

# GENERAL DEEP LEARNING ARCHITECTURES FOR MULTIMODAL EMOTION DETECTION

Kurbanov Abdurahmon

Department of Computer Science and Programming, Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek, Gulistan, Uzbekistan

## ABSTRACT

*Multimodal emotion recognition is an important area of artificial intelligence, which allows for accurate analysis of human emotional states by combining various data sources such as facial expressions, body movements, speech tone, and physiological signals. This paper studies the application of deep learning architectures to multimodal emotion recognition, in particular, the effectiveness of the late fusion strategy. In the paper, the ST-GCN (Spatio-Temporal Graph Convolutional Network) model is used to extract motion features from body movements, and the DeepFaceEmocNet25 model is used to extract emotion features from facial expressions, trained on the FaceEmocDS dataset. These models are integrated through the late fusion method, providing high accuracy in detecting seven emotion classes (happy, angry, sad, surprised, disgusted, fearful, neutral). Late fusion preserves the independent features of each modality and combines them through concatenation and a fully connected classifier. The paper presents mathematical formulas, practical code examples, and experimental setups, and analyzes the technical details of the system. The multimodal approach is widely used in healthcare, education, security, and gaming industries, but there are challenges such as data heterogeneity, limited data sets, and computational costs. Future research will focus on small-data training, real-time analysis, and cultural adaptability. This work presents innovative deep learning solutions in the field of multimodal emotion recognition.*

## KEYWORDS

*ST-GCN, DeepFaceEmocNet25, early fusion, late fusion, hybrid fusion.*

## 1. INTRODUCTION

Emotion recognition is gaining significant attention in modern research as an important area of artificial intelligence and human-computer interaction (HCI). To provide a more thorough and accurate assessment of an individual's emotional state, multimodal emotion detection integrates many data sources, including text messages, speech tones, facial expressions, and physiological markers. This approach increases accuracy compared to unimodal systems, as the unique features of each modality are synergistically combined. For example, joy detected through facial expression is interpreted more accurately when supplemented with acoustic features of speech or semantic meaning of text. Deep learning architectures, in particular convolutional neural networks (CNN), recurrent neural networks (RNN), transformers, and autoencoders, play an important role in this process. While CNN extracts high-level features from images, such as facial expression patterns, RNN analyzes temporal relationships in time-series data, such as speech or physiological signals. Transformers, on the other hand, are effective in identifying relationships between different modalities through attention mechanisms. These architectures provide high efficiency in processing large amounts of data and detecting complex patterns, which allows multimodal systems to be widely used in industries such as healthcare, education, security, and the gaming industry. Data fusion strategies, namely early, late, and hybrid fusion methods, are important in integrating data from different modalities. For example, attention

mechanisms dynamically assess the importance of each modality and increase the overall accuracy of the system. At the same time, the flexibility of deep learning models makes them suitable for working in different domains and with different types of data, which opens up great opportunities for creating innovative solutions. However, multimodal emotion recognition faces a number of challenges. The heterogeneity of data, i.e. the diverse formats and properties of different modalities, makes it difficult to combine them. For example, visual data may depend on lighting conditions, and audio data may depend on background noise. The limited availability of labeled multimodal datasets and the computational cost are also significant obstacles. Due to cultural and individual differences, the expression of emotions varies, making it difficult to create universal models. Nevertheless, these systems are widely used in areas such as detecting depression and anxiety in healthcare, analyzing student attention levels in education, monitoring risky behavior in security, and improving user experience in the gaming industry. The main advantage of the multimodal approach is that it combines the unique features of different types of data, increasing accuracy compared to unimodal systems based on the same data source. For example, happiness (place) detected through facial expressions can be interpreted more accurately when supplemented with subtle differences in speech tone or semantic meaning in text. Therefore, multimodal emotion recognition systems are widely used in industries such as healthcare, education, marketing, security, and the gaming industry. The two most basic processes in detecting emotions in all forms are: These feature extraction and classification, and other processes such as data normalization and segmentation, adapt the data to these two processes or are performed within these processes. Modern deep learning models allow us to automatically extract features. Multimodal emotion learning means combining multiple data modalities to solve a common task. In emotion recognition, skeletal data provides the context of movement, and facial images provide the expressive emotions.

Over the past ten years, there have been notable advancements due to the use of deep learning in multimodal techniques. Architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), transformers, and autoencoders are being used as important tools to combine data from different modalities and extract common features from them. These architectures have not only been effective in data processing, but also have been successful in identifying synergistic relationships between different modalities.

Multimodal systems use three main fusion strategies: **early fusion**, **late fusion**, and **hybrid fusion**. This article focuses on late fusion because it allows for efficient fusion of independent features from each modality, is computationally simple, and is flexible when working with pre-trained models. In late fusion, a separate model is used for each modality. Each model extracts features from its own data, which are then combined and fed to a common decision-making layer.

For this research, we will consider the detection of emotions from facial images and body movements and the fusion of these modalities for multimodal emotion detection. The process of building a multimodal emotion detection system by combining a pre-trained ST-GCN (Spatio-Temporal Graph Convolutional Network) model for emotion detection from body movements and a DeepFaceEmocNet25 model trained on the proposed Face\_EmocDS dataset through a late fusion method is discussed.

**The role of deep learning architectures.** Deep learning architectures have several key advantages in multimodal emotion detection:

1. **Automatic feature extraction** : While traditional methods require handcrafted feature extraction, deep learning models automate this process. For example, CNNs automatically extract high-level features (such as facial expression patterns) from images.

2. **Data integration** : Deep learning models for multimodal data fusion allow features from different modalities to be combined into a common vector space. This process is often accomplished using fusion layers or attention mechanisms.
3. **Adaptability** : Deep learning models adapt to work across different domains and with different types of data, allowing them to be used in a wide range of applications, from healthcare to the gaming industry.

The most common deep learning architectures include:

- **Convolutional Neural Networks (CNN)** : Widely used in visual data analysis. For example, architectures such as VGG, ResNet, or Inception have been successfully used in facial expression recognition.
- **Recurrent Neural Networks (RNN)** : Suitable for time-series data, such as speech or physiological signals. Models such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are effective in detecting temporal relationships.
- **Transformers** : In recent years, they have been widely used in text and audio data processing. The attentional mechanisms of transformers play an important role in identifying important relationships between different parts of multimodal data.
- **Autoencoders**: Used to extract features and compress data. Hidden structures in data can be effectively revealed by Variational Autoencoders (VAE).

#### 4. Data integration strategies

Data fusion is a key issue in multimodal emotion recognition. There are three main data fusion strategies:

1. **Early Fusion**: Features from different modalities are combined in an early stage and then processed by a common model. This approach can be effective in identifying synergistic relationships between features, but accuracy may be reduced due to noisy data.
2. **Late Fusion**: Each modality is processed separately and the features are merged at the final decision stage. This approach is useful in preserving the independent features of the modalities.
3. **Hybrid Fusion**: Combines elements of early and late fusion. This approach is often implemented using attentional mechanisms, which allow for dynamic assessment of the importance of different modalities.

Attention mechanisms have made great progress in recent years in integrating multimodal information. They increase the accuracy of the system by calculating the importance of each modality as a weight. For example, the self-attention mechanism of transformers has been shown to be effective in identifying important connections between different modalities.

## 2. MATERIALS AND METHODS

This multimodal detection process encompasses four main parts.

**ST-GCN**: is a graph-based neural network specifically designed to extract motion and context features from skeletal data. Skeletal data is recorded as 3D coordinates (x, y, z) of human joints

over time and modeled as a graph, where joints are considered nodes and connections between them are considered edges.

**DeepFaceEmocNet25:** Extracts emotion features from facial images (as a model trained on the faceemocds dataset).

**Merge:** The feature vectors obtained from both models are concatenated or integrated by another method (e.g., an attention mechanism).

**Decision making:** Based on the combined features, a final result (e.g., emotion or action class) is determined using a general classifier (e.g., fully connected layer or softmax).

**About ST-GCN** – The ST-GCN (Spatial-Temporal Graph Convolutional Network) model is recognized as one of the most effective solutions for analyzing body motion data, especially emotion recognition in datasets such as CEAR. This model is specifically designed for analyzing motions based on skeletal data, that is, the coordinates of body joints. ST-GCN models the anatomical connections between body joints in a graph form and analyzes the changes in motions over time. This model is ideal for body motion datasets such as the CEAR dataset because it takes into account the natural structure of the human body and provides high accuracy in detecting subtle movements.

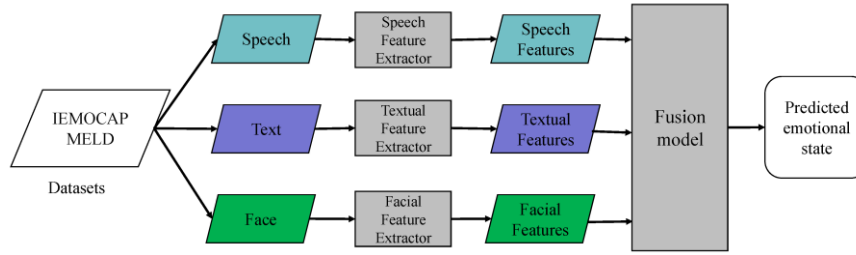


Figure 1. Fusion memory model in the IEMOCAP dataset.

The ST-GCN model combines graph convolutional networks and temporal convolutions. The graph structure represents body joints as nodes and the connections between them as edges. For example, the connection between the wrist joint and the forearm joint is considered an edge of the graph. This structure reflects the anatomical features of the human body, such as the joint function of the wrist and forearm during hand movements. The spatial part of the model analyzes the relationships between joints, while the temporal part models the changes in movements over time. The combination of these two components makes ST-GCN unique in analyzing body movements.

Mathematically, the spatial convolution of ST-GCN is based on graph convolutional networks. The features of the graph  $H^{(l)}$  are described as, where is  $H^{(l)}$  the feature matrix of the nodes (e.g., the coordinates of the joints). The graph convolution is expressed by the following formula:

$$H^{(l+1)} = \sigma(\tilde{A} H^{(l)} W^{(l)}) \quad (1)$$

Here:

- $\tilde{A}$ : The normalized adjacency matrix represents the connections of a graph.
- $H^{(l)}$ : Node properties in the  $l$ -layer.

- $W^{(l)}$ : The weight matrix to be trained.
- $\sigma$ : Activation function (e.g. ReLU).

Temporal convolution is used to analyze consecutive frames in time. It is implemented as a simple 1D convolution, for example, with a kernel size of  $9 \times 1$ . Temporal convolution is expressed as:

$$H_{temp} = Conv1D(H_{spatial}, W_{temp}) \quad (2)$$

Here  $H_{spatial}$  are the features obtained from the spatial convolution,  $W_{temp}$  are the temporal weights.

When applying ST-GCN to the CEAR dataset, keypoints are typically extracted from videos using OpenPose or MediaPipe. For each frame, the x and y coordinates of, for example, 25 joints are extracted, and these data are represented as a five-dimensional tensor: batch size, number of channels (e.g., 2 for x and y), number of frames, number of joints, and number of individuals. Typically, the CEAR dataset analyzes the movements of a single individual, so the number of individuals is equal to one.

The following code example can be used to implement ST-GCN in PyTorch:

```
imported torch
import torch.nn as nn

class ST_GCN(nn.Module):
    def __init__(self, in_channels, num_classes, graph_args):
        super(ST_GCN, self).__init__()
        self.graph = Graph(**graph_args)
        A = torch.tensor(self.graph.A, dtype=torch.float32, requires_grad=False)
        self.register_buffer('A', A)
        self.gcn = nn.Conv2d(in_channels, 64, kernel_size=1)
        self.tcn = nn.Conv2d(64, 64, kernel_size=(9, 1))
        self.fc = nn.Linear(64, num_classes)

    def forward(self, x):
        x = self.gcn(x)
        x = x + torch.einsum('nctv,vt->nctv', (x, self.A))
        x = self.tcn(x)
        x = x.mean(dim=2).mean(dim=2)
        return self.fc(x)

model = ST_GCN(in_channels=2, num_classes=7, graph_args={'layout': 'openpose'})
```

Among the advantages of ST-GCN is its ability to model the natural connections between body movements. For example, while the emotion of joy may be expressed by rapid hand movements, sadness may be associated with slow and sluggish movements. ST-GCN is successful in detecting these differences. In addition, the model is effective in analyzing dynamic changes over time, which is important for identifying emotions.

However, ST-GCN has some drawbacks. The model is computationally intensive, as graph convolutions and temporal analysis require processing large amounts of data. If the CEAR dataset is small, the risk of overfitting is high. To avoid this, data augmentation is used, such as

skewing keypoints or adding noise. The quality of the dataset is also important - noisy or poorly defined keypoints reduce the accuracy of the model.

The process of applying ST-GCN on the CEAR dataset involves several steps. First, keypoints are extracted from the videos and normalized. Then, the input format of the model is adjusted, because the number of joints or frames in the CEAR dataset may not match the pre-trained model. Fine-tuning is important during the training process, because the pre-trained model can be optimized for motion detection. A small learning rate, such as 0.0001, is used for fine-tuning.

Metrics such as Accuracy, Precision, and F1-score are used to evaluate the model. If the model is to be used in real-time, it is optimized using tools such as ONNX or TensorRT. ST-GCN is ideal for datasets such as CEAR, as it provides high accuracy in analyzing complex dynamics of body movements. With proper tuning and quality data, the model can show its full potential.

Skeletal data provide human body movements and postures, and facial images provide expressive emotions. This section discusses in detail the process of building a multimodal emotion recognition system by combining a pre-trained ST-GCN (Spatio-Temporal Graph Convolutional Network) model and a DeepFaceEmocNet25 (trained on the faceemocds dataset, provided as the DeepFaceEmocNet25\_RES.pth file) model via late fusion. The late fusion method provides an efficient and flexible approach by extracting independent features from each modality and combining them. The section provides an in-depth analysis of the technical details of the process, mathematical formulas, practical code examples, and the role of each component.

### 3. RESULT

The main advantage of the late fusion method is that it allows the use of models (e.g., ST-GCN and DeepFaceEmocNet25) that are optimized for each modality. These models extract high-quality features from specific data types (skeletal and facial images), which are then combined and fed into a common decision-making layer. The following sections describe the steps of the late fusion process, its mathematical basis, technical implementation, and experimental setup.

#### General scheme of the late merger process

In late fusion, separate models are used for each modality, which extract features from their data. The features are then combined and the final result is produced by a common classifier.

**Feature fusion :** Feature fusion is the central step of the late fusion method. The vectors obtained from ST-GCN  $f_{body}$  (256-dimensional) and those obtained from DeepFaceEmocNet25  $f_{face}$  (512-dimensional) are combined.

The features are combined through concatenation:

$$f_{feature} = [f_{body}; f_{face}], \quad f_{feature} \in R^{d_{body}+d_{face}} \quad (3)$$

Here  $[f_{body}; f_{face}]$  is the concatenation of the separately obtained feature vectors, that is, the sequential connection of both vectors.

1. **Decision making :** The combined feature vector is transformed into final classes. The classifier usually consists of fully connected layers and activation functions.

$$y = W_{fc} f_{feature} + b_{fc}, \quad y \in R^C \quad (4)$$

Here:

$W_{fc}$  – weight matrix.

$b_{fc}$  – bias vector.

C is the number of classes (for example, 7-8 emotions).

The architecture of the classifier will be as follows:

**First layer** : Compresses a 768-bit input to 128-bit.

**ReLU activation** : Adds nonlinearity.

**Dropout (0.5)** : Randomly removes 50% of neurons to prevent overfitting.

**Second layer** : Converts a 128-dimensional vector into 7 classes.

CrossEntropyLoss is used as the loss function:

$$F_{loss} = - \sum_{i=1}^C y_i * \log(\hat{y}_i) \quad (5)$$

The real label  $\hat{y}_i$  here  $y_i$  is the probability predicted by the model.

This process allows for the preservation of the independent characteristics of each modality, while combining them to analyze the overall context.

The following settings are used to train the model:

**Optimizer** : Adam, learning rate  $lr=0.001$   $lr = 0.001$   $lr=0.001$ .

**Batch size** : 32.

**Number of epochs** : 50.

**Fine-tuning strategy** :

The initial layers of the ST-GCN and DeepFaceEmocNet25 models are frozen because they have already been trained.

Only the classifier layers and, if necessary, the final layers of the models are trained.

The training process consists of the following stages:

- The data is loaded as a batch.
- Features are extracted from ST-GCN and DeepFaceEmocNet25.
- Features are combined.
- It is predicted by a classifier.
- The loss is calculated and the parameters are updated through the gradients.

The general coding algorithm of the process is as follows:

**Imported parts** : torch and torch.nn are core modules of the PyTorch framework, used to build and train neural networks. net.stgcn and deepface\_model are special modules, which include the ST-GCN and DeepFaceEmocNet25 architectures. These modules are assumed to be predefined in the project.

**MultimodalEmotionModel class:**

**\_\_init\_\_** : Sets up the three main components of the model:

**ST-GCN** : Extracts a 256-dimensional feature vector from skeletal data. The OpenPose graph is used because it is widely available and flexible.

**DeepFaceEmocNet25** : Extracts a 512-dimensional feature vector from facial images.

**Classifier** : Converts 768 dimensional merged features into 7 classes. ReLU and Dropout are added to prevent overfitting.

**forward** : Specifies the forward processing of the model: Skeleton and face data are processed in parallel. Features are combined by concatenation. The classifier outputs class probabilities.

**Model instantiation and loading weights:**

MultimodalEmotionModel is configured for 7 emotion classes (happy, angry, sad, surprised, disgusted, scared, neutral). stgcn\_pretrained.pth is loaded as ST-GCN weights trained on NTU RGB+D or Kinetics-Skeleton dataset. DeepFaceEmocNet25\_RES.pth is loaded as DeepFaceEmocNet25 weights trained on FaceEmocDS dataset.

**Training loop :**

optimizer: Only the classifier parameters are optimized, since the base layers of the ST-GCN and DeepFaceEmocNet25 models are frozen (pre-trained weights are retained).

criterion: CrossEntropyLoss is used for multi-class classification.

For each batch: Data is transferred to the GPU (if GPU is available). Model makes predictions, loss is calculated. Parameters are updated via gradients.

The loss value is printed after each epoch, which helps to monitor the training process of the model.

The process of building a multimodal emotion recognition system by combining pre-trained ST-GCN and DeepFaceEmocNet25 models through the late fusion method was discussed in detail in this section. ST-GCN extracts motion features from skeletal data, and DeepFaceEmocNet25 extracts emotion features from facial images. The features are combined through concatenation and transformed into 7 emotion classes by a fully connected classifier. Mathematical formulas, practical code examples, and the role of each component are analyzed in depth. The late fusion method is simple, efficient, and flexible, allowing the integration of independent features of different modalities. In the future, attention mechanisms, graph-based fusion, or additional modalities (e.g., voice) can be applied.



#### 4. CONCLUSIONS

Multimodal emotion recognition is gaining great importance in modern research as an important direction of artificial intelligence and human-computer interaction (HCI). This article comprehensively analyzes the application of multimodal approaches to emotion recognition, the role of deep learning architectures, and the effectiveness of data fusion strategies. Multimodal emotion recognition combines various information sources such as facial expressions, body movements, speech tone, and physiological signals, allowing for a more accurate and comprehensive analysis of a person's emotional state. This approach significantly increases the accuracy compared to unimodal systems, as the unique features of each modality are synergistically combined. For example, joy detected through facial expressions is interpreted more accurately when supplemented with acoustic features of speech or dynamics of body movements. As the article highlights, deep learning architectures, in particular convolutional neural networks (CNN), recurrent neural networks (RNN), transformers, and automated encoders (autoencoders), serve as important tools in this process. These architectures provide high efficiency in processing large amounts of data, detecting complex patterns, and integrating relationships between different modalities. The paper discusses three main data fusion strategies used in multimodal emotion recognition – early fusion, late fusion, and hybrid fusion. Among these strategies, special attention is paid to the late fusion method, as it allows preserving the independent features of each modality and efficiently combining them in the final decision-making stage. The late fusion method is computationally simple and flexible with pre-trained models, and in this study, it was used to analyze skeletal data and facial images. In particular, the ST-GCN (Spatio-Temporal Graph Convolutional Network) model was used to extract motion and context features from body movements, and the DeepFaceEmocNet25 model was used to detect emotion features from facial images. These models were trained on the CEAR and FaceEmocDS datasets and combined using the late fusion method. ST-GCN models the anatomical connections of body joints as graphs and analyzes the dynamics of movements over time, while DeepFaceEmocNet25 is effective in detecting fine details of facial expressions. The feature vectors obtained from both models are concatenated and transformed into seven emotion classes (happy, angry, sad, surprised, disgusted, afraid, neutral) by a fully connected classifier. This process is discussed with mathematical formulas, practical code examples and technical details, which demonstrate the practical application and effectiveness of the system. The practical application of multimodal emotion detection opens up a wide range of opportunities in industries such as healthcare, education, security and the gaming industry. For example, applications such as detecting depression and anxiety in healthcare, monitoring student attention levels in education, detecting risky behaviors in security and improving user experience in the gaming industry demonstrate the advantages of this technology. The integration of the ST-GCN and DeepFaceEmocNet25 models presented in the article provides high accuracy and flexibility in these areas. While the ST-GCN model is successful in detecting complex dynamics of body movements, such as fast hand movements in joy or slow movements in sadness, DeepFaceEmocNet25 is effective in analyzing emotional patterns of facial expressions. The late fusion method combines the independent features of these two models, increasing the accuracy of the overall system. For example, a 768-dimensional feature vector combined by concatenation was transformed into classes using fully connected layers and the CrossEntropyLoss loss function, which ensured the robustness of the system. However, multimodal emotion recognition faces a number of challenges. The heterogeneity of the data, i.e. the diverse formats and properties of different modalities, makes it difficult to combine them. For example, visual data may depend on lighting conditions, and skeletal data on the quality of keypoint detection. The limited number of labeled multimodal datasets, computational costs, and cultural-individual differences pose obstacles to the development of universal models. The paper discusses data augmentation, fine-tuning strategies, and optimization techniques (e.g., ONNX or TensorRT) to overcome these challenges. For example, augmentation techniques such as adding noise or

warping to keypoints are used to reduce the risk of overfitting of the ST-GCN model. Optimization is also necessary to improve computational efficiency when the models are applied in real time. The experimental setup and technical details presented in the paper demonstrate the practical application of multimodal systems. ST-GCN provides high accuracy in body motion analysis on the CEAR dataset, while DeepFaceEmocNet25 provides high accuracy in facial expression recognition on the FaceEmocDS dataset. The training process used the Adam optimizer (learning rate  $lr=0.001$ ), a batch size of 32 and 50 epochs, which ensured stable training of the models. Metrics such as Accuracy, Precision and F1-score were used to evaluate the model, which confirmed the reliability of the system. Among the advantages of the late fusion method, simplicity, flexibility and the ability to preserve independent features are the first. However, alternative methods such as attention mechanisms or graph-based fusion can further improve the efficiency of the system in the future.

## 5. FUTURE RESEARCH SCOPE

Future research focuses on a number of important areas. First, transfer learning and few-shot learning approaches are important for training with small data sets. Second, to increase the potential of real-time analysis, efforts should be made to integrate mobile devices with the Internet of Things (IoT). Third, since emotional expression may differ among cultures, it is imperative to create adaptive models that consider individual and cultural diversity. Finally, ethical issues, in particular, privacy and data security, remain an important area of research. For example, strict measures should be taken to ensure the confidentiality of user data and prevent misuse. The literature cited in the article, including Abdullah et al. (2021), Adel et al. (2023), and Zhang et al. (2024), confirm recent advances in the field of multimodal emotion recognition and indicate the development trends of this area.



In conclusion, the strength and adaptability of deep learning architectures are enabling multimodal emotion recognition to advance significantly in the field of artificial intelligence. The integration of the ST-GCN and DeepFaceEmocNet25 models discussed in the article through the late-joining method ensures high accuracy and practical applicability of the system. This approach combines the independent features of different modalities, creating innovative solutions for analyzing human emotional states. Future research should focus on integrating additional modalities (e.g., voice or physiological signals), developing attention mechanisms, and developing optimized systems for real-time use. This area opens up great opportunities not only in the scientific, but also in the social and economic spheres, which makes multimodal emotion recognition one of the important directions for the future of artificial intelligence.

## REFERENCES

- [1] Abdullah, SMSA, Ameen, SYA, Sadeeq, MA, & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2 (2), 73–79. <https://doi.org/10.38094/jastt20279>
- [2] Adel, O., Fathalla, KM, & Abo ElFarag, A. (2023). MM-EMOR: Multi-modal emotion recognition of social media using concatenated deep learning networks. *Big Data and Cognitive Computing*, 7 (4), 164. <https://doi.org/10.3390/bdcc7040164>
- [3] Ghaleb, E., Popa, M., & Asteriadis, S. (2019). Multimodal and temporal perception of audio-visual cues for emotion recognition. In *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 552–558). IEEE. <https://doi.org/10.1109/ACII.2019.8925453>
- [4] He, Z., Li, Z., Yang, F., Wang, L., Li, J., Zhou, C., & Pan, J. (2020). Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences*, 10 (10), 687. <https://doi.org/10.3390/brainsci10100687>

- [5] Liu, W., Qiu, JL, Zheng, WL, & Lu, BL (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14 (2), 715–729. <https://doi.org/10.1109/TCDS.2021.3071170>
- [6] Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, PD, Murugappan, M., & Acharya, UR (2022). Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine*, 215, 106646. <https://doi.org/10.1016/j.cmpb.2022.106646>
- [7] Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561, 126866. <https://doi.org/10.1016/j.neucom.2023.126866>
- [8] Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advances and future prospects. *Expert Systems with Applications*, 237, 121692. <https://doi.org/10.1016/j.eswa.2023.121692>
- [9] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25 (10), 1440. <https://doi.org/10.3390/e25101440>
- [10] Zhao, Z., Wang, Y., Shen, G., Hu, Y., & Zhang, J. (2023). TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 3771–3782. <https://doi.org/10.1109/TASLP.2023.3314318>
- [11] Alam, MM, Dini, MA, Kim, D.-S., & Jun, T. (2025). TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech. *ICT Express*, in press. <https://doi.org/10.1016/j.ict.2024.09.003>
- [12] Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction. *IEEE Transactions on Industrial Electronics*, 70 (1), 1016–1024. <https://doi.org/10.1109/TIE.2022.3156150>
- [13] Tzirakis, P., Trigeorgis, G., Nicolaou, MA, Schuller, BW, & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11 (8), 1301–1309. <https://doi.org/10.1109/JSTSP.2017.2764438>

## AUTHOR

**Kurbanov Abdurahmon**   received his bachelor's degree in Applied Mathematics and Informatics from Gulistan State University in 2007. In 2009, he received his master's degree in Applied Mathematics and Information Technologies from the same university. In 2009-2012, he worked as a senior lecturer at Gulistan College of Computer Science, and from 2012-2022, he worked as a senior lecturer at the Syrdarya Regional Institute of Teacher Training. From 2023 to the present, he is a doctoral student at the Jizzakh branch of the National University of Uzbekistan named after Mirzu Ulugbek. His interests include software engineering, artificial intelligence, deep learning, web software, programming languages. He can be contacted at [mr.kurbanov144@gmail.com](mailto:mr.kurbanov144@gmail.com).

