ISSN 2007-9737

Facial Expressions Recognition in Sign Language Based on a Two-Stream Swin Transformer Model Integrating RGB and Texture Map Images

Lourdes Ramirez-Cerna[1,*], Jose Rodriguez-Melquiades[2], Edwin Jonathan Escobedo-Cardenas[3], Guillermo Camara-Chavez[4], Dayse Garcia-Miranda[5]

[1] Universidad Nacional de Trujillo,
Escuela de Posgrado,
Peru

[2] Universidad de Nacional de Trujillo,
Departamento de Informática,
Peru

[3] Universidad de Lima,
Carrera de Ingeniería de Sistemas,
Peru

[4] Universidade Federal de Ouro Preto,
Departamento de Ciência da Computação,
Brazil

[5] Universidade Federal de Ouro Preto,
Departamento de Letras,
Brazil

{lramirezc, jrodriguez}@unitru.edu.pe, eescobed@ulima.edu.pe, {guillermo, dayse.miranda}@ufop.edu.br

Abstract. The study of facial expressions in sign language has become a significant research area, as these expressions not only convey personal states, but also enhance the meaning of signs within specific contexts. The absence of facial expressions during communication can lead to misinterpretations, underscoring the need for datasets that include facial expressions in sign language. To address this, we present the Facial-BSL dataset, which consists of videos capturing eight distinct facial expressions used in Brazilian Sign Language. Additionally, we propose a two-stream model designed to classify facial expressions in a sign language context. This model utilizes RGB images to capture local facial information and texture map images to record facial movements. We assessed the performance of several deep learning architectures within this two-stream framework, including Convolutional Neural Networks (CNNs) and Vision Transformers. In addition, experiments were conducted using public datasets such as CK+, KDEF-dyn, and LIBRAS. The two-stream architecture based on the Swin Transformer model demonstrated superior performance on the KDEF-dyn and LIBRAS datasets and achieved a second-place ranking on the CK+ dataset, with an accuracy of 97% and an F1-score of 95%.

Keywords. Facial expressions in sign language, RGBD data, texture map images, two-stream architecture, swin Transformer.

1 Introduction

According to data from the World Health Organization [59], approximately 5% (430 million) of

doi: 10.13053/CyS-29-2-5119

people worldwide have disabling hearing loss, a number projected to be 700 million people by 2050. Deaf people who cannot hear or have limited hearing rely on sign language to communicate with others. Unfortunately, there is a communication gap between hearing and hearing-impaired individuals due to the lack of knowledge of sign language and interpreters in the media, public and private institutions, and other areas [27].

Sign language is defined as nonverbal communication, where a sign is the basic unit that consists of non-manual parameters such as movement of the face, eyes, head and torso; and manual parameters such as configuration, orientation, location and movement of the hands [25]. Sign language is similar to spoken language because it presents grammatical structures that vary depending on the country or culture where it is used [52]. There are various sign languages in the world, such as American Sign Language, Brazilian Sign Language, Peruvian Sign Language, among others [14]. Additionally, both oral language and sign language share similarities in prosody. In oral language, individuals naturally vary their tone of voice, volume, and pause, while in sign language, facial expressions, body postures, and movements are used during communication. When interacting with each other within an oral environment, hearing-impaired individuals adopt forms of communication that involve gestural movements and facial expressions, utilizing visual and other senses [10].

Facial expressions of hearing-impairment individuals convey information about the emotional state of a person without any barriers, just as oral language does [30]. They are also part of the lexicon, grammar, syntax, and semantics of sign language, where they serve to emphasize or intensify signs as needed. In this way, facial expressions are combinations of facial behaviors and movements performed by humans during communication [26, 49].

However, it can be challenging to understand the nuances of facial expressions in sign language when learning; along with the absence of facial expressions during communication, it can produce an incorrect interpretation of the meaning or identification of facial expressions in a particular context, it can lead to inadequate reactions or misinterpretations because it does not convey the message, potentially resulting in misunderstandings [10, 14].

Many researchers are interested in research on sign language to facilitate communication [6, 21, 29, 32, 57]. However, authors sometimes tend to ignore facial expressions in the recognition task in the literature, instead focusing on gestures or splitting the structural components of signs, such as the configuration of the hand and movement type [25]. The investigations conducted in [31], assume that a language recognition system is incomplete without considering facial expressions. Therefore, facial expressions in sign language have become an emerging area of research due to scientific advances in human-computer interaction, security, and academia.

The main contributions of this research are as follows:

— We introduce the Facial-BSL dataset, a publicly available collection of continuous videos capturing eight facial expressions in Brazilian Sign Language (BSL).

— We develop a process to generate texture map images that encode facial changes and movements during sign performance. These texture map images are computed by analyzing the distances of movements between adjacent landmarks over time in isolated videos.

— We propose a two-stream architecture that uses the Swin-Transformer model as its backbone. The model inputs consist of both RGB and texture map images.

The rest of this paper is organized as follows: Section 2 reviews related works on facial expressions in sign language, Section 3 presents the existing facial expressions in sign language datasets. In Section 4 the proposed dataset is discussed in detail. Section 5 explains our proposed methodology. Section 6 presents the experimental results and discussion conducted on the proposed dataset and other datasets related to facial expressions. Section 7 mentions the strengths, limitations and future directions about the research. Finally, Section 8 provides the conclusions.

2 Related Work

The field of facial expression recognition in sign language has received significant attention from researchers. This section presents an overview of the methodologies proposed in the literature. Initially, Friesen and Ekman, [22] developed the Facial Action Coding System (FACS) to categorize facial movements based on their appearance on the face. The system identifies 46 facial action units (FAU), encoded as group or individual facial muscle movements. Therefore, numerous research studies have been conducted on facial expressions [5, 14, 17, 27, 33, 50, 54].

Deshpande et al., [18] identified the basic facial expressions from German Sign Language videos. The methodology involved using a pre-trained model, applying preprocessing techniques, and using machine learning models with K-fold cross-validation. The experiments showed no significant difference between the MobileNet and EfficientNet models in the recognition task. Similarly, Mukushev et al., [44] conducted a study on Kazakh-Russian Sign Language (K-RSL), focusing on signs with similar manual components but different non-manual components such as eyebrow height, facial expression, and head position. The authors compared the performance of manual features, and manual features combined with non-manual features extracted from 20 signs. The results showed that the addition of non-manual features improved the results by 5%, achieving 78.2% accuracy for 20 classes and 77.9% accuracy for two classes. Additionally, Javaid and Rizvi, [28] introduced a framework for recognizing non-manual features and manual gestures in American Sign Language using a multimodal approach. The framework utilizes a modified version of the YOLOv5 model to detect faces and hands, followed by a refined C3D architecture to extract features from both regions. These features are then concatenated and fed into an LSTM network. The authors conducted experiments on the RWTH-PHONIX-WEATHER-2014T, SILFA, and PkSLMNM datasets and obtained excellent results.

Facial expressions are also considered in sign language translation, along with manual gestures and body movements to emphasize the meaning of the signs. Irasiak et al., [27] proposed the Avatar2PJM project to translate the Polish Sign Language. The framework incorporated action unit recognition for annotating facial expressions, which served as inputs for machine learning models. Liu et al., [37] studied Grammatical Facial Expressions (GFE). They used action units and facial landmarks of facial expressions as input graphs for Graph Convolutional Networks (GCN) on two datasets: BUHMAP and LSE_GFE.

The study evaluated three CNN architectures (VGG, MobilenetV2, and a custom CNN), and found that GCN was effective for GFE. Guerra et al., [25] extended the work of [49], which focused on recognizing of Brazilian Sign Language (BSL) signs through facial expressions. The researchers re-labeled videos of ten signs with one of six facial expressions that closely resembled them.

They conducted experiments to recognize facial expressions, achieving an average accuracy of 89.33% with the KNN method, while Random Forest and Support Vector Machine achieved an average accuracy of 91.33%. Similarly, Cardoso et al., [11] focused on GFEs in BSL, and proposed a framework consisting of two modules: a module that identifies hand shapes, orientations, and movements; and a grammar module that combines the outputs of the previous module to give the meaning of a composition of elements.

A Multilayer Perceptron (MLP) neural network was used to classify six classes of GFEs among the eight expressions used in BSL. Other authors explored FACS to annotate facial actions of the facial muscles, and identify new FACS in BSL related to basic emotions; they conducted experiments using various models, including CNNs, AlexNet and VGG-16, as well as a hybrid CNN+LSTM, on a video dataset of twenty-three BSL sentences [14–16].

On the other hand, it is worth mentioning that vision transformers models (ViT) have demonstrated excellent performance in tasks such as recognition, segmentation, and object detection [20, 38]. As a result, many authors have incorporated them into their research of face and facial expression recognition [1, 3, 7, 12, 36, 41, 63]. Thus, we proposed to use two ViT models as the backbone of our proposed experimental schemes.

In addition, the literature review discusses facial expressions in sign language and presents various methodologies, datasets, and techniques for analyzing RGB images captured by Kinect sensors or cameras. Each proposal differs in how it obtains input data, such as landmarks, AUs, graphs, spatio-temporal data, key points, and geometric-based features. The authors also used different pre-trained deep learning models, CNNs, and SVMs to improve recognition accuracy. However, there is a lack of research on the use of vision transformers to study facial expressions in sign language.

Finally, sign language varies across countries and cultures, which limits the available datasets. Additionally, the meaning of a word can differ when used in a sentence. Most research focuses solely on recognizing basic emotions, omitting other types of facial expression present in sign language. In some research, the authors separate facial expressions from hand movements, forcing their recognition to be independent of the original meaning. This can result in relabeling them as basic emotions. Therefore, the research on facial expressions in sign language is an ongoing field of study. So, we will explore additional types of facial expressions in sign language beyond the basic emotions.

3 Facial Expressions in Sign Language Datasets

Table 1 displays the datasets related to facial expressions in sign language that are available in the literature. Some of these datasets are available for free download, while others require contacting the authors. The datasets listed in Table 1 provide RGB videos [2, 4, 11, 14, 25, 44], facial action units (FACs) [14, 47] and facial landmarks [14, 31, 47].

Some authors labeled facial expressions in manual signs to better understand the meaning of a sign [25, 31, 44]. Meanwhile, datasets proposed by [11, 14, 47] used grammatical facial expressions with FACs in sentences or phrases to convey desired meaning in discourse. In contrast, Aran et al., [4] labeled facial expressions in videos that include basic emotions with head movements to differentiate signs with similar manual components. Alaghband et al., [2] extracted facial images of seven basic emotions from the public TV station PHOENIX, including semi-blurry facial images with different head poses, orientations, and movements. However, the authors did not define any subjects.

These datasets provide different approaches to studying facial expressions in sign language. However, some are not accessible for free due to broken download links or non-functional email contacts. Additionally, in some cases, facial expressions have been isolated from the original sign, resulting in an incomplete representation of their meaning. It is important to note that not all of these datasets have been validated by a sign language expert. Furthermore, many datasets lack a clear categorization criterion and are often based on daily usage. In most cases, the datasets have low intra-class variance and high inter-class variance. However, Silva and Severo, [14] classifies Brazilian facial expressions associated with FACs and added some facial expressions that are not documented in the literature. To address these limitations, we propose a publicly available dataset for recognizing facial expressions in Brazilian Sign Language. The dataset shows both intra-class and inter-class variation. It has been validated by a sign language expert and provides data extracted from the Microsoft Kinect V1, including RGB-D data, to support future research.

4 Proposed Facial-BSL Dataset

We propose a new facial expression sign language dataset. The proposed Facial-BSL dataset includes recordings of ten subjects performing eight facial expressions in Brazilian Sign Language (angry, laugh, surprise, yawn, full, agree, cry

Table 1. Facial expressions in sign language datasets

Dataset author	Data provided			Vocabulary	Subjects	Type of sign language	Availability
	RGB	FACs	Facial landmarks				
Aran et al. [4]	✓		✓	8 non-manual signs	11	Turkish	Contact to the authors
Rezende et al. [49]	✓			10 facial expressions of signs	1	Brazil	Contact to the authors
Kumar et al. [31]			✓	51 sign word gestures	10	India	Contact to the authors
Cardoso et al. [11]	✓			9 GFEs	1	Brazil	Contact to the authors
Alaghband et al. [2]	✓			8 facial expressions	not defined	Germany	Free to download
Mukushev et al. [44]	✓			20 signs	5	Kazakh Russian	Free to download
da Silva et al. [14]	✓	✓	✓	23 sentences	10	Brazil	Request to the authors
Porta-Lorenzo et al. [47]		✓	✓	6 GFEs	106	Spanish	Free to download

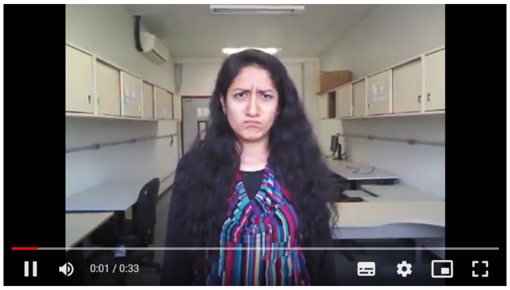

Fig. 1. An RGB video sample of angry facial expression from Facial-BSL dataset

and empty), showcasing both intra-class and inter-class variation of facial expressions that belong to signs. Each subject recorded six videos, each approximately 33 seconds long, with an average of 15 facial expressions per video. The facial expressions are repeated in different orders within the sequence of a video, and the subjects have no restrictions on recording the dataset. Moreover, subjects performed the same sign at different speeds in the videos, resulting in a varying number of frames per facial expression class. This enhances the diversity of our proposed dataset.

Figure 1 shows an RGB video sample of *angry* facial expression, and Figure 2 displays a depth sample of a surprised facial expression, including three views for the appreciation of facial features. Figure 3 shows the eight facial expressions included in the Facial-BSL dataset. The facial expressions share facial muscle movements, i.e. *angry* with *cry*, *laugh* with *agree*, and *surprise* with *yawn*. This indicates that the samples exhibit inter-class and intra-class variations, which present a challenge for the recognition task.

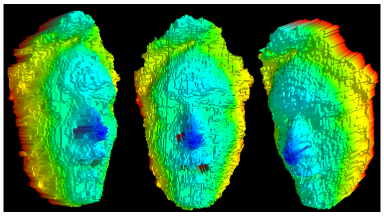

Fig. 2. Depth data for the surprised facial expression from Facial-BSL dataset

Table 2 shows the distribution of video samples across different facial expression classes in the Facial-BSL dataset. The dataset comprises 996 videos that have been manually labeled with the start and end times for each facial expression. The goal of this dataset is to provide both isolated and continuous facial expression videos, offering color and depth information along with precise labeling by an expert in Brazilian Sign Language.

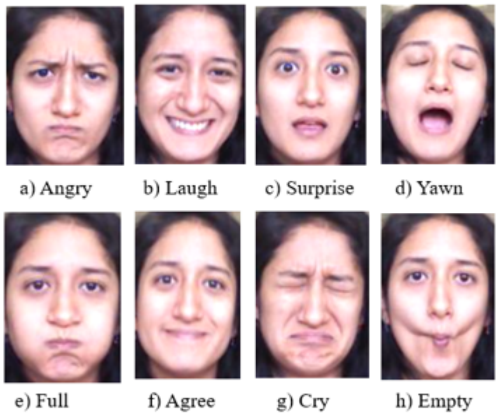

Fig. 3. Eight facial expressions in the Facial-BSL dataset

Table 2. Distribution of the number of samples for each type of facial expression

Type of facial expression	Number of video samples
0). Angry	121
1). Laugh	123
2). Surprise	125
3). Yawn	121
4). Cry	121
5). Full	121
6). Agree	122
7). Empty	142

To increase the number of samples in the dataset, we applied data augmentation techniques, which involved using various filters, rotations, and transformations on the videos. Consequently, the total number of videos increased significantly to $996 + 4979 = 5975$. This augmentation is essential for training deep learning models to ensure the generalization of the recognition task [35].

Table 3 shows the distribution of facial expression samples per class and subject before and after the data augmentation. Each class number is associated with the type of facial expression in Table 2. The dataset is partially imbalanced; however, data augmentation increases the number of samples available for the training stage.

5 Proposed Methodology

The proposed methodology pipeline comprises three stages: data preprocessing, texture map generation, and two-stream architecture design, as illustrated in Figure 4. Each stage will be explained in detail in the following subsections.

5.1 Data Preprocessing

The videos demonstrate the dynamic movement of facial muscles that represent facial expressions. To preprocess the videos, we extracted facial landmarks using the MediaPipe library [40]. The location of landmarks aids in segmenting faces from the background. In order to improve the processing speed, we selected 290 landmarks out of the 468 landmarks available. This resulted in creating separate sub-videos of faces corresponding to each facial expression. The subvideos contained different numbers of frames due to variations in the speed at which the subjects recorded their facial expressions. To ensure consistency, we have used an average of 30 frames per video. If the original subvideo contains a different number of frames, we adjust the frame count accordingly by decreasing or increasing the number of frames.

5.2 Texture Map Image Generation

Considering the findings of [19], we put forth a texture map method to capture facial movements in video. The researchers in [19], encoded the motion of 3D body joint landmarks into RGB texture images. In this study, we propose to use facial landmarks instead of body joints, since the goal is to capture facial motion. In contrast to body joints, landmarks are two-dimensional points extracted from faces in each frame and map subtle changes in features such as the eyes, eyebrows, mouth, among others. Accordingly, the landmarks were classified into three regions for computing texture map images (see Figure 5):

— Region 1 (R1) encompasses the forehead, eyes, and the upper portion of the nose.

Table 3. Distribution of samples per class before and after data augmentation

Subject	Video samples before the data augmentation								Video samples after the data augmentation							
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
p1	11	12	12	12	12	12	12	15	55	60	60	60	60	60	60	75
p2	13	12	13	12	12	12	12	14	65	60	65	60	60	60	60	70
p3	10	10	10	11	11	9	10	10	51	51	52	50	51	47	51	51
p4	12	13	13	12	12	12	12	15	60	65	65	60	60	60	60	75
p5	11	11	12	11	11	13	12	14	55	55	60	55	55	65	60	70
p6	12	12	12	12	12	12	12	13	60	60	60	60	60	60	60	65
p7	11	12	12	11	12	12	12	15	55	60	60	55	60	60	60	75
p8	17	16	16	16	15	15	16	21	85	80	80	80	75	75	80	105
p9	12	13	13	12	12	12	12	12	60	65	65	60	60	60	60	60
p10	12	12	12	12	12	12	12	13	60	60	60	60	60	60	60	65
Total per class	121	123	125	121	121	121	122	142	606	616	627	600	601	607	611	711
Total	996								4979							

— Region 2 (R2) extends from the eyes to the lower part of the nose.

— Region 3 (R3) includes the area from the bottom of the nose to the jaw.

These three types of features are calculated from all combinations of points as follows (Figure 6):

(a) Point-to-Point Distance (PoP) calculates the Euclidean metric between two landmarks. Equation 1 uses p_l and p_k to represent the landmark points, with t representing the frame number:

$$PoP = ||p_l^t - p_k^t||. \tag{1}$$

(b) Point-to-Line Distance (PoL) calculates the Euclidean metric between a landmark and a line formed by two adjacent landmarks. Equation 2 represents the landmark point P_l and t represents the frame number. L_k represents the line formed by two adjacent landmark points:

$$PoL_d = dist(P_l^t - L_k^t). \tag{2}$$

(c) Line-Line Distance (LoL) calculates the angle formed by two lines in a region. Equation 3, L_l and L_k represent two lines, each formed by two adjacent landmarks, and t is the frame number:

$$LoL_d = acosd(L_l^t - L_k^t). \tag{3}$$

The metric used to compute the distances between landmarks points was the Euclidean metric, following the approach of [19] who achieved better results in their proposal. In this study, we considered the possibility of using other well-known metrics in mathematics. However, we chose to use the Manhattan metric and the Euclidean metric for simplicity. After calculating the distance for each region (R1, R2, R3), the feature vectors are encoded into an RGB image. This is because the spatial feature vectors capture temporal information that represents facial motion. The RGB image has columns that represent the feature space at one frame and rows that represent the sequence of a particular feature. Next, the feature vectors are concatenated with PoP in the R channel, PoL in the G channel, and LoL in the B channel. Subsequently, the texture maps from each region (R1, R2, R3) were merged to create a single 224×224 texture map, as illustrated in Figure 5.

5.3 Proposed Two-stream Architecture

This paper presents an architecture for recognizing facial expressions in Brazilian sign language

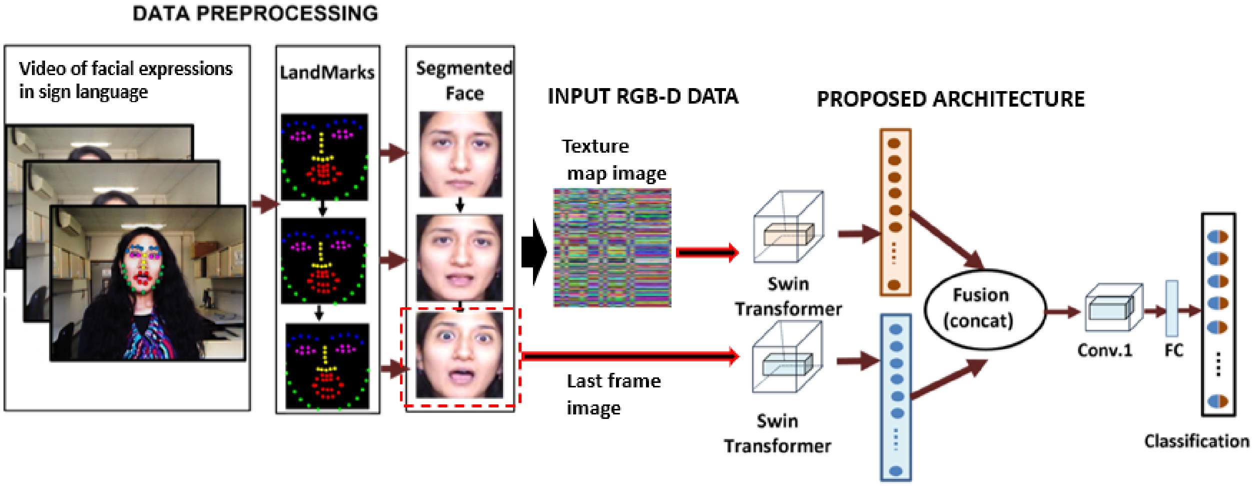

Fig. 4. Pipeline of the proposed methodology

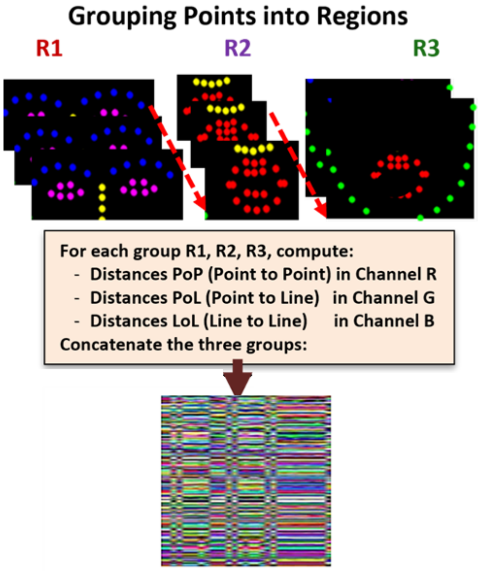
Grouping Points into Regions
R1
R2
R3
For each group R1, R2, R3, compute:
- Distances PoP (Point to Point) in Channel R
- Distances PoL (Point to Line) in Channel G
- Distances LoL (Line to Line) in Channel B
Concatenate the three groups:

Fig. 5. Texture map image generation

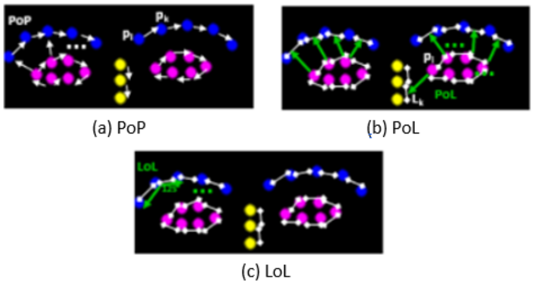

Fig. 6. a) PoP distance, b) PoL distance, and c) LoL distance from region R1 using some landmarks

videos (Figure 4). Unlike previous works, we used two types of images: an RGB image of the last frame of the video ($224 \times 224 \times 3$ pixels) to extract local information about the facial expression, and a texture map image ($224 \times 224 \times 3$ pixels) to extract information about the movement of the facial landmarks. These two images are input to a two-stream architecture. Each stream in the architecture is comprised of a *Swin Transformer* [38], which hierarchically represents non-overlapping small patches of raw pixels from an image. These patches undergo processing through multiple Transformer blocks with modified self-attention calculations, allowing for the handling of spatial and semantic information data. As the network deepens, the patches are merged. Subsequently, *Swin Transformer* blocks are applied to transform features, marking the initial stage of image patch fusion and feature transformation. This process is repeated twice, resulting in stages designated as stage 3 and stage 4. Collectively, these stages yield a hierarchical representation with feature map resolutions equivalent to those of typical convolutional networks. This makes it suitable for image classification, which is relevant to our proposed model. Finally, the outputs of each stream are combined in a fully connected layer to classify the facial expression corresponding to the input images.

6 Experimental Results and Discussion

This section presents the experiments conducted to validate the proposed methodology. The evaluation protocol is defined in section 6.1, and the results of the experiments are presented in section 6.2.

6.1 Evaluation Protocol

We used two state-of-the-art datasets related to facial expressions: CK+ from [39] and KDEF-dyn from [9] that provided sequences of frames related to basic emotions. In addition, we used a Brazilian Sign Language dataset proposed in [49], referred to as LIBRAS in this paper. Our proposed dataset, Facial-BSL, was also included. The remaining datasets in Table 1 were excluded from the experiments because they focused on FACS of grammatical facial expressions or non-manual signs. Additionally, the dataset from [2] was excluded because the partition was on gender.

To generate the texture map images, two distance metrics were employed: the Euclidean and Manhattan. Experiments were conducted for each metric to ascertain which one improved the performance of the proposed architecture. Moreover, five experimental schemes were compared to determine the optimal deep learning architecture:

— M2S_RESNET200 represented a two-stream architecture that took in input an RGB image and a texture map image. It employed the resnet200d.ra2_in1k model proposed by [58], as its backbone.

— M2S_GOOGLE_VIT represented a two-stream architecture that took in input an RGB image and a texture map image. It employed the google/vit-base-patch16-224 model proposed by [60], as its backbone.

— SWIN_BASE_224 represented a two-stream architecture that took in input an RGB image and a texture map image. It employed the swin-base-patch4-window7-224-in22k model proposed by [38], as its backbone.

— TM_GOOGLE_VIT was a single-stream architecture that took a texture map image as input. It used the google/vit-base-patch16-224 model proposed by [60], as its backbone.

— TM_SWIN_BASE_224 was a single-stream architecture that took a texture map image as input. It used the swin-base-patch4-window7-224-in22k model proposed by [38], as its backbone.

— IM_SWIN_BASE_224 was a single-stream architecture that took the last frame image as input. It used the swin-base-patch4-window7-224-in22k model proposed by [38], as its backbone.

All schemes used the same experimental configuration, which consisted of a batch size of 225 with the Adam optimizer and a learning rate of 0.001. The *ReduceLROnPlateau* method was applied if there was no improvement after four epochs, and the learning rate was reduced by a factor of 0.1. *Early stopping* was implemented if there was no improvement after 10 epochs in the training stage. To prevent overfitting in the fully connected layers, a *dropout* ratio of 0.4 was applied. The models were trained for 50 epochs, although this may vary depending on *early stopping*. The layers were partially unfrozen by 20% for fine-tuning. The model performance was evaluated using accuracy, precision, recall, and F1-score metrics. All experiments were carried out on Google Colab, using a GPU Testla V100-SXM2-16GB.

The models were evaluated on datasets with 10-fold person-independence cross-validation experiments (except for the KDEF-dyn dataset, which had 40 folds). This ensured a fair division and prevented subjects from the training set appearing in the test set. The data augmentation technique was used in the training samples.

6.2 Experimental Results

This section presents the results obtained from the proposed methodology. The experimental schemes were used to determine the best deep learning architecture and to evaluate the contribution of each type of information considering the RGB image and the texture map image generated by two different metrics, the Euclidean and Manhattan. Furthermore, experiments were conducted to demonstrate which metric enhanced the performance of the proposed architecture. The following sections will present the results for each dataset, applying the experimental configurations outlined in section 6.1.

6.2.1 Facial-BSL Dataset Experiments

Table 4 presents the results of the five evaluated schemes, including accuracy, precision, recall, and F1-score, for the 10-fold person-independence cross-validation. Additionally, it presents the average and standard deviation of each metric of the Facial-BSL dataset.

The SWIN_BASE_224 scheme performed better than the other schemes with an average precision of 95% and an average accuracy, recall, and F1-score of 94%. The average standard deviation for accuracy, precision, and recall was 0.08, and for F1-score was 0.09, indicating no variability in results for each fold. The proposed architecture took an RGB image and a texture map image (generated by the Euclidean metric) as input. Our analysis leaded us to conclude that the SWIN_BASE_224 scheme was the most effective for this image classification task. This is due to its hierarchical Swin Transformer backbone, which reduced computational complexity.

As previously stated, we conducted the same experiments by changing the type of texture map image generated by the Manhattan metric. Table 5 displays the results of the five schemes, and once again, the SWIN_BASE_224 scheme outperformed the others with 95% accuracy, recall, and F1-score, and 96% precision. The average standard deviation for accuracy, recall, and F1-score was 0.07, and for precision was 0.05. We concluded that the Manhattan metric slightly improved the performance of the scheme by 1%.

The outcomes of both experiments demonstrated that the F1-score metric was well-suited for the analysis of imbalanced datasets, such as Facial-BSL. The F1-score exhibited high values and exhibited slight differences in both experiments, as did the accuracy, precision, and recall. This indicated that the model was efficacious in recognizing facial expressions within their respective classes.

Table 6 presents the confusion matrix of the average results of all folds of the Facial-BSL dataset, calculated using the Manhattan metric. The class ‘cry’ exhibited the highest number of errors, with the model incorrectly identifying it with ‘angry’, ‘yawn’, and ‘agree’, despite the distinctiveness of their facial movements. ‘Agree’ also was misclassified with ‘full’. Furthermore, the labels ‘surprise’ and ‘yawn’ were confused due to their shared mouth opening. Additionally, the labels ‘laugh’ and ‘yawn’ were confused due to their similar mouth openings. Finally, the labels ‘yawn’, ‘full’, and ‘empty’ exhibited a few errors.

We conducted experiments on the other datasets using the SWIN_BASE_224 scheme and its variants (TM_SWIN_BASE_224 and IM_SWIN_BASE_224). Additionally, we evaluated the schemes using two types of texture map images generated by the Euclidean and Manhattan metrics separately. Finally, we compared our results with the state-of-the-art methods.

6.2.2 LIBRAS Dataset Experiments

The dataset proposed in [49] consists of 100 videos of facial expressions extracted from 10 signs, recorded by a single subject. The dataset was balanced, with 100 videos in total. Table 7 shows the results of the dataset using 10-fold cross-validation, employing the Euclidean and Manhattan metric for the generation of texture maps.

Table 4. Results of the 10-fold person-independence cross-validation on the Facial-BSL dataset using the Euclidean metric to generate texture maps

Scheme	Evaluation metric	FACIAL-BSL dataset -Euclidean metric											
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	std
SWIN_BASE_224	Accuracy	1.00	0.92	0.91	0.95	0.98	0.72	0.99	0.95	1.00	0.99	**0.94**	**0.08**
	Precision	1.00	0.93	0.91	0.96	0.98	0.74	0.99	0.95	1.00	0.99	**0.95**	**0.08**
	Recall	1.00	0.92	0.91	0.95	0.98	0.72	0.99	0.95	1.00	0.99	**0.94**	**0.08**
	F1-score	1.00	0.92	0.90	0.95	0.98	0.69	0.99	0.95	1.00	0.99	**0.94**	**0.09**
M2S_GOOGLE_VIT	Accuracy	0.98	0.88	0.99	0.86	0.84	0.99	0.91	0.78	0.79	0.90	0.89	0.08
	Precision	0.99	0.92	0.99	0.79	0.89	0.99	0.92	0.73	0.85	0.94	0.90	0.09
	Recall	0.98	0.87	0.99	0.86	0.85	0.99	0.91	0.77	0.79	0.90	0.89	0.08
	F1-score	0.98	0.87	0.99	0.81	0.84	0.99	0.90	0.73	0.77	0.88	0.88	0.09
M2S_RESNET200	Accuracy	0.85	0.86	0.95	0.85	0.96	0.99	0.95	0.82	0.91	0.94	0.91	0.06
	Precision	0.88	0.90	0.95	0.85	0.96	0.99	0.95	0.84	0.92	0.94	0.92	0.05
	Recall	0.85	0.87	0.95	0.84	0.96	0.99	0.95	0.82	0.91	0.93	0.91	0.06
	F1-score	0.85	0.86	0.95	0.83	0.96	0.99	0.95	0.81	0.91	0.94	0.91	0.06
TM_SWIN_BASE_224	Accuracy	0.71	0.43	0.64	0.71	0.64	0.68	0.69	0.56	0.63	0.65	0.63	0.08
	Precision	0.71	0.43	0.65	0.72	0.65	0.66	0.72	0.57	0.65	0.66	0.64	0.09
	Recall	0.71	0.42	0.63	0.71	0.63	0.69	0.70	0.56	0.63	0.64	0.63	0.09
	F1-score	0.69	0.39	0.63	0.71	0.61	0.64	0.69	0.56	0.63	0.64	0.62	0.09
IM_SWIN_BASE_224	Accuracy	0.98	0.86	0.89	0.99	0.97	0.76	0.98	0.95	0.99	0.98	0.94	0.08
	Precision	0.99	0.90	0.90	0.99	0.98	0.79	0.98	0.95	0.99	0.98	0.95	0.06
	Recall	0.98	0.85	0.89	0.99	0.97	0.76	0.98	0.95	0.99	0.98	0.93	0.08
	F1-score	0.98	0.85	0.89	0.99	0.97	0.75	0.98	0.95	0.99	0.98	0.93	0.08

Table 5. Results of the 10-fold person-independence cross-validation on the Facial-BSL dataset using the Manhattan metric to generate texture maps

Scheme	Evaluation metric	Facial-BSL - Manhattan metric											
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10	Mean	std
SWIN_BASE_224	Accuracy	1.00	0.89	0.93	1.00	0.98	0.78	0.98	0.97	1.00	0.98	**0.95**	**0.07**
	Precision	1.00	0.91	0.94	1.00	0.98	0.85	0.98	0.97	1.00	0.98	**0.96**	**0.05**
	Recall	1.00	0.89	0.93	1.00	0.98	0.78	0.98	0.97	1.00	0.98	**0.95**	**0.07**
	F1-score	1.00	0.89	0.93	1.00	0.98	0.77	0.98	0.97	1.00	0.98	**0.95**	**0.07**
M2S_GOOGLE_VIT	Accuracy	0.98	0.93	0.94	0.86	0.96	0.99	0.88	0.84	0.84	0.97	0.92	0.06
	Precision	0.98	0.93	0.95	0.89	0.96	0.99	0.91	0.75	0.87	0.97	0.92	0.07
	Recall	0.98	0.93	0.93	0.86	0.96	0.99	0.88	0.83	0.83	0.97	0.92	0.06
	F1-score	0.98	0.93	0.94	0.86	0.96	0.99	0.87	0.78	0.82	0.97	0.91	0.07
M2S_RESNET200	Accuracy	0.91	0.85	0.98	0.84	0.98	0.98	0.93	0.80	0.90	0.96	0.91	0.07
	Precision	0.93	0.91	0.98	0.84	0.98	0.99	0.93	0.82	0.91	0.96	0.93	0.06
	Recall	0.90	0.85	0.97	0.83	0.99	0.98	0.92	0.80	0.89	0.96	0.91	0.07
	F1-score	0.90	0.85	0.97	0.82	0.98	0.98	0.92	0.80	0.89	0.96	0.91	0.07
TM_SWIN_BASE_224	Accuracy	0.71	0.39	0.61	0.72	0.69	0.69	0.71	0.54	0.66	0.61	0.63	0.10
	Precision	0.70	0.46	0.64	0.73	0.68	0.66	0.73	0.54	0.69	0.65	0.65	0.09
	Recall	0.71	0.38	0.61	0.72	0.68	0.69	0.71	0.54	0.67	0.61	0.63	0.11
	F1-score	0.69	0.40	0.58	0.72	0.67	0.64	0.71	0.53	0.67	0.61	0.62	0.10
IM_SWIN_BASE_224	Accuracy	1.00	0.87	0.89	0.96	0.99	0.76	0.99	0.93	1.00	0.99	0.94	0.08
	Precision	1.00	0.90	0.89	0.97	0.99	0.84	0.99	0.94	1.00	0.99	0.95	0.06
	Recall	1.00	0.87	0.89	0.96	0.99	0.76	0.99	0.93	1.00	0.99	0.94	0.08
	F1-score	1.00	0.87	0.89	0.96	0.99	0.74	0.99	0.93	1.00	0.99	0.94	0.08

The SWIN_BASE_224 scheme represented a two-stream architecture that took in input an RGB image and a texture map image; this scheme outperformed the other variants in both types of texture map images (Table 9). In the case of the texture map generated by the Euclidean metric, this scheme obtained an average accuracy and recall of 94%, an average precision, and an F1-score of 92%. The average standard deviation was 0.05 for accuracy and precision, 0.07 for F1-score, and 0.08 for precision. Similarly, in the case of the texture map generated by the Manhattan metric, the SWIN_BASE_224 scheme outperformed the other variants. The accuracy and recall were 94%, while the average precision and F1-score were 93%. The average standard deviation was 0.05 for accuracy and recall, 0.06 for F1-score, and 0.07 for precision. Based on

Table 6. Confusion matrix on Facial-BSL dataset

Facial expressions	angry	laugh	surprise	yawn	cry	full	agree	empty
angry	**0.90**	0.00	0.00	0.00	0.07	0.01	0.03	0.00
laugh	0.00	**0.98**	0.00	0.02	0.00	0.00	0.00	0.00
surprise	0.00	0.00	**0.93**	0.07	0.00	0.00	0.00	0.00
yawn	0.00	0.00	0.01	**0.99**	0.00	0.00	0.00	0.00
cry	0.07	0.00	0.00	0.03	**0.87**	0.00	0.02	0.00
full	0.00	0.00	0.00	0.00	0.00	**0.99**	0.00	0.01
agree	0.00	0.00	0.00	0.00	0.00	0.04	**0.95**	0.00
empty	0.00	0.00	0.00	0.00	0.00	0.01	0.00	**0.99**

the results of both experiments, we concluded that the Manhattan metric improved the performance of the SWIN_BASE_224 scheme by 1% in terms of average precision and F1-score.

Table 7. The average results on the LIBRAS dataset of 10-fold cross-validation using the Euclidean and Manhattan metric for generating texture maps

Scheme	Evaluation Metric	Euclidean metric		Manhattan metric	
		Average	std	Average	std
SWIN_BASE_224	Accuracy	**0.94**	**0.05**	**0.94**	**0.05**
	Precision	**0.92**	**0.08**	**0.93**	**0.07**
	Recall	**0.94**	**0.05**	**0.94**	**0.05**
	F1-Score	**0.92**	**0.07**	**0.93**	**0.06**
TM_SWIN_BASE_224	Accuracy	0.79	0.10	0.83	0.09
	Precision	0.80	0.10	0.81	0.12
	Recall	0.79	0.10	0.83	0.09
	F1-Score	0.77	0.11	0.81	0.11
IM_SWIN_BASE_224	Accuracy	0.94	0.05	0.92	0.06
	Precision	0.91	0.07	0.91	0.07
	Recall	0.94	0.04	0.92	0.06
	F1-Score	0.93	0.06	0.91	0.07

Table 8 displays the confusion matrix of the average results of all folds of the LIBRAS dataset using the Manhattan metric. The SWIN_BASE_224 scheme was confused between the facial expressions 'to annihilate', 'to accuse', 'to gain weight', and 'angry', as they shared similar movements of the cheeks, mouth, and furrowed eyebrows. The facial expression 'to annihilate' had the highest number of errors. Also, 'to gain weight' involved inflating the cheeks with mouth movements, and the model misclassified it with 'slim' and 'angry'. The facial expression 'to accuse' was misclassified as 'angry', 'to gain weight', and 'to calm down'. The facial expression 'to calm down' was misclassified as 'surprise' due to the similarity between the two expressions, and as 'slim' despite exhibiting distinct facial movements. 'Surprise' was misclassified as 'happiness' due to the movement of the mouth, smile, and eyebrows. 'Slim' was misclassified as 'to calm down' despite involving different facial movements. 'Lucky' was confused with 'slim' despite the significant differences between the two. 'Love' was only confused with 'happiness' because they both share a smile and mouth movement. Finally, there were no predicted errors in the class 'happiness'.

We compared the proposed scheme with the original paper proposed by [49] and found that our scheme outperformed theirs with an average accuracy of 94%, while the authors obtained an average accuracy of 84%.

6.2.3 CK+ Dataset Experiments

The CK+ dataset proposed by [39] contained 327 sequences of frames from 118 subjects labeled with seven basic expressions (angry, contempt, disgust, fear, happiness, sadness, and surprise). We followed the 10-fold person-independence cross-validation, as explained in section A, according to [42].

In Table 9, the SWIN_BASE_224 scheme obtained an average accuracy of 96%, an average precision, recall, and an F1-score of 94% when employing the Euclidean metric to generate the texture maps. The average standard deviation was 0.04 for accuracy, 0.06 for precision, recall, and F1-score, indicating no variability in the data. Notice that the IM_SWIN_BASE_224 scheme slightly outperformed the SWIN_BASE_224 scheme; the IM_SWIN_BASE_224 scheme took an RGB image as input to the single stream with Swin Transformer as its backbone. This scheme achieved an average accuracy and precision of 96%, an average recall of 95%, and an average F1 score of 94%. We used the F1-score to compare both schemes because the CK+ dataset was unbalanced, so both schemes achieved 94%.

Similarly, in the case of the texture map generated by the Manhattan metric, the SWIN_BASE_224 scheme outperformed the other variants. The average accuracy was 97%,

Table 8. Confusion matrix on the LIBRAS dataset.

Facial expressions	to calm down	to accuse	to calm annihilate	to love	to gain weight	happiness	slim	lucky	surprise	angry
to calm down	**0.91**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.08	0.00
to accuse	0.03	**0.88**	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.02
to annihilate	0.00	0.09	**0.87**	0.00	0.01	0.00	0.00	0.00	0.00	0.03
to love	0.00	0.00	0.00	**0.99**	0.00	0.01	0.00	0.00	0.00	0.00
to gain weight	0.00	0.00	0.00	0.00	**0.97**	0.00	0.02	0.00	0.00	0.01
happiness	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
slim	0.03	0.00	0.00	0.00	0.00	0.00	**0.97**	0.00	0.00	0.00
lucky	0.00	0.00	0.00	0.00	0.00	0.00	0.03	**0.97**	0.00	0.00
surprise	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	**0.96**	0.00
angry	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.88**

the average precision was 96%, and the average recall and F1 score were 95%. The average standard deviation was 0.03 for accuracy and precision, 0.05 for recall and F1-score.

Table 9. The average results on the CK+ dataset of 10-fold cross-validation experiments using the Euclidean and Manhattan metric for generating texture maps

Scheme	Evaluation Metric	Euclidean metric		Manhattan metric	
		Average	std	Average	std
SWIN_BASE_224	Accuracy	**0.96**	**0.04**	**0.97**	**0.03**
	Precision	**0.94**	**0.06**	**0.96**	**0.03**
	Recall	**0.94**	**0.06**	**0.95**	**0.05**
	F1-Score	**0.94**	**0.06**	**0.95**	**0.05**
TM_SWIN_BASE_224	Accuracy	0.77	0.06	0.76	0.05
	Precision	0.70	0.07	0.68	0.10
	Recall	0.68	0.08	0.67	0.07
	F1-Score	0.68	0.07	0.66	0.08
IM_SWIN_BASE_224	Accuracy	0.96	0.04	0.96	0.03
	Precision	0.96	0.04	0.96	0.03
	Recall	0.95	0.05	0.94	0.04
	F1-Score	0.94	0.05	0.94	0.04

The results of both experiments indicated that the Manhattan metric enhanced the performance of the SWIN_BASE_224 scheme by 1% in terms of average accuracy, recall, and F1-score, and by 2% in terms of average precision. Consequently, the SWIN_BASE_224 scheme was employed in conjunction with the Manhattan metric to facilitate a comparative analysis with other methodologies proposed within the existing literature.

Table 10 displays the confusion matrix of the average results of all folds of the CK+ dataset using the Manhattan metric. 'Fear' had the highest number of errors. The model confused it with 'happy' despite the distinct facial movements, and with 'sad' due to shared facial movements. 'Sad' was confused with 'contempt' and 'angry' due to shared facial movements such as the mouth and cheeks. 'Angry' was only confused with 'disgust'. Finally, 'contempt', 'disgust', 'happy', and 'surprise' had no predictive errors.

Table 10. Confusion matrix on the Ck+ dataset

Facial expressions	angry	contempt	disgust	fear	happy	sad	surprise
angry	**0.98**	0.00	0.02	0.00	0.00	0.00	0.00
contempt	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
disgust	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
fear	0.00	0.00	0.00	**0.83**	0.03	0.13	0.00
happy	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
sad	0.02	0.07	0.00	0.00	0.00	**0.92**	0.00
surprise	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**

The CK+ dataset did not include the training/testing splits. Most literature papers divided the data into three types of experiments: train/test comparison [43, 48], 10-fold person-independence cross-validation, k-fold subject-independent cross-validation [43, 56], and 10-fold cross-validation of randomly sampled data [3, 23, 37, 51, 55, 56, 61]. Additionally, some authors chose to use the 6 classes by removing the contempt class [13, 34, 45, 48], while others used 7 classes (standard), and still others used 8 classes by adding neutral facial expressions.

To ensure a fair comparison, the training data samples differed from those in the test data. For instance, each fold was constructed by using 10 subsets of sampling ID in ascending order with a step size of 10, as proposed by [42]. We conducted a comprehensive search of papers published between 2018 and 2024 that utilized the same experimental protocol for recognizing the seven standard facial expressions from the CK+ dataset. Table 11 compares our model with previous works on the extended CK+ dataset. Our proposed scheme ranked among the top methods.

6.2.4 KDEF-dyn Dataset Experiments

The KDEF-dyn dataset, proposed in [9], consisted of videos of six facial expressions: 'laugh', 'sad', 'angry', 'fear', 'disgust', and 'surprise'. The videos display only the face without ears against a black background, with the expression starting from neutral and progressing to the peak expression. This dataset was balanced, with 40 subjects recording one video of each of the six facial expressions, resulting in a total of 240 videos.

We employed 40-fold person-independent cross-validation to ensure a fair comparison with state-of-the-art methods. Two experiments were conducted as in the previous subsections, differentiating by Euclidean and Manhattan metrics for generating texture map images.

Table 12 shows the results of the 40-fold person-independent cross-validation using the Euclidean and Manhattan metrics for generating texture maps. In the case of the Euclidean metric, the SWIN_BASE_224 scheme achieved an average accuracy and recall of 95%, an average precision of 93%, and an average F1-score of 94%. The average standard deviation was 0.07 for accuracy and recall, 0.11 for precision, and 0.09 for F1-score, indicating no variability in the data. Note that the IM_SWIN_BASE_224 scheme slightly outperformed the SWIN_BASE_224 scheme. The IM_SWIN_BASE_224 scheme used the RGB image as input to the single stream with the Swin Transformer as its backbone. The scheme achieved an average accuracy, precision, and recall of 95% and an average F1 score of 94%. The only difference was in the average precision, which was 2% lower in the IM_SWIN_BASE_224. This difference can be attributed to the dataset containing only facial expressions without any head movement.

Whereas in the Manhattan metric for generating texture maps, the SWIN_BASE_224 scheme achieved an average accuracy, precision, and recall of 94% and an average F1-score of 93% as shown in Table 12. The average standard deviation was 0.07 for accuracy and recall, 0.09 for precision and F1-score, indicating no variability in the data. It is worth noting that the IM_SWIN_BASE_224 scheme slightly outperformed the SWIN_BASE_224 scheme. The IM_SWIN_BASE_224 scheme took as input an RGB image to the single stream with Swin Transformer as its backbone. The scheme achieved an average accuracy, precision, and recall of 95% and an average F1-score of 94%. Based on the experiments, the Euclidean metric slightly outperformed the Manhattan metric in terms of accuracy, recall, and F1-score by 1% in this dataset. However, the precision of the Manhattan metric is 1% higher than that of the Euclidean metric.

Table 13 displays the confusion matrix of the average results of all folds of the KDEF-dyn dataset using the Manhattan metric. Facial expressions were analyzed based solely on facial motion, excluding head movements. The facial expression 'fear' had the highest number of errors, as the model often confused it with 'surprise' and 'sad' due to shared facial movements. Similarly, 'angry' was often confused with 'sad' and 'disgust' due to shared movements of the mouth, eyebrows, and cheeks. 'Disgust' was confused with several other facial expressions, including 'angry', 'surprise', 'sad', and 'fear'. 'Sad' was confused with 'angry' and 'disgust'. Finally, 'surprise' was only confused with 'fear', while 'laugh' had no errors.

We compared our proposed scheme with the work proposed by [46]. We followed the same experimental protocol as the authors to make a fair comparison. Table 14 shows the obtained results, and both the Euclidean and Manhattan metrics

Table 11. Comparison with the state-of-the-art methods on the CK+ dataset

Proposed Method	Features	Accuracy (%)
Spatial-temporal RNN [62]	Multi-channel EEG signals	95.40
Island loss CNN [8]	Facial images	94.35
Score fusion (base line) [42]	Fixed dimension	94.80
Frame Attention networks [42]	Feature representation	99.69
The algorithm combines gentle boost decision trees and neural networks [24]	Local binary features	96.48
Hybrid 3D-CNN and ConvLSTM [53]	Aligned facial images	95.10
SWIN_BASE_224 two-stream model (our)	Texture map images generated by Manhattan metric and RGB images	**97.00**

Table 12. The average and std results on the KDEF-dyn dataset of 40-fold person-independence cross-validation experiments using the Euclidean and Manhattan metric for generating texture maps

Scheme	Evaluation Metric	Euclidean metric		Manhattan metric	
		Average	std	Average	std
SWIN_BASE_224	Accuracy	**0.95**	**0.07**	**0.94**	**0.07**
	Precision	**0.93**	**0.11**	**0.94**	**0.09**
	Recall	**0.95**	**0.07**	**0.94**	**0.07**
	F1-Score	**0.94**	**0.09**	**0.93**	**0.09**
TM_SWIN_BASE_224	Accuracy	0.49	0.14	0.55	0.13
	Precision	0.46	0.18	0.56	0.18
	Recall	0.49	0.14	0.55	0.13
	F1-Score	0.44	0.16	0.52	0.15
IM_SWIN_BASE_224	Accuracy	0.95	0.07	0.95	0.07
	Precision	0.95	0.09	0.95	0.09
	Recall	0.95	0.07	0.95	0.07
	F1-Score	0.94	0.09	0.94	0.09

Table 13. Confusion matrix on the KDEF-dyn dataset

Facial expressions	angry	happy	surprise	sad	fear	disgust
angry	**0.93**	0.00	0.00	0.03	0.00	0.05
happy	0.00	**1.00**	0.00	0.00	0.00	0.00
surprise	0.00	0.00	**0.99**	0.00	0.01	0.00
sad	0.01	0.00	0.00	**0.96**	0.00	0.03
fear	0.00	0.00	0.08	0.01	**0.91**	0.00
disgust	0.03	0.00	0.02	0.01	0.01	**0.94**

of the SWIN_BASE_224 scheme outperformed the results.

Table 14. Comparison of the proposed method with others on the KDEF-dyn dataset

Proposed Method	Features	Accuracy (%)
Deep stacked autoencoder (DSA) [46]	Fusion of local neighborhood patterns and LBP.	88.50
SWIN_BASE_224 Two-stream model (our)	Texture maps generated by Manhattan metric.	94.00
	Texture maps generated by Euclidean metric.	95.00

7 Strengths, Limitations and Future Directions

Our research has strengths like the proposed Facial-BSL dataset validated by an expert. The videos depicted various facial expressions executed continuously, contributing to the state of the art. We also propose a new method called texture map image, which uses the Pop, PoL, and LoL distances to detect variations between adjacent landmarks over time. The study tested the Euclidean and Manhattan metrics to calculate the distance between landmarks. The results indicated that the Manhattan metric improved the quality of the results.

Finally, the experiments conducted with different schemes demonstrated that the two-stream approach based on the Swin Transformer model achieved better results. This approach was excellent for recognizing facial expressions involving facial movement.

The work is limited to using literature recommendations for transfer learning and data augmentation in our pipeline due to the limited number of samples in the datasets. Additionally, we only compared our approach with different versions of the schemes using the Facial-BSL dataset. Future developments in this work will explore the depth data of facial movement in the Facial-BSL dataset and test other fusion methods. Moreover, other attention models should be modified and tested to improve the recognition of facial expressions that share similar movements, such as anger, sadness, and crying. Finally, the dataset should increase the number of subjects, samples, and various categories of facial expressions.

8 Conclusion

The research proposed a two-stream architecture based on Swin Transformer model for the recognition of facial expressions in sign language. The proposed architecture takes as input an RGB image and a texture map of facial expressions in sign language. Moreover, the Facial-BSL dataset was introduced for the purpose of recognizing facial expressions in Brazilian Sign Language. It should be noted that the videos in the Facial-BSL dataset exhibited significant movements of the head, jaw, and mouth, while the images in the CK+ and KDEF-dyn dataset showed minimal movements of the mouth and jaw, but not of the head.

Several experimental schemes were tested on the Facial-BSL dataset, including the use of both Euclidean and Manhattan metrics for generating texture maps. The SWIN_BASE_224 scheme exhibited superior performance relative to the other schemes employing the Manhattan metric. Moreover, it demonstrated superior performance to M2S_GOOGLE_VIT and M2S_RESNET200. The SWIN_BASE_224, which employs the Manhattan metric, demonstrated superior performance relative to the other schemes and the original paper on the LIBRAS dataset. Furthermore, it outperformed the other schemes on the CK+ dataset, although it was ranked second in terms of state-of-the-art results. Nevertheless, the IM_SWIN_BASE_224 demonstrated a 1% improvement in performance compared to the SWIN_BASE_224 in the KDEF-dyn dataset due to the absence of head movements and outperformed the original paper.

These findings indicated that the SWIN_BASE_224 with the Manhattan metric exhibited enhanced recognition performance and outperformed CNN models.

Acknowledgments

We appreciate the volunteer's willingness to collaborate in constructing the dataset.

References

1. **Acenjo, B. X. T., Pariona, M. A. T., Cárdenas, E. J. E. (2023).** Comparativa entre RESNET-50, VGG-16, Vision Transformer y Swin Transformer para el reconocimiento facial con oclusión de una mascarilla. Interfases, , No. 017, pp. 56–78. DOI: 10.26439/interfases2023.n017.6361.
2. **Alaghband, M., Yousefi, N., Garibay, I. (2020).** Facial Expression Phoenix (FePh): An Annotated Sequenced Dataset for Facial and Emotion-Specified Expressions in Sign Language. arXiv, pp. 1–9. DOI: 10.48550/arXiv.2003.08759.
3. **Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K., Seguier, R. (2021).** Learning Vision Transformer With Squeeze and Excitation for Facial Expression Recognition. arXiv, pp. 1–13. DOI: 10.48550/arXiv.2107.03107.
4. **Aran, O., Ari, I., Guvensan, A., Haberdar, H., Kurt, Z., Turkmen, I., Uyar, A., Akarun, L. (2007).** A Database of Non-Manual Signs in Turkish Sign Language. IEEE 15th Signal Processing and Communications Applications, pp. 1–4. DOI: 10.1109/SIU.2007.4298708.

5. **Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I. R., Movellan, J. R. (2006).** Automatic Recognition of Facial Actions in Spontaneous Expressions. Journal of Multimedia, Vol. 1, No. 6, pp. 22–35. DOI: 10.4304/jmm.1.6.22-35.

6. **Bejarano, G., Huamani-Malca, J., Cerna-Herrera, F., Alva-Manchego, F., Rivas, P. (2022).** PeruSIL: A Framework to Build a Continuous Peruvian Sign Language Interpretation Dataset. Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, pp. 1–8.

7. **Bobojanov, S., Kim, B. M., Arabboev, M., Begmatov, S. (2023).** Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. Applied Sciences, Vol. 13, No. 22, pp. 1–14. DOI: 10.3390/app132212271.

8. **Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., Tong, Y. (2018).** Island Loss for Learning Discriminative Features in Facial Expression Recognition. 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 302–309. DOI: 10.48550/arXiv.1710.03144.

9. **Calvo, M. G., Fernández-Martín, A., Recio, G., Lundqvist, D. (2018).** Human Observers and Automated Assessment of Dynamic Emotional Facial Expressions: KDEF-dyn Database Validation. Frontiers in Psychology, Vol. 9, pp. 1–12. DOI: 10.3389/fpsyg.2018.02052.

10. **Cardoso, E. C., et al. (2018).** A Interface Prosódia/Pragmática Nas Expressões Faciais Das Emoções Dos Surdos. , pp. 1–58.

11. **Cardoso, M., Freitas, F., Barbosa, F. V., Lima, C., Peres, S. M., Hung, P. (2020).** Automatic Segmentation of Grammatical Facial Expressions in Sign Language: Towards an Inclusive Communication Experience. Proceedings of the 53rd Hawaii International Conference on System Science, pp. 1499–1508. DOI: 10.24251/HICSS.2020.184.

12. **Chaudhari, A., Bhatt, C., Krishna, A., Mazzeo, P. L. (2022).** ViTFER: Facial Emotion Recognition With Vision Transformers. Applied System Innovation, Vol. 5, No. 4, pp. 1–16. DOI: 10.3390/asi5040080.

13. **Connie, T., Al-Shabi, M., Cheah, W. P., Goh, M. (2017).** Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator. International Workshop on Multi-Disciplinary Trends in Artificial Intelligence, Springer, pp. 139–149. DOI: 10.1007/978-3-319-69456-6_12.

14. **da Silva, E. P., Costa, P. D. P., Kumada, K. M. O., De Martino, J. M. (2020).** Silfa: Sign Language Facial Action Database for the Development of Assistive Technologies for the Deaf. 15th IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, pp. 688–692. DOI: 10.1109/FG47880.2020.00059.

15. **da Silva, E. P., Costa, P. D. P., Kumada, K. M. O., De Martino, J. M. (2022).** Facial Action Unit Detection Methodology with Application in Brazilian Sign Language Recognition. Pattern Analysis and Applications, pp. 1–17. DOI: 10.1007/s10044-021-01024-5.

16. **da Silva, E. P., Costa, P. D. P., Kumada, K. M. O., De Martino, J. M., Florentino, G. A. (2020).** Recognition of Affective and Grammatical Facial Expressions: A Study for Brazilian Sign Language. Computer Vision – ECCV Workshops, pp. 218–236. DOI: 10.1007/978-3-030-66096-3_16.

17. **Deng, D., Chen, Z., Shi, B. E. (2020).** Multitask Emotion Recognition with Incomplete Labels. 15th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 592–599. DOI: 10.48550/arXiv.2002.03557.

18. **Deshpande, N., Nunnari, F., Avramidis, E. (2022).** Fine-Tuning of Convolutional Neural Networks for the Recognition of Facial Expressions in Sign Language Video Samples. Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives, pp. 29–38.

19. **Ding, Z., Wang, P., Ogunbona, P. O., Li, W. (2017).** Investigation of Different Skeleton Fea-

tures for CNN-based 3D Action Recognition. IEEE International Conference on Multimedia & Expo Workshops, pp. 617–622. DOI: 10.1109/ICMEW.2017.8026286.

20. **Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020).** An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, pp. 1–22. DOI: 10.48550/arXiv.2010.11929.

21. **Escobedo, E., Ramirez, L., Camara, G. (2019).** Dynamic Sign Language Recognition Based on Convolutional Neural Networks and Texture Maps. 32nd SIBGRAPI Conference on Graphics, Patterns and Images, pp. 265–272. DOI: 10.1109/SIBGRAPI.2019.00043.

22. **Friesen, E., Ekman, P. (1978).** Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Vol. 3, No. 2, pp. 5. DOI: 10.1037/t27734-000.

23. **Ghazouani, H. (2021).** A Genetic Programming-Based Feature Selection and Fusion for Facial Expression Recognition. Applied Soft Computing, Vol. 103, pp. 1–14. DOI: 10.1016/j.asoc.2021.107173.

24. **Gogić, I., Manhart, M., Pandžić, I. S., Ahlberg, J. (2020).** Fast Facial Expression Recognition Using Local Binary Features and Shallow Neural Networks. The Visual Computer, Vol. 36, No. 1, pp. 97–112. DOI: 10.1007/s00371-018-1585-8.

25. **Guerra, R. R., Rezende, T. M., Guimaraes, F. G., Almeida, S. G. M. (2018).** Facial Expression Analysis in Brazilian Sign Language for Sign Recognition. Anais do XV Encontro Nacional de Inteligência Artificial e Computacional, pp. 1–18. DOI: 10.17648/educare.v13i28.18372.

26. **Guimarães, C., Pereira, R. C., Labes, M. G., Fernandes, S. F. (2018).** A Expressao Facial É Parte Integrante da Língua de Sinais - Libras como L2. Educere et Educare, Vol. 13, No. 28, pp. 1–18. DOI: 10.17648/educare.v13i28.18372.

27. **Irasiak, A., Kozak, J., Piasecki, A., Steclik, T. (2023).** Processing Real-Life Recordings of Facial Expressions of Polish Sign Language Using Action Units. Entropy, Vol. 25, No. 1, pp. 1–18. DOI: 10.3390/e25010120.

28. **Javaid, S., Rizvi, S. (2023).** Manual and Non-Manual Sign Language Recognition Framework Using Hybrid Deep Learning Techniques. Journal of Intelligent & Fuzzy Systems, Vol. 45, No. 3, pp. 3823–3833. DOI: 10.3233/JIFS-230560.

29. **Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y. (2021).** Skeleton Aware Multi-Modal Sign Language Recognition. CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 3408–3418. DOI: 10.48550/arXiv.2103.08833.

30. **Ko, B. C. (2018).** A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors, Vol. 18, No. 2, pp. 1–20. DOI: 10.3390/s18020401.

31. **Kumar, P., Roy, P. P., Dogra, D. P. (2018).** Independent Bayesian Classifier Combination Based Sign Language Recognition Using Facial Expression. Information Sciences, Vol. 428, pp. 30–48. DOI: 10.1016/j.ins.2017.10.046.

32. **Laines, D., Gonzalez-Mendoza, M., Ochoa-Ruiz, G., Bejarano, G. (2023).** Isolated Sign Language Recognition Based on Tree Structure Skeleton Images. CVF Conference on Computer Vision and Pattern Recognition, pp. 276–284. DOI: 10.48550/arXiv.2304.05403.

33. **Leong, S.-M., Phan, R. C.-W., Baskaran, V. M. (2023).** Emotion-Specific AUs for Micro-Expression Recognition. Multimedia Tools and Applications, Vol. 83, No. 1, pp. 1–38. DOI: 10.1007/s11042-023-16326-5.

34. **Li, M., Li, X., Sun, W., Wang, X., Wang, S. (2021).** Efficient Convolutional Neural Network with Multi-Kernel Enhancement Features for Real-Time Facial Expression Recognition. Journal of Real-Time Image Processing, Vol. 18, No. 2, pp. 1–12. DOI: 10.1007/s11554-021-01088-w.

35. **Li, S., Deng, W. (2020).** Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, Vol. 13, No. 3, pp. 1195–1215. DOI: 10.48550/arXiv.1804.08348.

36. **Liu, C., Hirota, K., Dai, Y. (2023).** Patch Attention Convolutional Vision Transformer for Facial Expression Recognition with Occlusion. Information Sciences, Vol. 619, pp. 781–794. DOI: 10.1016/j.ins.2022.11.068.

37. **Liu, D., Zhang, H., Zhou, P. (2021).** Video-Based Facial Expression Recognition Using Graph Convolutional Networks. 25th International Conference on Pattern Recognition, IEEE, pp. 607–614. DOI: 10.1109/ICPR48806.2021.9413094.

38. **Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021).** Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. CVF International Conference on Computer Vision, pp. 9992–10002. DOI: 10.1109/ICCV48922.2021.00986.

39. **Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010).** The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

40. **Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019).** Mediapipe: A Framework for Building Perception Pipelines. arXiv, pp. 1–9. DOI: 10.48550/arXiv.1906.08172.

41. **Ma, F., Sun, B., Li, S. (2021).** Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion. IEEE Transactions on Affective Computing, Vol. 14, No. 2, pp. 1236–1248. DOI: 10.1109/TAFFC.2021.3122146.

42. **Meng, D., Peng, X., Wang, K., Qiao, Y. (2019).** Frame Attention Networks for Facial Expression Recognition in Videos. IEEE International Conference on Image Processing, pp. 3866–3870. DOI: 10.1109/ICIP.2019.8803603.

43. **Minaee, S., Minaei, M., Abdolrashidi, A. (2021).** Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. Sensors, Vol. 21, No. 9, pp. 1–16. DOI: 10.3390/s21093046.

44. **Mukushev, M., Sabyrov, A., Imashev, A., Koishibay, K., Kimmelman, V., Sandygulova, A. (2020).** Evaluation of Manual and Non-manual Components for Sign Language Recognition. Proceedings of The 12th Language Resources and Evaluation Conference, pp. 6073—6078.

45. **Niu, B., Gao, Z., Guo, B. (2021).** Facial Expression Recognition with LBP and ORB Features. Computational Intelligence and Neuroscience, Vol. 2021, No. 7, pp. 1–10. DOI: 10.1155/2021/8828245.

46. **Pitchaiyan, S., Savarimuthu, N. (2022).** Deep Stacked Autoencoder-Based Automatic Emotion Recognition Using an Efficient Hybrid Local Texture Descriptor. Journal of Information Technology Research, Vol. 15, No. 1, pp. 1–26. DOI: 10.4018/JITR.2022010103.

47. **Porta-Lorenzo, M., Vázquez-Enríquez, M., Pérez-Pérez, A., Alba-Castro, J. L., Docío-Fernández, L. (2022).** Facial Motion Analysis Beyond Emotional Expressions. Sensors, Vol. 22, No. 10, pp. 1–21. DOI: 10.3390/s22103839.

48. **Punuri, S. B., Kuanar, S. K., Kolhar, M., Mishra, T. K., Alameen, A., Mohapatra, H., Mishra, S. R. (2023).** Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning. Mathematics, Vol. 11, No. 3, pp. 1–24. DOI: 10.3390/math11030776.

49. **Rezende, T. M., Castro, C., Almeida, S. (2016).** An Approach for Brazilian Sign Language (BSL) Recognition based on Facial Expression and k-NN Classifier. 29th SIBGRAPI, pp. 1–4.

50. **Saito, J., Kawamura, R., Uchida, A., Youoku, S., Toyoda, Y., Yamamoto, T., Mi, X., Murase, K. (2020).** Action Units Recog-

nition by Pairwise Deep Architecture. arXiv, pp. 1–4. DOI: 10.48550/arXiv.2010.00288.

51. **Shi, C., Tan, C., Wang, L. (2021).** A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network. IEEE Access, Vol. 9, pp. 39255–39274. DOI: 10.1109/ACCESS.2021.3063493.

52. **Silva, K. A. d., Severo, J. (2014).** Que língua é essa?: Crenças e preconceitos em torno da língua de sinais e da realidade surda. SciELO Brasil, Vol. 14, No. 4. DOI: 10.1590/1984-639820145507.

53. **Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., Singh, S. (2023).** Facial Expression Recognition in Videos Using Hybrid CNN & ConvLSTM. International Journal of Information Technology, Vol. 15, No. 4, pp. 1819–1830. DOI: 10.1007/s41870-023-01183-0.

54. **Tian, Y.-l., Kanade, T., Cohn, J. F. (2001).** Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 2, pp. 97–115. DOI: 10.1109/34.908962.

55. **Tuncer, T., Dogan, S., Subasi, A. (2023).** Automated Facial Expression Recognition Using Novel Textural Transformation. Journal of Ambient Intelligence and Humanized Computing, Vol. 14, No. 7, pp. 9435–9449. DOI: 10.1007/s12652-023-04612-x.

56. **Umer, S., Rout, R. K., Pero, C., Nappi, M. (2022).** Facial Expression Recognition with Trade-Offs Between Data Augmentation and Deep Learning Features. Journal of Ambient Intelligence and Humanized Computing, pp. 1–15. DOI: 10.1007/s12652-020-02845-8.

57. **Wadhawan, A., Kumar, P. (2020).** Deep Learning-Based Sign Language Recognition System for Static Signs. Neural Computing and Applications, Vol. 32, No. 12, pp. 7957–7968. DOI: 10.1007/s00521-019-04691-y.

58. **Wightman, R., Touvron, H., Jegou, H. (2021).** Resnet strikes back: An improved training procedure in timm. NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future.

59. **World Health Organization (2023).** Deafness and Hearing Loss. URL: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

60. **Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P. (2020).** Visual Transformers: Token-based Image Representation and Processing for Computer Vision. arXiv, pp. 1–12. DOI: 10.48550/arXiv.2006.03677.

61. **Zhang, H., Huang, B., Tian, G. (2020).** Facial Expression Recognition Based on Deep Convolution Long Short-Term Memory Networks of Double-Channel Weighted Mixture. Pattern Recognition Letters, Vol. 131, pp. 128–134. DOI: 10.1016/j.patrec.2019.12.013.

62. **Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y. (2018).** Spatial–Temporal Recurrent Neural Network for Emotion Recognition. IEEE Transactions on Cybernetics, Vol. 49, No. 3, pp. 839–847. DOI: 10.1109/TCYB.2017.2788081.

63. **Zhao, S., Liu, C., Liu, G. (2022).** Facial Expression Recognition Based on Visual Transformers and Local Attention Features Network. 7th International Conference on Computer and Communication Systems, pp. 228–231. DOI: 10.1109/ICCCS55155.2022.9846106.

Article received on 14/08/2024; accepted on 15/02/2025.
**Corresponding author is Lourdes Ramirez-Cerna.*