesen, W. V., and Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier.
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., and Scherer, K. (2011). Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, 2(2):106–118.
- Gunes, H. and Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE.
- Gunes, H. and Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345.
- Gunes, H. and Piccardi, M. (2008). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):64–84.
- Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., and Moeslund, T. B. (2018a). Facial expression recognition for traumatic brain injured patients. In *VISIGRAPP (4: VISAPP)*, pages 522–530.
- Ilyas, C. M. A., Nasrollahi, K., Rehm, M., and Moeslund, T. B. (2018b). Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2291–2295. IEEE.
- Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 410–415. IEEE.
- Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaoui, A., Malatesta, L., and Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artificial intelligence for human computing*, pages 91–112. Springer.
- Khorrani, P., Le Paine, T., Brady, K., Dagli, C., and Huang, T. S. (2016). How deep neural networks can improve emotion recognition on video data. In *2016 IEEE international conference on image processing (ICIP)*, pages 619–623. IEEE.
- Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083.
- Kim, K., Cha, Y.-S., Park, J.-M., Lee, J.-Y., and You, B.-J. (2011). Providing services using network-based humanoids in a home environment. *IEEE Transactions on Consumer Electronics*, 57(4):1628–1636.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lang, K., Dapelo, M. M., Khondoker, M., Morris, R., Surguladze, S., Treasure, J., and Tchaturia, K. (2015). Exploring emotion recognition in adults and adolescents with anorexia nervosa using a body motion paradigm. *European Eating Disorders Review*, 23(4):262–268.
- Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2017). Bilinear cnns for fine-grained visual recognition. In *Transactions of Pattern Analysis and Machine Intelligence (PAMI)*.
- Mano, L. Y., Façal, B. S., Nakamura, L. H., Gomes, P. H., Libralon, G. L., Meneguete, R. I., Geraldo Filho, P., Giancristofaro, G. T., Pessin, G., Krishnamachari, B., et al. (2016). Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition. *Computer Communications*, 89:178–190.
- Martínez-Rodrigo, A., Zangróniz, R., Pastor, J. M., Latorre, J. M., and Fernández-Caballero, A. (2015). Emotion detection in ageing adults from physiological sensors. In *Ambient Intelligence-Software and Applications*, pages 253–261. Springer.
- Mehrabian, A. et al. (1971). *Silent messages*, volume 8. Wadsworth Belmont, CA.
- Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., and Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding*, 174:33–42.
- Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE transactions on affective computing*.
- Piana, S., Stagliano, A., Odone, F., Verri, A., and Camurri, A. (2014). Real-time automatic emotion recognition from body gestures. *arXiv preprint arXiv:1402.5047*.
- Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective

- physiological state. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1175–1191.
- Shan, C., Gong, S., and McOwan, P. W. (2007). Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, pages 1–10.
- Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.
- Sorbello, R., Chella, A., Calí, C., Giardina, M., Nishio, S., and Ishiguro, H. (2014). Telenoid android robot as an embodied perceptual social regulation medium engaging natural human–humanoid interaction. *Robotics and Autonomous Systems*, 62(9):1329–1341.
- Sun, B., Cao, S., He, J., and Yu, L. (2018). Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks*, 105:36–51.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Yao, A., Shao, J., Ma, N., and 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*, pages 451–458. ACM.

