Information, Computing and Intelligent Systems, 2025, No. 7, 127 – 141

UDC 519.688; 004.89; 004.9 https://doi.org/10.20535/2786-8729.7.2025.333629

METHOD FOR COMBINING CNN-BASED FEATURES WITH GEOMETRIC FACIAL DESCRIPTORS IN EMOTION RECOGNITION

Liudmyla Zinchenko
https://orcid.org/0009-0009-3956-5854

National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

zinchenko.liudmyla@gmail.com

This study presents a method for combining CNN-based visual features with geometric facial descriptors to improve the accuracy of emotion recognition in static images. The method integrates deep convolutional embeddings extracted from a pre-trained `ResNetV2_101` model within the ML.NET framework with handcrafted geometric features computed from facial landmarks. Open-source datasets containing labeled emotional categories were used for experiments. At the first stage, deep image embeddings were obtained through transfer learning. At the second stage, 68 facial landmarks were detected to calculate distances and proportional relationships such as interocular distance, mouth width, eyebrow height, and other geometry-based indicators. These visual and geometric representations were concatenated into a unified feature space and classified using a multiclass linear model. The hybrid method achieved approximately 4% higher accuracy than the baseline CNN model relying solely on pixel-level features (from about 63% to 67%), confirming that combining heterogeneous features enhances generalization and robustness. The results also highlight that geometric descriptors act as stabilizing factors, compensating for noise, occlusions, and lighting variations that degrade CNN-only models. The developed pipeline demonstrates the feasibility of integrating interpretable geometric cues with deep embeddings directly in C# using ML.NET. The research novelty lies in proposing an interpretable hybrid model for emotion recognition that improves reliability while maintaining compatibility with .NET-based applications. The approach offers an accessible solution for developers working within enterprise .NET ecosystems, enabling direct deployment without cross-language integration. Future research will focus on extending the model toward multimodal emotion analysis that incorporates speech, gesture, and physiological signals to enhance contextual understanding of affective states. Additionally, the hybrid model can serve as a diagnostic tool for studying emotion dynamics in psychological or behavioral research.

Keywords: emotion recognition, facial landmarks, convolutional neural networks, .NET framework, feature fusion.

1. Introduction

Emotion recognition has become one of the key challenges in the field of affective computing, artificial intelligence, and human-computer interaction. In an era of increasing integration between humans and intelligent systems – from virtual assistants and educational tools to telemedicine and security applications – the ability of machines to interpret emotional cues accurately is crucial for natural and effective communication.

Among various modalities, facial expression analysis has proven to be one of the most direct and universal ways of decoding emotions. The human face conveys rich information through subtle changes in the eyes, eyebrows, and mouth, which can be mapped to emotional states using systems such as the Facial Action Coding System (FACS). Advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved the ability to extract visual features from facial images. However, despite this progress, emotion recognition systems still face challenges due to variations in lighting, occlusion, individual facial structures, and contextual factors.

Traditional appearance-based approaches often fail to capture deeper structural information or context-sensitive nuances that influence emotion. Additionally, psychological and environmental factors – such as personality traits or current stress levels – can modulate emotional expression in ways not detectable through pixels alone. This complexity calls for more robust and interpretable approaches that combine low-level visual cues with higher-level geometric or contextual descriptors.

This study focuses on the images-based component of emotion recognition and explores a hybrid approach that combines CNN-based deep visual features with manually engineered geometric features derived from facial landmarks. By integrating these complementary feature sets within a unified ML.NET-based pipeline, the research aims to enhance recognition accuracy while maintaining interpretability.

Thus, the relevance of the topic lies in the growing demand for emotionally intelligent systems and the scientific need to improve robustness and effectiveness in facial emotion recognition models.

2. Literature review and problem statement

Emotion recognition is a key task in affective computing, allowing intelligent systems to interpret and respond to human emotions through multiple channels, including visual, auditory, and textual signals. Emotion recognition remains a significant area of interdisciplinary research at the intersection of artificial intelligence, computer vision, and behavioral science. Numerous recent studies focus on building multimodal systems that can recognize emotions using three primary channels: video, audio, and text. Each of these modalities contributes unique information about a person's emotional state, and their combination allows for a more robust and accurate analysis.

Audio-based methods analyze tone, pitch, volume, and rhythm to determine emotional states. According to Mehrabian, up to 38% of emotional communication is conveyed through voice intonation. This finding underscores the importance of the auditory modality in overall emotion analysis [1]. Deep learning methods such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are widely applied to process time-series audio data. These models successfully capture emotional dynamics and temporal dependencies in human speech [2].

The video modality is often considered the most expressive non-verbal communication channel. Facial expressions, such as movements of the eyes, eyebrows, and mouth, have long been established as reliable indicators of emotions. Automated emotion detection based on video data commonly employs computer vision algorithms such as the FACS. This framework maps facial muscle activity to specific emotions and enables the systematic annotation of expressive behavior [2]. CNN-based models like `ResNet` or Visual Geometry Group (VGG) have demonstrated effectiveness in classifying basic emotions using large datasets. Moreover, recent work has explored complex emotion recognition even in cases where verbal and voice signals are unavailable [2].

Text-based emotion recognition has gained momentum due to the growing volume of written communication through emails, chats, and social media. Natural Language Processing (NLP) models such as Bidirectional Encoder Representations from Transformers (BERT) and LSTM analyze both word semantics and syntax. This also consider context to reveal emotional undertones in textual data [3]. These models are capable of identifying complex, subtle emotions in unstructured text formats.

Despite significant progress in unimodal systems, most current approaches still process video, audio, and text independently. As a result, they fail to leverage the complementary relationships that exist between these modalities. Moreover, recent findings highlight the influence of contextual factors – such as weather conditions, external events, or personal psychological traits – on emotional expression. These influences are often implicit, making them difficult to detect yet highly impactful on the accuracy of emotion prediction [4]. The phenomenon referred to as the "bird outside the window" illustrates how small, seemingly unrelated events or environmental factors can shift emotional tone [5]. Studies confirm that such contextual effects, though often overlooked in current technical models, play a meaningful role in shaping emotional communication [4, 5].

Traditional research in emotion recognition follows two primary paradigms. Appearance-based approaches rely on CNNs that extract features directly from raw image pixels. These methods automatically capture complex visual patterns but are often sensitive to external conditions, such as illumination or facial orientation. Moreover, CNNs, while powerful, frequently act as “black-box” models, providing limited interpretability regarding the specific facial cues driving their decisions. In contrast, geometry-based approaches describe facial configuration using numerical relationships between predefined landmarks. By measuring distances and angles between facial points – such as eye spacing, mouth width, or brow elevation – these methods produce interpretable geometric descriptors. Such descriptors are invariant to lighting changes and provide insight into which specific muscle movements correlate with particular emotional states. Nevertheless, geometry-based systems alone often fail to capture texture-level information or micro-expressions that are important for distinguishing subtle affective differences.

Recent literature highlights the increasing interest in hybrid emotion recognition models that combine visual and geometric representations. Studies demonstrate that integrating CNN-derived features with handcrafted geometric descriptors enhances model robustness and interpretability. For instance, feature fusion methods have been applied in medical image analysis and facial biometrics, showing that heterogeneous feature spaces yield better generalization across datasets and demographic variations.

However, despite promising results, most implementations rely on Python-based deep learning frameworks such as TensorFlow or PyTorch, limiting accessibility for integration into enterprise .NET systems. Given this context, the present research aims to evaluate whether ML.NET, Microsoft’s native machine learning framework, can be effectively used to construct such hybrid pipelines. ML.NET provides functionality for both image-based deep learning and traditional feature-based algorithms, allowing seamless fusion of CNN embeddings and geometric inputs in a unified environment. This approach is particularly valuable for software ecosystems built on C#, enabling direct deployment of emotion recognition models without external dependencies.

Beyond purely visual information, researchers increasingly recognize that emotion perception is influenced by broader contextual factors. While facial expressions are the most immediate cues, emotions do not exist in isolation from the environment or personal characteristics. Psychological studies have shown that factors such as personality traits, mood, fatigue, or even weather conditions can significantly alter emotional expression and perception. This phenomenon is referred to in the present study as the “bird outside the window” effect – symbolizing subtle, external influences that may not be visible in the image yet still affect emotional state and facial appearance [5].

Contextual influences can be categorized as external or internal. External factors include environmental variables such as temperature, brightness, or situational events that modulate affective response. Internal factors refer to personality characteristics, emotional predisposition, and momentary psychological state. For example, individuals with higher levels of stress or anxiety may display reduced facial mobility, whereas relaxed individuals might show more pronounced expressions. Similarly, environmental brightness or social context may change the expressiveness of facial gestures captured in the dataset.

These considerations highlight the importance of developing models capable of interpreting not only static visual data but also implicit contextual cues. Although the present study focuses primarily on static facial images, its conceptual framework acknowledges that emotion recognition in real-world conditions benefits from considering multimodal inputs such as audio tone or linguistic content.

To overcome these challenges, recent research proposes feature fusion techniques that integrate CNN embeddings with geometric or statistical descriptors. Such methods achieve higher accuracy by combining the expressive power of deep neural representations with the interpretability of handcrafted measurements. In practice, this means leveraging pre-trained architectures like `ResNet` or `EfficientNet` for global visual representation and augmenting them with geometric metrics that encode local facial dynamics.This hybrid paradigm aligns with broader trends in explainable artificial intelligence (XAI), which emphasize transparency and human interpretability in model decisions. By introducing explicit geometric variables tied to known facial movements, researchers can better understand how specific configurations of features correspond to particular emotions. This transparency is particularly valuable in applications such as psychology, healthcare, or education, where interpretability is essential for ethical and practical use.

Although some efforts have been made to include geometric facial features (like distances between landmarks), their integration into CNN-based pipelines remains limited. There is a notable lack of hybrid approaches that combine low-level features learned from raw images with high-level interpretable descriptors.

The analysis of current literature shows that individual modalities have been extensively explored in previous research. However, the integration of CNN-based image features with handcrafted geometric descriptors within a unified image-based recognition pipeline remains an open problem. This justifies conducting a study focused on developing a method that fuses CNN-based and geometric features to improve accuracy and interpretability in emotion recognition tasks.

In summary, the reviewed literature underscores the following research gaps:

1. Existing CNN-based models achieve high accuracy but lack interpretability and robustness under real-world conditions.
2. Geometry-based systems offer explainability but insufficient performance for subtle or mixed emotional states.
3. UFew studies successfully combine these two paradigms within production-ready frameworks such as ML.NET.

The current research addresses these limitations by exploring a hybrid emotion recognition method for static images, integrating CNN-based visual embeddings with geometric features derived from facial landmarks.

This approach aims to improve the balance between accuracy, interpretability, and deployability – laying the foundation for future multimodal emotion recognition systems capable of incorporating contextual and environmental data.

3. The aim and objectives of the study

The aim of this study is to substantiate and experimentally evaluate a hybrid approach to emotion recognition in images that combines CNN features with geometric facial descriptors.

This approach is intended to improve the accuracy and robustness of emotion classification compared to models that rely solely on visual pixel-based features.

The following task has been set to achieve this goal: to develop and experimentally validate a hybrid method that implements the proposed hybrid approach by integrating CNN-derived embeddings with handcrafted geometric features extracted from facial landmarks.

4. The study materials and methods for designing an optimized syntax concept

4.1. The object and hypothesis of the research

The object of this research is the process of automatic facial emotion recognition in image data using machine learning techniques.

The research hypothesis assumes that integrating CNN visual representations with manually engineered geometric features from facial landmarks increases classification accuracy compared with models relying solely on CNN-derived features in ML.NET.

4.2. System description

The task of recognizing emotions in an image is to determine the emotional state of a person based on image material containing pictures and, often, a voice. The main goal is to categorize the emotions expressed by a person at a certain point in time.

The input is an image to be analyzed, as well as a set of pre-selected features and feature weight.

The output is the value of the value function (emotional scores) obtained at different moments of time from the image in question.

4.3. Model description

Emotion recognition in images was performed using a linear convolution method. Each feature, such as facial expressions, is assigned a weight reflecting its contribution to the overall emotional state assessment. This model allows us to efficiently calculate an emotional score based on several parameters. For each feature, such as eye movements or facial expressions, weights are calculated separately to ensure maximum accuracy in recognizing emotions in the image.

For recognizing emotions in images, an important tool is the FACS, a system designed to encode facial muscle movements that are associated with specific emotional states. FACS allows you to analyze facial movements, which is the basis for determining emotions such as joy, anger, or sadness. Each muscle movement, or *action unit* (AU), corresponds to a specific feature and has a numerical value that reflects the intensity of this movement [6].

For example, AUs such as raised corners of the mouth or wrinkles around the eyes can be used to determine the emotion of joy. Each of these AUs receives a numerical intensity score that will be included in the overall emotional score formula [7].

The main features of FACS:

1. AU. Each unit of action corresponds to the movement of a specific facial muscle group, e.g:

1.1. AU1 – augmentation of the inner part of the eyebrows (forehead muscles).

1.2. AU2 – raising the outer part of the eyebrows.

1.3. AU6 – contraction of the muscles around the eyes, which closes the wrinkles around the eyes (typical for a sincere smile).

1.4. AU12 – an increase in the corners of the lip (also associated with a smile).

1.5. AU15 – lowering of the corners of the lips (often when expressing sadness).

2. Combinations of AU. Many emotions are displayed through a combination of several action units, e.g:

2.1. Genuine smile, also known as a "Duchenne Smile": connection of AU6 (wrinkles around the eyes) and AU12 (lifting of the corners of the lips).

2.2. Pod: AU1 + AU2 (raising eyebrows), AU5 (open eyes), AU26 (open mouth).

2.3. Anger: AU4 (narrowing of eyebrows), AU5 (eye tension), AU23 (mouth tension).

3. Intensity of single actions. FACS provides intensity levels for each action unit (for example, from A to E), which describes the degree of severity of a particular facial reaction.

4. Temporal patterns:

4.1. Onset – the beginning of the action unit (when the muscle starts moving).

4.2. Apex – peak of movement (maximum intensity of facial movement).

4.3. Offset – the end of the action unit (when the muscle returns to its initial state).

Adjusting the weights can be done based on empirical data or by training a model that adapts the weights based on the results. This can be done using optimization algorithms, such as gradient descent, which adjusts the weights according to the accuracy of emotion classification based on test data. Alternatively, weights can also be set by experts who analyze the significance of features or through classified data using methods like Bayes classification, which assigns probabilities to features based on prior knowledge.

In this study, the features $\boldsymbol{F_j}$ of the FACS system, which represent specific facial muscle activations, are calculated from geometric measurements of facial landmarks. These landmarks correspond to characteristic points on the face (such as the corners of the eyes, mouth, and eyebrows), and the derived features capture the intensity and configuration of facial expressions.

The weights of each attribute are considered equivalent across all features and are computed using the formula:

$$W_j = \frac{1}{m}, \tag{1}$$

where W_j – weights of each feature F_j, m – amount of features, with $\sum\limits_{j=1}^{m} W_j = 1$.

The concept of an emotional score for image analysis corresponds to an emotion at a specific moment for a given image. The emotional score is calculated using a linear convolution [8] using the formula:

$$E_{image} = \sum_{j=1}^{m} \left(F_j \cdot W_j\right), \tag{2}$$

where F_j – FACS features are calculated for the image, W_j– the respective weights of each feature, showing their contribution to the overall emotional score for this image.

In the event of a situation where several emotions are identified at once at one time, the one with the highest emotional score is selected from all the emotions at the time.

The proposed model produces an array of emotional scores. Each element represents either the maximum or the mean emotional value computed for a particular group of images. This structure enables quantitative comparison of emotional intensity across multiple samples. For example, if a dataset contains 600 images grouped into batches of 10, the model calculates the emotional score for each batch, resulting in an array of 60 elements.

Thus, the result looks like an array in the formula:

$$\left[E_{image\ 1}, \ldots, E_{image\ 60}\right], \tag{3}$$

where $E_{image\ k}$ – emotional score, which represents the average or maximum intensity of a particular emotion for the k-th group of images ($k = \overline{1, 60}$).

This approach allows the analysis of a person's emotional state across multiple images, which is especially useful when studying the dynamics of emotions within large image datasets.

It should be noted that this model requires control over the accuracy of emotion recognition.

4.4. Suggested algorithm

The following approaches are proposed to improve the accuracy of emotion recognition in image. Important step is to execute *data preprocessing* that aims to:

1. *Extraction of key facial points:* using algorithms to detect key facial features (eyes, mouth, eyebrows) enhances facial expression recognition.

2. *Changing noise in the data:* filtering and normalization of the image (blur removal, education normalization) improve the quality of the data for analysis.

Processing involves several basic steps (Fig. 1) to prepare images for accurate emotion analysis:

STEP 1: Face detection. The first step is to identify the face region in each image. This allows you to focus on analyzing facial characteristics, ignoring other elements of the image that are not relevant to emotion recognition. For face detection, algorithms such as Haar Cascades or Histogram of Oriented Gradients (HOG) [9] in combination with Support Vector Machines (SVM) [10] are often used.

STEP 2: Normalization of face images. After face detection, each image goes through a normalization process. This is important because faces can have different sizes, angles, and positions in the frame. Normalization includes scaling the face to the same size and aligning the orientation so that all images have the same format.

STEP 3: Extraction of facial features using FACS. In the next step, the FACS system identifies specific facial movements, or so-called AUs, which are indicators of certain emotions.

STEP 4: Training the model using CNN [11]. After data processing, the obtained features are fed to a neural network, usually a CNN. In the case of emotion recognition, CNN can learn to detect facial features associated with specific emotional states, such as raised corners of the mouth for joy or furrowed brows for anger.

STEP 5: Empirical experiments to tune the model. To tune the model that calculates the emotional score, it is important to conduct a series of empirical experiments to optimize its accuracy and efficiency. The basis of this process is testing on well-known emotion recognition databases, such as `Cohn-Kanade Plus` (`CK+`) [12] and `Facial Expression Recognition 2013` (`FER2013`) [13].

STEP 6: Final tuning and optimization. After training, the model is optimized to improve accuracy and performance by integrating the algorithm with an adaptive approach.

All the steps are visualized on Fig. 1.

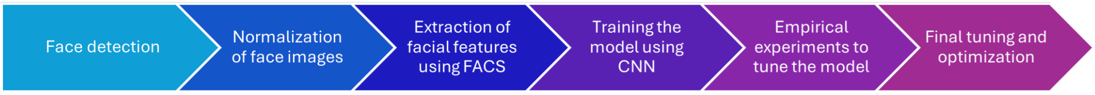
Face detection
Normalization of face images
Extraction of facial features using FACS
Training the model using CNN
Empirical experiments to tune the model
Final tuning and optimization

Fig. 1. Learning and data processing algorithm

4.5. Research workflow

Most academic studies employ Python-based deep learning frameworks, including `TensorFlow` and `PyTorch`. The present research, however, aims to evaluate the potential of ML.NET, Microsoft's native machine learning framework. ML.NET supports direct integration of ML models into C# applications. Such integration is especially valuable for maintaining consistency within .NET-centric software infrastructures.

The rationale behind this choice lies in the increasing industrial adoption of C# for enterprise software systems, where direct integration of machine learning components without external dependencies is of significant practical value.

4.5.1. Justification for using ML.NET

ML.NET offers a cohesive environment that supports both classical and deep learning algorithms, enabling the construction of hybrid pipelines. The framework provides:

– seamless integration with .NET applications, ensuring full compatibility with enterprise-grade architectures;

– pre-trained image classification architectures, including `ResNet`, `Inception` and `MobileNet`, which facilitate transfer learning without the need for GPU-specific Python libraries;

– support for custom feature engineering, allowing the injection of domain-specific numerical features (e.g., geometric measurements derived from facial landmarks) into deep learning pipelines;

– unified data processing workflows, which integrate data loading, transformation, training, and evaluation within a single environment.

This configuration makes ML.NET particularly suitable for research scenarios where reproducibility, performance portability, and ease of deployment in production systems are key objectives [14].

The goal of our research is to explore whether combining CNN with geometry-based features extracted from facial landmarks leads to improved emotion classification accuracy, using only the tools available within the ML.NET ecosystem.

4.5.2. Justification for using ResNet

The choice of `ResNetV2_101` was motivated by its proven efficiency in capturing hierarchical visual features through residual connections, which mitigate vanishing gradient problems in deep networks. This architecture contains over one hundred convolutional layers interconnected by identity shortcuts, allowing the network to learn both low-level and high-level spatial representations critical for distinguishing subtle emotional cues.

ResNet introduced *residual connections* that enable the training of very deep networks while mitigating vanishing gradient problems, thus achieving stable convergence and improved generalization performance [11]. Its hierarchical convolutional layers effectively capture both low-level (edges, textures) and high-level (semantic and structural) facial patterns, which are crucial for emotion recognition [13].

Compared to more recent architectures, ResNet maintains a favorable balance between accuracy, interpretability, and compatibility with existing ML frameworks, including ML.NET. According to [13], ResNet-based models trained on the FER-2013 dataset demonstrate competitive results across various convolutional depths while retaining model stability and explainability. Additionally, its structure allows for seamless integration with handcrafted geometric features, which can be concatenated at the feature-fusion stage without major architectural modifications.

Although newer architectures such as EfficientNet and Vision Transformer (ViT) have shown improvements in efficiency and global contextual understanding [4], their deployment in ML.NET remains less direct. EfficientNet achieves higher accuracy-to-computation ratios due to compound scaling [4], while ViT captures long-range spatial dependencies through self-attention mechanisms [4]; however, both require larger datasets and more complex preprocessing.

The EfficientNet family of models introduces compound scaling, which jointly optimizes network depth, width, and input resolution to achieve a better balance between accuracy and computational cost. EfficientNet V2 outperformed several conventional convolutional networks, including ResNet variants, on the FER-2013 dataset, demonstrating superior classification accuracy and faster convergence during training [15]. These findings suggest that EfficientNet represents a strong candidate for lightweight yet high-performance emotion recognition pipelines within the ML.NET environment, particularly where computational resources are limited.

Similarly, ViT architectures replace convolutional operations with self-attention mechanisms, enabling them to model long-range dependencies across facial regions more effectively. ViT-based models outperform CNNs in robustness to occlusion and variations in facial pose or viewing distance, while maintaining competitive computational efficiency [16]. Integrating transformer-based backbones into the existing hybrid pipeline through ONNX Runtime could enhance contextual modeling and improve emotion recognition accuracy, offering a valuable direction for the further evolution of this research.

In this study, ResNetV2_101 was therefore selected as a baseline architecture due to its mature implementation within ML.NET, robust transfer-learning performance on facial datasets, and compatibility with the hybrid pipeline combining CNN-based embeddings and geometric facial descriptors. This foundation ensures reproducibility, interpretability, and computational feasibility within the constraints of the current experimental environment [14].

In future research, it would be reasonable to extend the current experiments by evaluating more recent architectures such as EfficientNet and ViT. These models offer potential advantages in terms of computational efficiency and global context modeling, respectively. EfficientNet could provide improved accuracy with lower resource consumption through compound scaling, while ViT might enhance recognition robustness by capturing long-range spatial dependencies across facial regions. Integrating these architectures within the ML.NET environment via ONNX Runtime represents a promising direction for further exploration and comparison.

4.5.3. Experimental setup

All experiments were conducted on a Windows 11 workstation equipped with an Intel i7-13700H processor, 32 GB of RAM, and an NVIDIA RTX 4060 GPU. The ML.NET version used was 4.0.2, with Microsoft.ML.Vision and Microsoft.ML.Dnn libraries for image-based training.

The implementation was done in C# 12 using Visual Studio 2022.

The dataset structure followed the FER-2013 format and contained approximately 28,000 grayscale facial images resized to 48×48 pixels. The emotion labels covered seven standard categories: angry, disgusted, fearful, happy, neutral, sad, surprised.

The dataset contained approximately 35,000 facial images distributed across seven emotion categories: anger (5,000), disgust (4,500), fear (5,200), happiness (6,000), sadness (5,000), surprise (4,800), and neutral (4,500). The class distribution was moderately balanced, with the happiness class being slightly overrepresented.

The dataset was divided into 80% training and 20% validation subsets. Data normalization was performed using the `ExtractPixels` transformer, which converted image tensors into normalized byte arrays with mean scaling.

All experiments were repeated from three up to five times to ensure stability of the results, and average metrics were reported. The same random seed was used for reproducibility.

Below, the experimental workflow is presented and was used to evaluate this hypothesis.

STEP 1: Baseline CNN Model. At the first stage, a baseline CNN was trained using ML.NET's `ImageClassificationTrainer`, configured to use the `ResNetV2_101` architecture as the backbone. Transfer learning was employed – the model reused pre-trained weights and fine-tuned them on the target emotion dataset. The training ran for 30 epochs, with early stopping enabled via the `EarlyStoppingCriteria` (in `ImageClassificationTrainer.Options`) parameter to prevent overfitting.

During the training process, ML.NET automatically managed the learning rate schedule, batch normalization, and checkpoint saving. The model's performance improved steadily up to epoch 28–31, after which accuracy plateaued. The final checkpoint represented the baseline performance and served as a benchmark for subsequent hybridization experiments. The model employed a pre-trained `ResNetV2_101` convolutional architecture, which was subsequently fine-tuned on grayscale facial images derived from a `FER-2013`-like dataset.

The dataset used for fine-tuning consisted of facial images representing seven universal emotion categories – angry, disgusted, fearful, happy, neutral, sad, and surprised – which are widely adopted in affective computing research. The dataset structure followed the `FER-2013` format, originally introduced for large-scale facial expression recognition challenges. Each image was converted to grayscale to minimize the influence of illumination variance and color bias, focusing the learning process on structural facial features rather than chromatic attributes.

To ensure consistency, all images were standardized to a uniform resolution of 48×48 pixels, normalized by pixel intensity, and aligned based on eye coordinates. Such preprocessing guaranteed scale and orientation invariance, facilitating stable convergence during model fine-tuning. The resulting input tensors were used to train the classification layers of `ResNetV2_101` within the ML.NET framework, adapting the model to domain-specific emotion data.

STEP 2: Performance evaluation. The baseline model achieved an average accuracy of 63% on the validation dataset, which is consistent with comparable implementations in Python-based frameworks using `ResNet`. However, the analysis of Cross-Entropy Loss revealed that the model struggled with fine-grained distinctions between similar emotional expressions, particularly between fear and surprise, or neutral and sad. These observations motivated the introduction of geometric features to capture spatial dependencies that pure CNN embeddings might overlook. Accuracy and loss metrics were monitored during all training runs and logged to the console using ML.NET's built-in callback function for transparency of convergence behavior.

STEP 3: Geometric feature extraction via facial landmarks. The next phase involved the extraction of handcrafted geometric descriptors using the Dlib facial landmark detector. Each facial image was processed to identify 68 characteristic points corresponding to eyes, eyebrows, nose, mouth, and jawline. From these landmarks, we computed several spatial distances and proportional relationships that correspond to emotion-relevant facial cues, including:

- eye distance (between landmarks 36–45);
- mouth width (between 48–54);
- brow distance (between 21–22).

These metrics were normalized to the interocular distance to minimize the influence of face size or image scaling. The resulting feature vectors contained six continuous variables per image and were later merged with CNN-based embeddings for hybrid classification.

STEP 4: Hybrid feature fusion pipeline. The final stage of the experiment focused on constructing a hybrid ML.NET pipeline that fused deep and geometric representations. This was achieved by:

1. Loading image data through `LoadImages` and converting them into pixel tensors using `ExtractPixels`.
2. Feeding the tensors into a pre-trained `ResNetV2_101` model using the `DnnFeaturizeImage` transformer to generate 2048-dimensional CNN embeddings.
3. Concatenating these embeddings with six engineered geometric features (e.g., mouth width, brow height, eye-to-mouth distance) using the `Concatenate` transformer.
4. Normalizing the combined feature vector with `NormalizeMinMax`.
5. Training a classification head based on the Stochastic Dual Coordinate Ascent (SDCA) algorithm with a maximum entropy loss function.

The full pipeline was serialized into a model file for later deployment, ensuring reproducibility and compatibility with .NET inference APIs.

STEP 5: Evaluation of the hybrid method. The hybrid model was trained under the same conditions as the baseline and evaluated on the same validation set to ensure fair comparison.

It achieved an average accuracy of 67%, marking an improvement of approximately 4 percentage points over the baseline model.

Cross-Entropy Loss decreased more rapidly during training, suggesting that the inclusion of geometric descriptors facilitated faster convergence.

Additionally, confusion matrix analysis revealed that the hybrid model more accurately distinguished between visually similar emotions (fear vs. surprise) and reduced misclassifications for low-intensity expressions (neutral vs. sad).

STEP 6: Validation and reproducibility. To ensure reliability, the experiments were repeated with shuffled data partitions, confirming consistent accuracy improvement across all runs (±0.04). All model configurations, feature extraction parameters, and transformation pipelines were saved and can be fully reproduced on any Windows environment using ML.NET 4.0+. This reproducibility aspect underlines one of ML.NET's strongest advantages – deterministic model training, achieved through managed execution and fixed random seed initialization.

4.5.4. Summary of the workflow

The described workflow demonstrates a complete, end-to-end emotion recognition pipeline within ML.NET, from image preprocessing and feature extraction to model training and evaluation.

The hybrid approach not only improved accuracy but also offered better interpretability through explicit geometric features.

This confirms the feasibility of conducting deep learning research entirely within the .NET ecosystem, bridging the gap between academic experimentation and industrial deployment.

5. Results of investigating of method for combining CNN-based features with geometric facial descriptors

This section presents the results of the experimental study. This section presents the results of the experimental evaluation of the hybrid method developed to implement the proposed hybrid approach for emotion recognition in images.

The proposed model combines deep learning techniques based on convolutional neural networks (CNNs) with handcrafted geometric features derived from facial landmarks.

To evaluate the performance of the proposed model, we compared the Baseline and Hybrid approaches in terms of accuracy and loss dynamics during training. The results of the experiments are presented below.

For the Baseline model, accuracy also increases initially and later stabilizes, but at a lower level compared to the Hybrid approach. Cross-Entropy decreases over time, though the overall accuracy remains lower, highlighting the limited capacity of the Baseline model. These trends are shown in Fig. 2.

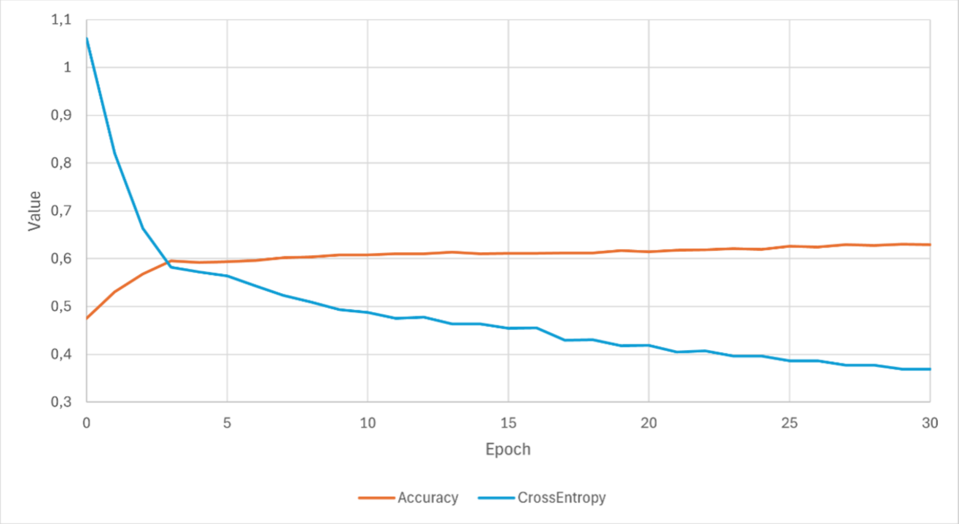

Fig. 2. Accuracy and Cross-Entropy dynamics (Baseline model)

The relationship between Accuracy and Cross-Entropy during training is presented for the Hybrid model. Accuracy increases steadily across epochs, while Cross-Entropy decreases rapidly at the beginning and then continues to decline gradually, indicating successful optimization and reduced error. These dynamics are shown in Fig. 3.

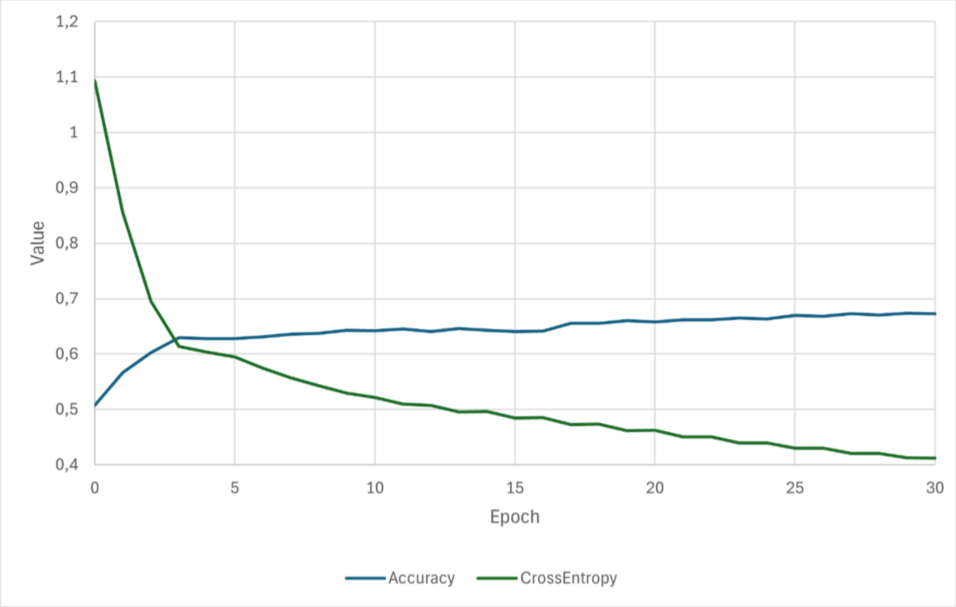

Fig. 3. Accuracy and Cross-Entropy dynamics (Hybrid model)

The hybrid model demonstrated a significant improvement in emotion classification accuracy, outperforming the baseline CNN by ~4% points on the validation set. This confirms the hypothesis that combining geometric features with learned image representations enhances emotion recognition performance. The comparison of two approaches for facial emotion recognition are shown on Fig. 4:

– utilizing CNN-based on the `ResNetV2_101` architecture within the ML.NET framework (Baseline `ResNet` Model);

– combining CNN-based embeddings with geometric features extracted from facial landmark analysis using the `Dlib` library (Hybrid Model).

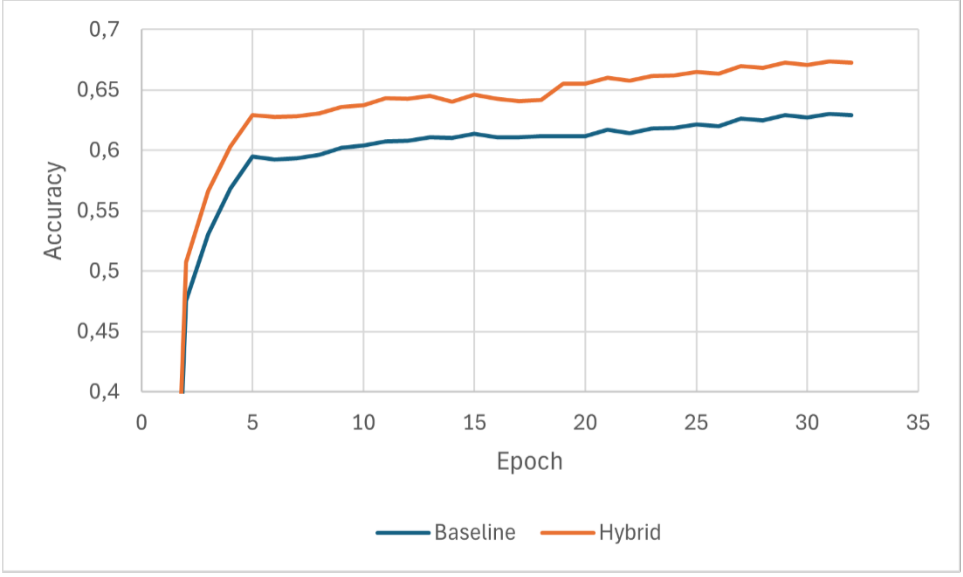

Fig. 4. Accuracy comparison between Baseline and Hybrid models

This shows the accuracy progression across epochs for two models: Baseline (blue) and Hybrid (orange):

– hybrid model consistently outperforms the Baseline, reaching higher accuracy values at each stage of training;

– both models demonstrate rapid growth in accuracy during the first few epochs, after which the growth slows down and stabilizes.

The progression of validation accuracy over 30 training epochs highlights the consistent advantage of the hybrid model over the baseline CNN. Accuracy is calculated by standard `MulticlassClassificationMetrics` metric in ML.NET, calculated by formula:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + NP_i}, \qquad (4)$$

where N – the number of classes in a multiclass classification task, TP_i – the number of correctly classified examples of class i, FP_i – the number of examples that were incorrectly assigned to the class i, NP_i – the number of examples of class i that were incorrectly assigned to other classes.

This hybrid methodology not only improves performance but also offers interpretability, as geometric distances relate directly to human-understandable facial movements. The approach is modular, lightweight, and deployable in local or edge environments using .NET technologies.

The scientific novelty lies in integrating a CNN-based visual model with manually extracted geometric descriptors, which:

– enhances model accuracy;

– maintains computational efficiency;

– and enables deployment in resource-constrained environments (e.g., edge devices).

The methodology is reproducible: all data transformations, model configurations, and evaluation metrics are standard and documented. The hybrid approach can be reused with other pre-trained architectures and adapted to other domains.

6. Discussion of the results of the combined method of facial emotion recognition

The obtained results confirmed the hypothesis regarding the effectiveness of a hybrid approach to emotion recognition in images, which combines CNNs with geometric features extracted from facial landmarks. Compared to the baseline model using only pixel-level representations (`ResNetV2_101` in ML.NET), the proposed hybrid model demonstrated an accuracy improvement of approximately 4% points. The proposed hybrid configuration not only improved recognition accuracy but also enhanced interpretability, as the inclusion of geometric cues provided anatomically meaningful insights that the baseline CNN failed to capture. The hybrid approach reached an accuracy confirms the hypothesis that combining heterogeneous features (deep visual and interpretable geometric) leads to better generalization and more robust emotion recognition.

This indicates the importance of incorporating spatial facial features such as eye distance, mouth width, and eyebrow position – elements that might not be fully captured even by powerful CNN architectures.

This improvement can be attributed to the complementarity of different feature types. CNNs effectively extract abstract, high-dimensional visual representations. However, they may overlook precise spatial configurations of the face that are critical for distinguishing emotions. By explicitly incorporating features such as eye distance, mouth width, and eyebrow separation, the model gains interpretable cues linked to facial anatomy. These cues correspond to emotion-specific expressions, such as raised eyebrows for surprise or compressed lips for anger.

Furthermore, the enhanced performance indicates the limitations of using pre-trained architectures alone. Models such as `ResNetV2_101` may fail to capture all the information required for accurate emotion classification. This limitation becomes especially evident when working with real-world data affected by noise, occlusions, or varied lighting conditions. The geometric descriptors act as a regularizing factor, providing stable references that anchor the classification even when visual quality is imperfect.

Conclusions

The results of the study confirm the effectiveness of the hybrid method developed to implement the proposed hybrid approach to facial emotion recognition in static images.

The method integrates deep convolutional embeddings obtained from a pre-trained `ResNetV2_101` model within the ML.NET framework with handcrafted geometric features extracted from facial landmarks.

This finding validates the initial hypothesis that combining heterogeneous features – deep visual and interpretable geometric – enhances generalization and robustness in emotion recognition. By incorporating geometric features that correspond to human facial anatomy (such as raised eyebrows for surprise or compressed lips for anger), the model obtains interpretable and stable cues that strengthen its predictive capability.

There was done the following task: a hybrid method was developed and experimentally validated, implementing the proposed hybrid approach by integrating CNN-derived embeddings with handcrafted geometric features extracted from facial landmarks. The experimental evaluation demonstrated 4% improvement in classification accuracy compared to the baseline model. The baseline model achieved an accuracy of approximately 63%. In comparison, the hybrid approach reached about 67% accuracy on the validation set. This result confirms the hypothesis that combining heterogeneous features – deep visual and interpretable geometric – enhances generalization and strengthens the robustness of emotion recognition. By explicitly integrating interpretable features tied to human facial anatomy (e.g., raised eyebrows for surprise, compressed lips for anger), the model gains robust and semantically meaningful cues that enhance recognition accuracy.

However, the achieved accuracy ~67% on the validation set still leaves room for improvement. Potential future enhancements include:

– increasing the size and diversity of the training dataset;
– exploring more advanced CNN architectures (e.g., `EfficientNet`);
– expanding the set of geometric features;
– optimizing the model using modern ensemble or regularization techniques.

Moreover, the results indicate that pre-trained architectures such as `ResNetV2_101` have certain limitations when applied independently. Used alone, they may fail to capture all the information required for accurate emotion classification. This limitation becomes more pronounced under real-world conditions affected by noise, occlusions, or variable lighting. In such cases, geometric descriptors serve as stabilizing factors. They anchor the classification process and compensate for imperfections in image quality.

Future work will expand this research toward testing alternative architectures and improving the hybrid pipeline. In particular, evaluating `EfficientNet` and transformer-based models through ONNX integration will serve as the next step in advancing the proposed method. Such continuation of the study is expected to enhance both accuracy and generalization while maintaining interpretability within the ML.NET environment.

The developed hybrid method thus fully achieves the stated research objective: it substantiates the feasibility of the hybrid approach and provides an experimentally validated implementation that improves accuracy and robustness in image-based emotion recognition.

References

[1] A. Mehrabian, *Silent Messages.* Wadsworth Publishing Company, 1971.

[2] T. K. Arora, “Optimal facial feature based emotional recognition using deep learning algorithm,” 2022. https://doi.org/10.1155/2022/8379202.

[3] F. Ye, “Emotion recognition of online education learners by convolutional neural networks,” 2022. https://doi.org/10.1155/2022/4316812.

[4] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, “MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild,” in *Proc. 30th ACM Int. Conf. Multimedia (MM’22)*, Lisbon, Portugal, Oct. 2022. https://doi.org/10.1145/3503161.3548190, pp. 1–9.

[5] J. J. A. Denissen, L. Butalid, L. Penke, and M. A. G. van Aken, “The effects of weather on daily mood: a multilevel approach,” *Emotion*, vol. 8, no. 5, pp. 662–667, Oct. 2008. https://doi.org/10.1037/a0013497.

[6] P. Ekman and W. V. Friesen, *Facial Action Coding System.* American Psychological Association (APA), 1978. https://psycnet.apa.org/doi/10.1037/t27734-000.

[7] O. Gavrylenko and L. Zinchenko, “Detection of human emotions using machine learning tools,” in *Proc. Global Learning: Problems, Causes, Solutions, and Theories*, 2024. https://eu-conf.com/wp-content/uploads/2024/10/GLOBAL-LEARNING-PROBLEMS-CAUSES-SOLUTIONS-AND-THEORIES.pdf.

[8] B. Xiao, H. Zhu, S. Zhang, Z. OuYang, T. Wang, and S. Sarvazizi, “Gray-related support vector machine optimization strategy and its implementation in forecasting photovoltaic output power,” 2022. https://doi.org/10.1155/2022/3625541.

[9] C. Rahmad, R. A. Asmara, D. R. H. Putra, I. Dharma, H. Darmono, and I. Muhiqqin, “Comparison of viola-jones haar cascade classifier and histogram of oriented gradients (hog) for face detection,” in *IOP Conference Series: Materials Science and Engineering*, vol. 732, no. 1, 2020. https://doi.org/10.1088/1757-899X/732/1/012038.

[10] R. Bridgelall, “Tutorial on support vector machines,” *Preprints*, 2022. https://doi.org/10.20944/preprints202201.0232.v1.

[11] P. Purwono, A. Ma’arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. ul Haq, “Understanding of convolutional neural network (CNN): A review,” vol. 2, no. 4, pp. 739–748.

[12] A. E. Nada Elgendy and T. Päivärinta, “Decas: a modern data-driven decision theory for big data and analytics,” vol. 31, no. 4, pp. 337–373.

[13] H. Song, “Comparison of different depth of convolutional neural network deep and shallow cnn comparison based on fer-2013,” in *8th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2023)*, vol. 41, 2023. https://doi.org/10.54097/hset.v41i.6746.

[14] Özge Cömert and N. Yücel, “Review mate: A cutting-edge model for analyzing the sentiment of online customer product reviews using ml.net,” vol. 5, no. 2, pp. 74–88.

[15] C. Qian, J. A. Lobo Marques, A. R. de Alexandria, and S. J. Fong, “Application of multiple deep learning architectures for emotion classification based on facial expressions,” vol. 25, no. 5, 2025. https://doi.org/10.3390/s25051478.

[16] C. C. Marcos Rodrigo and N. García, “Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks,” vol. 14, no. 21392, 2024. DOI:https://doi.org/10.1038/s41598-024-72254-w.

УДК 519.688; 004.89; 004.9

МЕТОД ПОЄДНАННЯ CNN-ОЗНАК З ГЕОМЕТРИЧНИМИ ХАРАКТЕРИСТИКАМИ ОБЛИЧЧЯ ДЛЯ РОЗПІЗНАВАННЯ ЕМОЦІЙ

Людмила Зінченко
https://orcid.org/0009-0009-3956-5854

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Київ, Україна

У дослідженні представлено метод поєднання візуальних ознак, отриманих із згорткових нейронних мереж (CNN), із геометричними дескрипторами обличчя для підвищення точності розпізнавання емоцій на статичних зображеннях. Метод інтегрує глибокі згорткові вектори ознак, отримані з попередньо натренованої моделі `ResNetV2_101` у середовищі ML.NET, із вручну розрахованими геометричними параметрами, визначеними на основі ключових точок обличчя. Для експериментів використано відкриті набори даних, що містять зображення облич із відповідними емоційними категоріями. На першому етапі глибокі візуальні ознаки отримано з попередньо натренованої мережі, а на другому – на основі 68 ключових точок обличчя обчислено метричні та пропорційні характеристики (відстань між очима, ширину рота, висоту брів тощо). Отримані візуальні та геометричні ознаки об'єднано в єдиний простір і класифіковано за допомогою багатокласової лінійної моделі. Гібридний метод продемонстрував покращення точності приблизно на 4% у порівнянні з базовою CNN-моделлю, що використовувала лише піксельні ознаки (з 63% до 67%). Це підтвердило, що поєднання гетерогенних ознак підвищує узагальнювальну здатність і стійкість моделі. Результати показали, що геометричні дескриптори стабілізують процес класифікації, компенсуючи вплив шумів, перекриттів і варіацій освітлення. Розроблений програмний код ML.NET демонструє можливість інтеграції інтерпретованих геометричних ознак із глибокими векторами ознак безпосередньо у середовищі C#. Наукова новизна полягає у створенні інтерпретованої гібридної моделі, що підвищує надійність класифікації та зберігає сумісність із .NET-орієнтованими застосунками. Подальші дослідження спрямовуватимуться на розширення моделі до мультимодального аналізу емоцій, який поєднує мовні, жестикуляційні та фізіологічні сигнали для глибшого розуміння емоційних станів. Також гібридна модель може бути використана як діагностичний інструмент у психологічних і поведінкових дослідженнях.

Ключові слова: розпізнавання емоцій, ключові точки обличчя, згорткові нейронні мережі, .NET фреймворк, метод поєднання ознак.