

Semantic-based visual emotion recognition in videos-a transfer learning approach

Vaijyanthi Sekar, Arunnehru Jawaharlalnehru

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India

Article Info

Article history:

Received Mar 30, 2020

Revised Dec 18, 2021

Accepted Jan 25, 2022

Keywords:

AlexNet

Convolutional neural network

Dense optical flow

Human motion analysis

Transfer learning

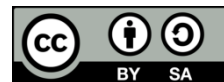
VGG-16

VGG-19

ABSTRACT

Automatic emotion recognition is active research in analyzing human's emotional state over the past decades. It is still a challenging task in computer vision and artificial intelligence due to its high intra-class variation. The main advantage of emotion recognition is that a person's emotion can be recognized even if he is extreme away from the surveillance monitoring since the camera is far away from the human; it is challenging to identify the emotion with facial expression alone. This scenario works better by adding visual body clues (facial actions, hand posture, body gestures). The body posture can powerfully convey the emotional state of a person in this scenario. This paper analyses the frontal view of human body movements, visual expressions, and body gestures to identify the various emotions. Initially, we extract the motion information of the body gesture using dense optical flow models. Later the high-level motion feature frames are transferred to the pre-trained convolutional neural network (CNN) models to recognize the 17 various emotions in Geneva multimodal emotion portrayals (GEMEP) dataset. In the experimental results, AlexNet exhibits the architecture's effectiveness with an overall accuracy rate of 96.63% for the GEMEP dataset is better than raw frames and 94% for visual geometry group-19 VGG-19, and 93.35% for VGG-16 respectively. This shows that the dense optical flow method performs well using transfer learning for recognizing emotions.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Vaijyanthi Sekar

Department of Computer Science and Engineering, SRM Institute of Science and Technology

No. 1, Jawaharlal Nehru Road, Vadapalani, Chennai, Tamilnadu, India

Email: vaijyanthisekar@gmail.com

1. INTRODUCTION

The research perspective lies in a sound understanding of human emotions; emotion recognition offers various opportunities for different occurrences. Human-computer interfaces provide better friendly communication by increasing emotional intelligence and reducing computational complexity [1]. Nowadays, advanced progress in applying computing technologies [2]–[7] has significant progress in artificial intelligence. A person converses both in verbal form and non-verbal form of communication with others. Sharing of information in the form of sound, words, language and speech denote verbal communication [8]–[10]. Mannerism speaks louder than vocal speech. In contrast, sharing information in the form of facial expressions, visual postures and emotions through body language (Kinesics) is called non-verbal communication [11], [12].

Gestures are the most compelling way to communicate as they understand extreme movements of the head and other parts of the body by which it is easy to realize the whole spectrum of emotions and feelings [13]–[15]. The qualities of movement have fewer examples by relating to specific emotional forms

such as joy, surprise, cheerfulness, anger, and fear that exhibit a contraction in the body. Here movement of forearms and openness may bring joy; sadness and fear make the human body turn away [16]. The non-verbal component has a wide range of applications in healthcare systems [17], transfer learning [18], human-robot interaction [19], and other security-related to automatic recognition of emotions using body language [20].

Many deep learning models have recently emerged based on facial recognition and gained advantages over traditional emotion recognition techniques [21]. Still, this scenario lacks in recognizing the facial micro-expressions with high robustness. Convolutional neural network (CNN) has a wide area of applications in recognizing emotions from visual body postures in raw images and video sequences. It helps in extracting the high-level dynamic features from the optical flow sequence. The primary motivation of this research lies in crime prevention, the customer-based review response by recognizing the visual body gestures. Since the camera is far from the human in surveillance monitoring, it is challenging to identify the facial emotional expression alone with high robustness. This scenario works better by adding visual body clues (facial actions, hand posture, body gestures) [22], [23]. This paper mainly focuses on identifying the visual posture of human emotions in video sequences by optical flow features. These dense optical visualizations encode the magnitude and directional information by mapping the image to RGB color space. Finally, the motion feature frames are passed to transfer learning-based architectures to classify the emotions.

The paper is structured as follows: section 2 describes the methodology of the proposed work in detail. Section 3 provides a detailed sketch of transfer-learning based architectures. Section 4 reports the experimental results and the performance metrics of transfer learning-based dense optical flow (DOF). Section 5 discusses the comparative study, and finally, section 6 concludes the work with future studies.

2. PROPOSED VISUAL GESTURE RECOGNITION USING DENSE OPTICAL FLOW

Optical flow [24] is a handcrafted feature involved in detecting the motion of moving objects in a video sequence. This algorithm involves the Gradient-based method for varying displacement in a motion introduced by Gibson [25]. At different time intervals, the motion between any two frames, say t and $t + \Delta t$ is given by its brightness constancy. The Lucas-Kanade flow algorithm [26] estimates the global displacement values in the facial expression feature points between two consecutive frames. The obtained global displacement vectors help identify different actor's facial emotions from the neutral state to the apex state. Consider $(I_1 \dots I_N)$ as the spatial image positions of the facial feature point I at frames 1, ..., N . Hence the global displacement vector of V_I of point I is:

$$\vec{v}_I = \sum_{x=1}^{N-1} \vec{v}_{Ix} = \sum_{x=1}^{N-1} (I_{x+1} - I_x) \quad (1)$$

The module $|\vec{v}_I|$ in pixel and angle $\theta_{\vec{v}_I}$ are

$$|\vec{v}_I| = \sqrt{I_m^2 + I_n^2}, \theta_{\vec{v}_I} = \tan^{-1} \left(\frac{I_n}{I_m} \right)$$

Figure 1 (in appendix) provides an overall view of the proposed emotion recognition architecture. I_m and I_n are the m and n components of $|\vec{v}_I|$. The module and angle help in normalization for tracking the facial feature point. The grid size involves computing the efficiency of the dense optical flow-based emotions. In this work, the input is a video file, which comprises various body gesture emotions in the continuous image frames. Here we consider leaving one out of frame strategy for finding emotions from the dense optical flow method. For example, the frames have the sequence number from 1 to 10, and we consider frameset as $\{1, 3\}$, $\{2, 4\}$, $\{5, 7\}$, $\{6, 8\}$, $\{7, 9\}$ and so on for obtaining dense flow generation of emotions.

3. TRANSFER LEARNING-BASED PRE-TRAINED EMOTION RECOGNITION

Transfer learning adapts in multiple ways in emotion recognition. The method of transferring knowledge from previously proposed approaches extracts the different classes of emotion-specific information under similar collected datasets [27]. This paper proposes a unique state of the art transfer learning model for recognizing video-based emotions with high-level motion frames using a CNN trained on a significant source of GEMEP corpus dataset (AlexNet, VGG-16, and VGG-19). The CNN architectures [28] is fine-tuned with the parameters of the pre-trained frontal view portion of the training dataset so that it is easy for the model to learn with the weight and bias. Transferring weights can be carried out to other networks for future training and testing a similar new model. Therefore, instead of training from scratch, the system can be pre-trained.

3.1. Convolutional neural network (CNN)

In general, CNN [29] is helpful to detect objects and to classify the images based on analyzing the visual imagery. It consists of a convolution layer, a max-pooling layer followed by one or more completely connected layers. The convolution layer performs the stack of arithmetical operations and forwards the output to the next layer. The pooling/subsampling layer obtains the feature maps depending on the width and height of the input image and performs the kernel-based operation. The fully connected layer uses certain activation functions such as rectified linear unit (ReLU), sigmoid, and tanh to optimize the networks because of the gradient. Furthermore, several regularization units, such as overlapping and dropout, aid the model in optimizing CNN's performance [30].

Convolution network layers play an essential role in developing a new architecture by achieving optimization and batch normalization via various activation functions [31]. This paper proposes an impressive convolutional neural network to extract the in-depth features [32] from the input frames. The trained CNN models include a set of labelled datasets and provide motion frames and the resulting flow area to recognize the emotions. Here, the already obtained optical flow displacement images are passed through the pre-trained CNN architecture to form a feature vector and precisely classify the emotions. We attempt to reduce the loss function in the network using the backpropagation algorithm for result optimization [33].

3.1.1. AlexNet

Alex Krizhevsky trained a deeper neural network with 1.2 million object images followed by eight convolution layers primarily used for image classification applications [34]. It includes five convolution layers followed by subsampling layers, feed-forward layers and a SoftMax layer. Initially, the convolution layers undergo certain filter operations and include ReLU as its non-linear activation function for recognizing the emotions precisely. It uses local response normalization (LRN) and stochastic gradient descent (SGD) to learn the algorithm and provides the best results with an ensemble approach. The main advantage over AlexNet is that; it does the computation faster and intakes the principle of convolution layer, transfer learning and is finely tuned to recognize the face and gesture of the motion frames to predict the emotions accordingly.

3.1.2. Visual geometry group (VGG)

Simonyan and Zisserman [35] pioneered a new architecture named the visual geometry group (VGG) at the University of Oxford. Typically, the system covers VGG-16 and VGG-19. The network has undergone training for the ImageNet large scale visual recognition challenge (ILSVRC) 2012 image classification task using 1.2 million image features of 1,000 different object categories. VGG-16 [36] is the best image recognition system consisting of 224×224 color object images as the input to the convolutional layer, followed by pooling, fully connected and a SoftMax layer. The reduced filter size and computational complexity help to increase the efficiency of the architecture. Spatial resolution is maintained using both row and column padding. Compared to the earlier network, this architecture significantly enhances the localization problem and improves the classification of images. VGG-19 is one of the deep architectures for object detection and image classification challenge for image net classification. In a typical CNN model, various layers are connected for training various tasks. In the end, the network fits several levels of features with its layers.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were implemented in Windows 10 with Intel Core i7 processor using MATLAB 2019b with NVIDIA support. As explained in section 2, the dense optical flow features are extracted from the input video file in one left-of-frame manner and trained with different transfer learning models to identify the particular emotional states. The proposed optical flow model's performance is tested on AlexNet, VGG-16, and VGG-19 deep architectures to identify the accuracy rate of 17 different emotional states of the GEMEP dataset.

4.1. GEMEP dataset

The Geneva multimodal emotion portrayals (GEMEP) corpus [37] is a pair of video and audio records. The dataset extracted from the video consists of 10 subjects, including the non-verbal and emotional speech command of frontal view portions of both male and female characters. The frontal view portion of the pseudo-linguistic sequence dataset simulated for this work is shown in Figure 2.

In total, the dataset consists of 3,042 image rows, portraying 17 different body gesture emotional expressions, including 142 admiration, 227 amusement, 209 anger, 157 anxiety, 122 contempt, 204 despair, 107 disgust, 228 interest, 240 irritations, 200 joy, 147 panic_fear, 312 pleasure, 153 pride, 200 relief,

197 sadness, 95 surprises, and 102 tenderness. The sample dense optical flow frames obtained from the input video file for the emotions of anger, disgust, joy, sadness, and surprise are shown in Figure 3. The recorded video has a 720×576 for a distinct actor, and each video has 25 frames per second. In the first stage, DOF is applied in the GEMEP dataset, which comprises a video of 25 fps of facial and body gestures. In the second stage, fine-tuning on pre-trained CNN based on the videos trained within the network weights increase the performance in each fine-tuning stage.

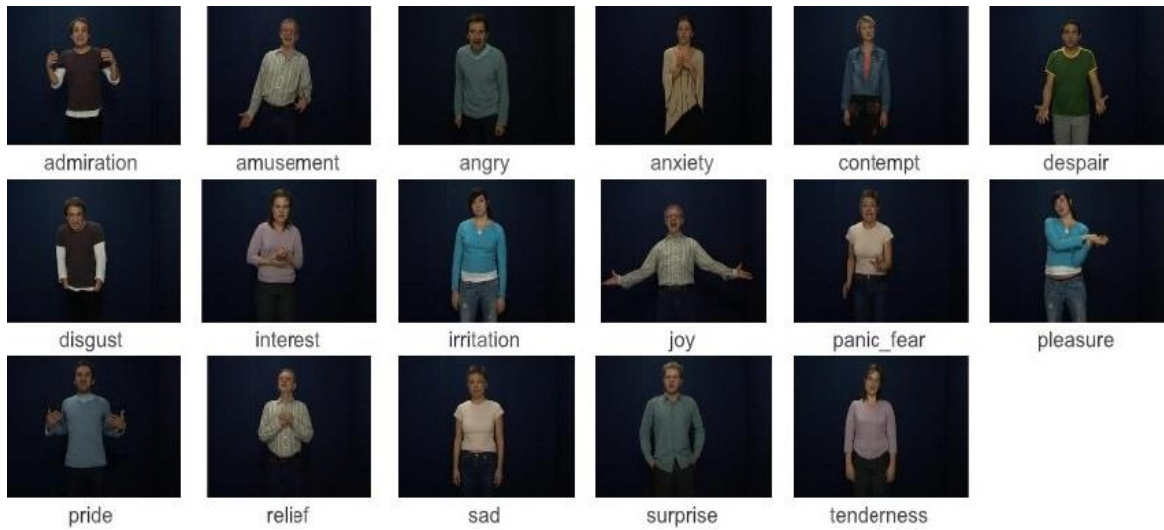


Figure 2. Sample frames of the 17 emotions from the GEMEP dataset

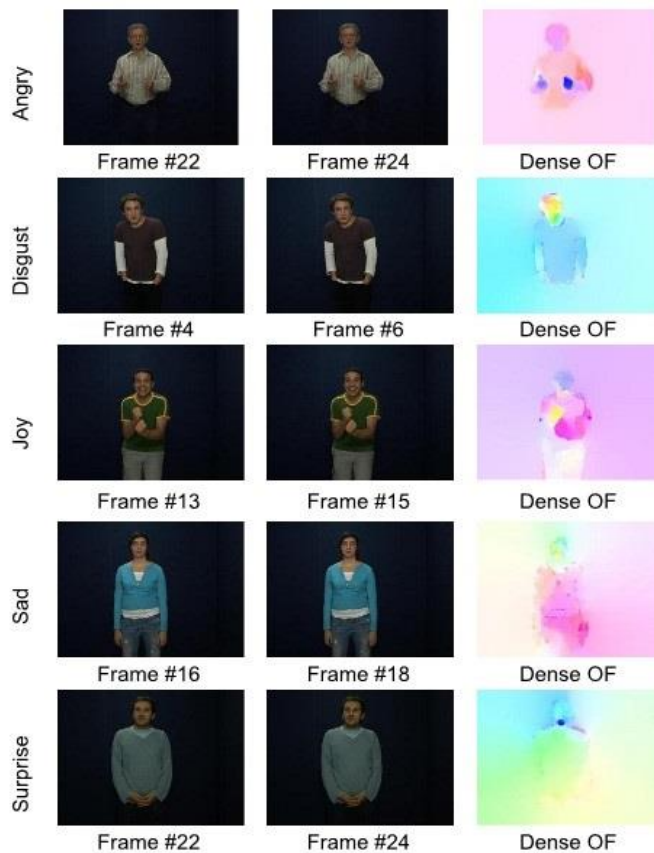


Figure 3. Sample and corresponding dense optical flow frames from GEMEP dataset

4.2. Performance evaluation metrics

The proposed approach uses the dense optical flow method with pre-trained CNN uses 70% for testing and 30% for training the network. The statistical metrics like accuracy (A), precision (P), recall (R) and F-Score has opted for evaluating the execution performance. Here tp and tn are genuine positive, and genuine negative, fp and fn are false positive and false negative predictions.

Accuracy (A) = $\left[\frac{tp+tn}{tp+fp+tn+fn} \right]$ gives the overall excellence of the emotion using DOF. Precision (P) = $\left[\frac{tp}{tp+fp} \right]$ is the degree of perfection. Recognizing emotions in the exact way defined by Recall (R) = $\left[\frac{tp}{tp+fn} \right]$. Specificity (S) = $\left[\frac{tn}{tn+fp} \right]$ gives the accuracy measure of negative emotion. F-Score = $2 \frac{P \times R}{P+R}$ provides the detailed mean of both precision and recall. The AlexNet, VGG-16 and VGG-19 performance measures are shown in Tables 1 to 3. From the results, AlexNet performs well when compares to VGG-16 and VGG-19.

Table 1. Performance measure for AlexNet architecture

Emotions	Precision	Recall	Specificity	F-Score
Admiration	0.925	0.969	0.996	0.947
Amusement	0.968	0.902	0.998	0.934
Angry	0.994	0.952	1.000	0.973
Anxiety	0.951	0.972	0.997	0.961
Contempt	1.000	1.000	1.000	1.000
Despair	1.000	0.880	1.000	0.936
Disgust	1.000	1.000	1.000	1.000
Interest	0.990	1.000	0.999	0.995
Irritation	0.964	0.986	0.997	0.975
Joy	0.994	0.923	1.000	0.957
Panic_fear	0.948	0.970	0.997	0.959
Pleasure	0.979	1.000	0.998	0.989
Pride	0.964	0.957	0.998	0.960
Relief	0.962	0.972	0.997	0.967
Sad	0.978	1.000	0.998	0.989
Surprise	0.875	0.988	0.995	0.928
Tenderness	0.850	1.000	0.994	0.919
mean (μ)	0.961	0.969	0.998	0.964

Table 2. Performance measure obtained for VGG-16

Emotions	Precision	Recall	Specificity	F-Score
Admiration	0.814	0.820	0.991	0.817
Amusement	0.868	0.995	0.988	0.927
Angry	0.983	0.931	0.999	0.956
Anxiety	0.964	0.943	0.998	0.953
Contempt	1.000	1.000	1.000	1.000
Despair	0.966	0.903	0.998	0.933
Disgust	0.835	1.000	0.993	0.910
Interest	0.967	1.000	0.997	0.983
Irritation	0.979	0.852	0.998	0.911
Joy	0.993	0.818	1.000	0.897
Panic_fear	0.801	0.947	0.988	0.868
Pleasure	0.979	1.000	0.998	0.989
Pride	0.991	0.783	1.000	0.874
Relief	0.868	0.912	0.990	0.889
Sad	0.912	1.000	0.993	0.954
Surprise	0.988	0.988	1.000	0.988
Tenderness	0.989	1.000	1.000	0.995
mean (μ)	0.935	0.935	0.996	0.932

Table 3. The performance measure obtained for VGG-19

Emotions	Precision	Recall	Specificity	F-Score
Admiration	0.992	0.984	1.000	0.988
Amusement	0.942	0.961	0.995	0.951
Angry	0.963	0.835	0.998	0.895
Anxiety	0.977	0.915	0.999	0.945
Contempt	1.000	0.836	1.000	0.911
Despair	0.969	0.853	0.998	0.908
Disgust	0.980	1.000	0.999	0.990
Interest	0.958	0.995	0.996	0.976
Irritation	0.982	1.000	0.998	0.991
Joy	1.000	0.867	1.000	0.929
Panic_fear	0.930	1.000	0.996	0.964
Pleasure	0.996	1.000	1.000	0.998
Pride	0.963	0.949	0.998	0.956
Relief	0.918	0.923	0.994	0.920
Sad	0.876	0.994	0.990	0.931
Surprise	0.963	0.929	0.999	0.946
Tenderness	0.645	1.000	0.981	0.784
mean (μ)	0.944	0.944	0.997	0.940

4.3. Results obtained from AlexNet

Our framework extracts dense features using the Farneback method that results in a color-coded image sequence fed to a pre-trained AlexNet architecture to extract image and motion features from optical flow fields. These features passed through fully connected layers for representation adaptation and reduced dimensional to predict emotions. The fully connected layer weights are learned during the training part and produce better test results. The confusion matrix of the CNN based AlexNet model with the dense flow on emotion dataset is shown in Figure 4, where the main diagonal denotes the correct response and most of the emotion classes like contempt, despair, disgust, interest, pleasure, sad and tenderness are predicted well.

The average recognition rate of AlexNet with the dense flow on the GEMEP dataset is 96.63% is better than AlexNet with raw frames of 91.8%. Some of the emotions, like amusement, are misclassified as admiration, and despair is misclassified as tenderness. Angry is partially confused with surprise, pride, and relief. Table 1 represents the performance measure, which shows that the proposed dense optical flow-based transfer learning for AlexNet architecture has a good precision for contempt, despair and disgust. Identifying the actions correctly is given by recall for contempt, disgust, interest, pleasure, sad, tenderness. Specificity states the measure of identifying negative emotions like anger, contempt, despair, disgust, and Finally F-Score values for the GEMEP corpus dataset.

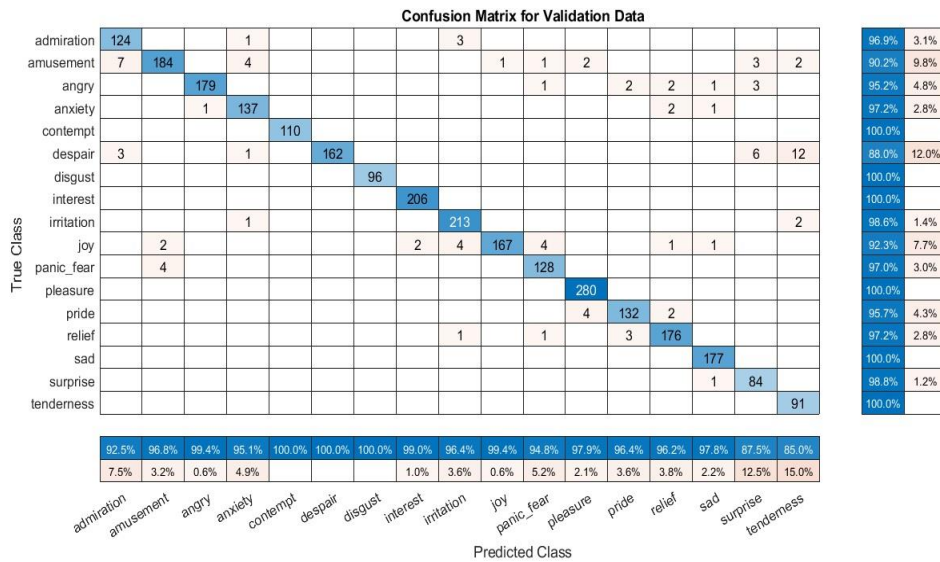


Figure 4. Confusion matrix obtained for the GEMEP dataset using AlexNet

4.4. Results obtained from VGG-16

The confusion matrix of the VGG-16 model with the dense flow on the emotion dataset is shown in Figure 5, where the main diagonal represents the instances classified correctly. The rows in the confusion matrix represent the 17 different emotional class instances, and the column symbolizes the visual posture emotion class predicted by the VGG net.

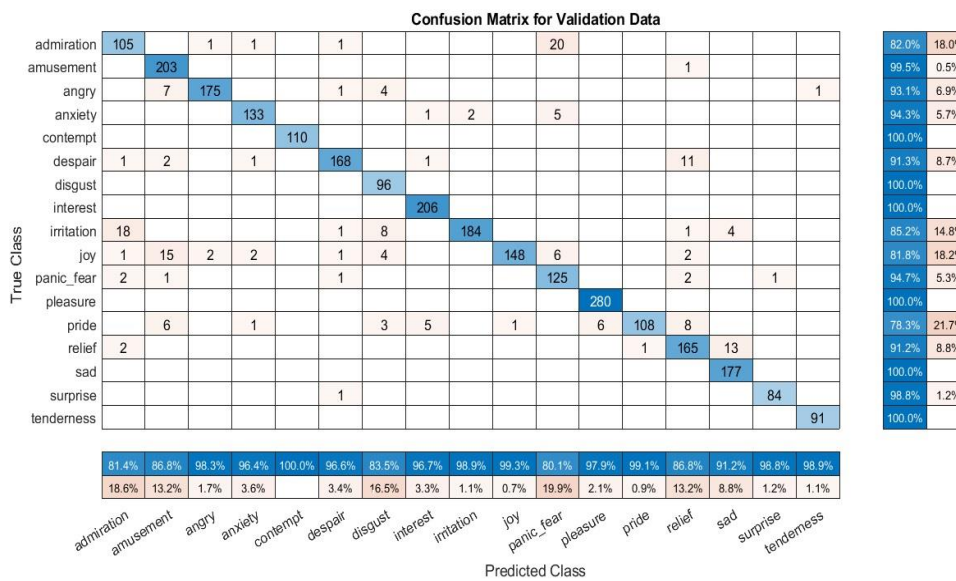


Figure 5. Confusion matrix obtained for the GEMEP dataset using VGG-16

Most emotion classes like contempt, disgust, interest, pleasure, sad and tenderness are predicted with greater accuracy. The average recognition rate of the VGG-16 model with the dense flow on the GEMEP dataset is 93.35% is better than VGG-16 with raw frames of 87.5%. Some emotions like panic_fear and irritation are misclassified as admiration, and amusement are misclassified as joy since it is hard to distinguish the emotions, hence needing further attention. Table 2 represents the performance measure of precision, recall, specificity and F-Score values for dense optical flow-based transfer learning on the GEMEP dataset for VGG-16 architecture.

4.5. Results obtained from VGG-19

The confusion matrix of the VGG-19 model with the dense flow on emotion dataset is shown in Figure 6, where the main diagonal denotes the correct response and most of the emotion classes like disgust, irritation, pleasure, and tenderness are predicted well. An average recognition rate of VGG-19 with the dense flow on the GEMEP dataset is 94.48% is better than VGG-19 with raw frames of 89.7%. Some of the emotions like anger, contempt and despair are misclassified as tenderness. Relief is partially confused with anger, anxiety, and despair. Table 3 represents the performance measure of precision, recall, specificity, and F-score value for dense optical flow-based transfer learning on the GEMEP dataset for VGG-19 architecture. From the results, AlexNet performs well in identifying the 17 emotions when compared to VGG nets.

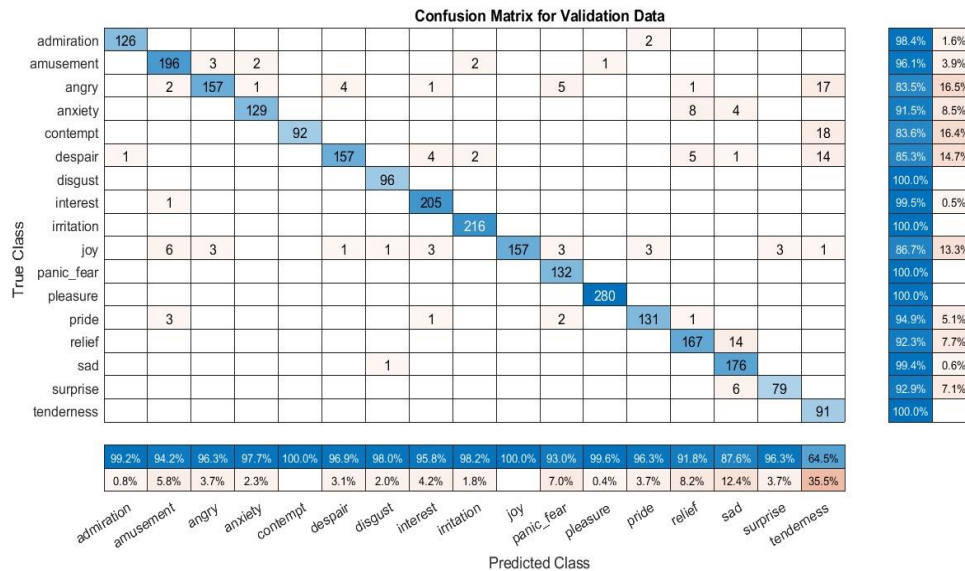


Figure 6. Confusion matrix obtained for the GEMEP dataset using VGG-19

5. COMPARATIVE STUDY

As shown in Table 4, our method based on DOF with the pre-trained CNN model outperformed the recognition of raw frames and other accuracy methods. For comparison, we used techniques that classified the entire human body. Based on the study in the GEMEP dataset, multi blocks maximum intensity code (MBMIC), Santhoshkumar and Geetha [38] used frame differences to extract temporal features. Santhoshkumar and Geetha [39] uses different Bin HoG features (DBLHoG) and spatio-temporal interest points (STIP) from the frontal human body movements. Additionally, Santhoshkumar [11] proposes a deep learning architecture for classifying a video sequence using images. It is observed that our technique improves recognition accuracy by 96.63% and achieves excellent results on the GEMEP dataset.

Table 4. State-of-the-art results on the GEMEP dataset

Methods	Accuracy (%)
MBMIC [38]	94.6%
DBL HoG with KNN [39]	94.8%
FDCNN [11]	95.4%
Ours+Raw image	91.8%
Ours+Dense optical flow	96.6%

Note that we only chose methods that consider the whole human body

6. CONCLUSION

This paper presented a novel approach for recognizing emotions based on visual posture and body gesture movements using dense optical flow (DOF). Building a recent transfer learning-based convolutional neural network (CNN) helps train the network and explicitly predicts the optical flow generation of input motion features. Experiments were carried out on the GEMEP dataset to identify the various emotional states. The results show that AlexNet gives an overall recognition accuracy of 96.63% compared to other traditional approaches for the GEMEP dataset and 94.48% for VGG-19 and 93.35% for VGG-16. The average accuracy rate of various quantitative evaluations is figured with metrics like precision, recall, specificity, and F-Score. The findings concluded that the system could not highly recognize admiration, amusement, joy, and tenderness with greater precision. The proposed work requires less processing time comparing to the existing models. Hence, this model is perfect for recognizing visual emotions, body clues and helps in increasing the system's efficiency and accuracy with a better recognition rate. Our further research extends recognizing emotional states by combining the facial with body gestures using real-time datasets to increase the system's robustness.

APPENDIX

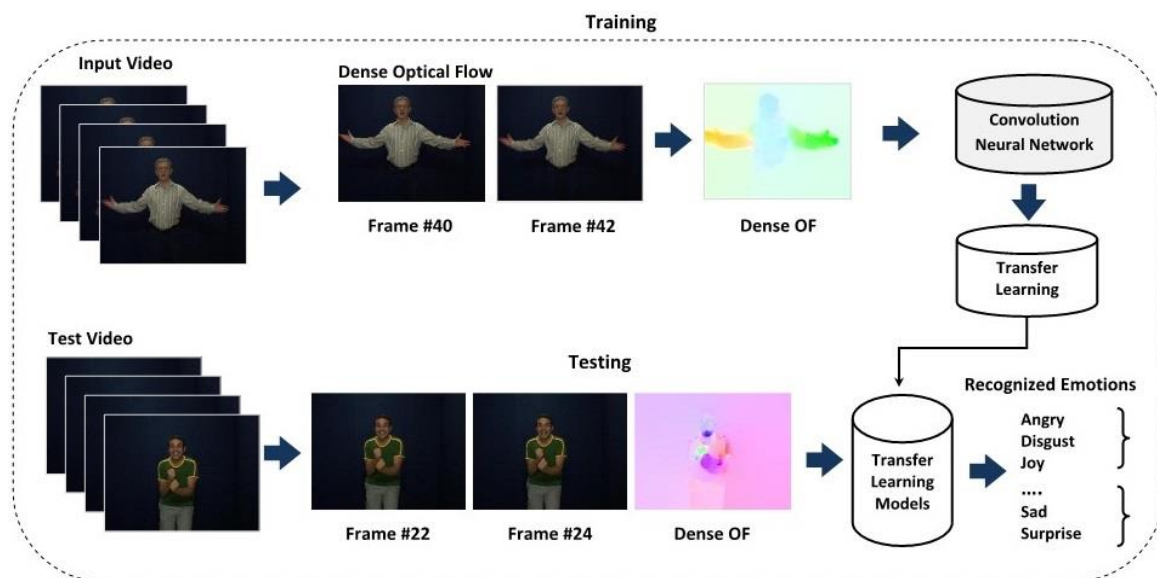






Figure 1. The proposed architecture for emotion recognition

REFERENCES





- [1] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.
- [2] M. Zidan, "A novel quantum computing model based on entanglement degree," *Modern Physics Letters B*, vol. 34, no. 35, Dec. 2020, doi: 10.1142/S0217984920504011.
- [3] M. Zidan, H. Eleuch, and M. Abdel-Aty, "Non-classical computing problems: Toward novel type of quantum computing problems," *Results in Physics*, vol. 21, Feb. 2021, doi: 10.1016/j.rinp.2020.103536.
- [4] M. Zidan *et al.*, "A quantum algorithm based on entanglement measure for classifying Boolean multivariate function into novel hidden classes," *Results in Physics*, vol. 15, Dec. 2019, doi: 10.1016/j.rinp.2019.102549.
- [5] A. Sagheer, M. Zidan, and M. M. Abdelsamea, "A novel autonomous perceptron model for pattern classification applications," *Entropy*, vol. 21, no. 8, Aug. 2019, doi: 10.3390/e21080763.
- [6] M. Zidan, A. H. Abdel-Aty, A. El-Sadek, E. A. Zanaty, and M. Abdel-Aty, "Low-cost autonomous perceptron neural network inspired by quantum computation," in *AIP Conference Proceedings*, 2017, vol. 1905, doi: 10.1063/1.5012145.
- [7] M. Zidan *et al.*, "Quantum classification algorithm based on competitive learning neural network and entanglement measure," *Applied Sciences (Switzerland)*, vol. 9, no. 7, Mar. 2019, doi: 10.3390/app9071277.
- [8] S. Vaijayanthi and J. Arunehru, "Synthesis approach for emotion recognition from cepstral and pitch coefficients using machine learning," in *Lecture Notes in Electrical Engineering*, vol. 733 LNEE, Springer, Singapore, 2021, pp. 515–528.
- [9] D. Suryani, V. Ekaputra, and A. Chowanda, "Multi-modal Asian conversation mobile video dataset for recognition task," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 4042–4046, Oct. 2018, doi: 10.11591/ijece.v8i5.pp4042-4046.
- [10] V. P. Tank and S. K. Hadia, "Creation of speech corpus for emotion analysis in Gujarati language and its evaluation by various speech parameters," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 4752–4758, Oct. 2020, doi: 10.11591/ijece.v10i5.pp4752-4758.

- [11] R. Santhoshkumar, "Deep learning approach: emotion recognition from human body movements," *Journal of Mechanics of Continua And Mathematical Sciences*, vol. 14, no. 3, pp. 182–195, Jun. 2019, doi: 10.26782/jmcs.2019.06.00015.
- [12] A. Agrima, I. Mounir, A. Farchi, L. Elmaazouzi, and B. Mounir, "Emotion recognition from syllabic units using k-nearest-neighbor classification and energy distribution," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5438–5449, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5438-5449.
- [13] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11188 LNCS, no. 10, 2019, pp. 153–163.
- [14] P. Kulkarni and R. T. M., "Analysis on techniques used to recognize and identifying the Human emotions," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 3307–3314, Jun. 2020, doi: 10.11591/ijece.v10i3.pp3307-3314.
- [15] F. Z. Salmam, A. Madani, and M. Kissi, "Emotion recognition from facial expression based on fiducial points detection and using neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 1, pp. 52–59, 2018, doi: 10.11591/ijece.v8i1.pp52-59.
- [16] J. Arunnehr and M. K. Geetha, "Automated complex activity recognition in multiple person interaction," *International Journal of Imaging and Robotics*, vol. 16, no. 3, pp. 71–85, 2016.
- [17] M. K. Geetha, J. Arunnehr, and A. Geetha, "Early recognition of suspicious activity for crime prevention," in *Computer Vision*, IGI Global, 2018, pp. 2139–2165.
- [18] J. C. Hung, K.-C. Lin, and N.-X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Applied Soft Computing*, vol. 84, Nov. 2019, doi: 10.1016/j.asoc.2019.105724.
- [19] R. B. Knapp, J. Kim, and E. André, "Physiological signals and their use in augmenting emotion recognition for human-machine interaction," in *Cognitive Technologies*, no. 9783642151835, Springer Verlag, 2011, pp. 133–159.
- [20] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, 2017, vol. 2017-Janua, pp. 701–704, doi: 10.1109/ICECA.2017.8203631.
- [21] G. Bhargavi, S. Vajayanthi, J. Arunnehr, and P. R. D. Reddy, "A survey on recent deep learning architectures," Springer, Singapore, 2021, pp. 85–103.
- [22] J. Arunnehr and M. Kalaiselvi Geetha, "Difference intensity distance group pattern for recognizing actions in video using support vector machines," *Pattern Recognition and Image Analysis*, vol. 26, no. 4, pp. 688–696, Oct. 2016, doi: 10.1134/S1054661816040015.
- [23] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, Nov. 2007, doi: 10.1016/j.jnca.2006.09.007.
- [24] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636–642, Jun. 1996, doi: 10.1109/34.506414.
- [25] K. N. Ogle, "The perception of the visual world. James J. Gibson; Leonard Carmichael, Ed. Boston: Houghton Mifflin, 1950. 235 pp. \$4.00," *Science*, vol. 113, no. 2940, pp. 535–535, May 1951, doi: 10.1126/science.113.2940.535.
- [26] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision (IJCAI)," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, 1981, vol. 2, pp. 674–679.
- [27] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers in Computer Science*, vol. 2, no. 9, Feb. 2020, doi: 10.3389/fcomp.2020.00009.
- [28] V. Nekrasov, J. Ju, and J. Choi, "Global deconvolutional networks for semantic segmentation," in *Proceedings of the British Machine Vision Conference 2016*, 2016, vol. 2016-Septe, pp. 124.1-124.14, doi: 10.5244/C.30.124.
- [29] J. Walker, A. Gupta, and M. Hebert, "Dense optical flow prediction from a static image," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, vol. 2015 Inter, pp. 2443–2451, doi: 10.1109/ICCV.2015.281.
- [30] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.
- [31] B. Zoph and Q. V. Le, "Searching for activation functions," in *6th International Conference on Learning Representations, ICLR 2018-Workshop Track Proceedings*, 2018, pp. 1–13.
- [32] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, vol. 2016-May, pp. 5200–5204, doi: 10.1109/ICASSP.2016.7472669.
- [33] M. S. P. Wandale, P. P. A. Tijare, P. S. N. Sawalkar, and W. S. Email, "Principal component analysis(PCA) with back propagation neural network(BPNN) for face recognition system," *Computer Science*, vol. 2, no. 4, pp. 410–415, 2013.
- [34] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPRW.2009.5206848.
- [35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations*, Sep. 2014.
- [36] P. Networks, "VGG16-convolutional network for classification and detection," *Neurohive*, 2018. .
- [37] T. Baltrusaitis *et al.*, "Real-time inference of mental states from facial expressions and upper body gestures," in *Face and Gesture 2011*, Mar. 2011, pp. 909–914, doi: 10.1109/FG.2011.5771372.
- [38] R. Santhoshkumar and M. K. Geetha, "Emotion recognition on multi view static action videos using multi blocks maximum intensity code (MBMIC)," in *New Trends in Computational Vision and Bio-inspired Computing*, Cham: Springer International Publishing, 2020, pp. 1143–1151.
- [39] R. Santhoshkumar and M. K. Geetha, "A study on discrete action sequences using deep emotional intelligence," in *Deep Learning in Data Analytics*, vol. 4, no. 4, 2022, pp. 3–24.

BIOGRAPHIES OF AUTHORS

Vaijyanthi Sekar     received her B.E. degree in Computer Science and Engineering in 2012 from IFET College of Engineering and Technology, Villupuram and M.Tech. Degree in Computer Science and Engineering in 2014, from Pondicherry Engineering College, Pondicherry. Presently, she is working as a Full-Time PhD scholar in the Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Chennai - 26, Tamilnadu. She has published 3 papers in reputed international journals, and she has presented 4 international conference papers. Her area of specialization includes image and video processing, pattern classification and machine learning. She can be contacted at email: vaijyanthisekar@gmail.com.



Arunnehru Jawaharlalnehru     received his Diploma in Computer Technology from Muthiah Polytechnic College, Annamalai Nagar, in 2004. He obtained his B.E. degree in Computer Science Engineering in 2008, M.E. degree in Computer Science and Engineering in 2010, and the PhD degree in Computer Science and Engineering in 2017 from Annamalai University, Annamalai Nagar. Presently, he is working as an Assistant Professor in the Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu. He has published more than 45 papers in reputed international journals and presented 20 international conference papers. His area of specialization includes image and video processing, pattern classification, machine learning and deep learning. He is a life member of the Indian Society of Technical Education (ISTE), IET and IAENG. He received the best paper award at the Springer international conference in 2014. He is a reviewer for Elsevier and other internationally reputed journals. He can be contacted at email: arunnehru.aucse@gmail.com. Further info on his homepage: <https://www.srmist.edu.in/vadapalani/faculty/cse-arunnehru>