Photogram Classification-Based Emotion Recognition

JUAN MIGUEL LÓPEZ-GIL AND NESTOR GARAY-VITORIA
Faculty of Informatics, University of the Basque Country UPV/EHU, 20018 Donostia-San Sebastián, Gipuzkoa province, Spain
Corresponding author: Nestor Garay-Vitoria (nestor.garay@ehu.eus)

This work was supported in part by the Department of Education, Universities, and Research of Basque Government, under Grant IT980-16.

ABSTRACT This paper presents a method for facial emotion recognition based on parameterized photograms and machine learning techniques. Videos of people displaying emotions are parameterized by a facial feature-based emotional category association process to determine whether a given photogram expresses emotions by comparing the facial action units displayed with findings in the literature about facial emotion. To test the proposed approach, two strategies are adopted. First, photograms displaying emotions are gathered, and then different machine learning classifiers are applied to check the goodness of the obtained set of categorized emotional photograms. Second, classifiers trained on the sets of emotional photograms were then used to emotionally classify all the videos in each database, using all the photograms with no preprocessing or photogram selection. The presented method was tested using the OpenFace parameterizer with emotional videos gathered from Multimedia Understanding Facial Expression (MUG) and Cohn-Kanade (CK+) databases. The outcomes achieved for emotional photogram classification on the sets of emotional photograms reached maximums of 99.80% and 99.63% in the MUG and CK+ databases, respectively. The videos were classified using different voting strategies regarding the outcome of each photogram in the video with all the photogram emotion recognition classifiers obtained results reflecting recognition rates of 70.71% and 66.36% for the videos in MUG and CK+ databases, and reached up to 72.55% and 88.37% when classifier combination strategies were used. The work carried out opens the door to follow-up work concerning data preprocessing and the use of different classifier combination methods in facial emotion recognition.

INDEX TERMS Affective computing, emotion recognition, classification algorithms, image classification, image sequence analysis, machine learning algorithms.

I. INTRODUCTION
Human beings are mostly emotional, and our social interaction is measured by taking into consideration our ability to communicate emotions and to perceive the emotional states of others. Affective computing provides computing systems with mechanisms that emulate and/or interpret human emotions [1]. Its main objective is to make communication with computing systems easier and more natural.

Emotions are multimodal [2], and currently, they are recognized within the human-computer interaction (HCI) area using the following factors, both separately and in combination, starting from facial expressions [3], oral intonation [4], psycho-physiological information [5], or even the texts used [6]–[8]. To make them known to users, it is usual to employ avatars [9] and speech synthesis [10], frequently combining the two. Although, in general, people are experts in recognizing and expressing emotions, sometimes there is misunderstanding when transmitting them. This may be caused by ambient issues (noise, lighting, or distance between interlocutors), or even personal issues (concentration or the behavior or confidence with the interlocutor). This is why emotional resources are frequently validated by people, in order to ascertain whether they really express the correct emotion or if the interlocutors are able to perceive them adequately [11]. Many times, resources are not very expressive or not correctly understood by humans; therefore, computing
systems still fall far short of identifying emotions with a 100% recognition rate.

Facial expression is one of the most natural and immediate means for human beings to communicate their emotions, as the human face can express emotions sooner than people verbalize or even realize their feelings. Automatic facial expression recognition (FER) has become an increasingly important research area that involves computer vision, machine learning, and behavioral sciences. Much progress has been made in building computer systems to understand and use this natural form of human communication, although most of these systems attempt to recognize only a small set of prototypical emotional expressions [12]. FER can be used for many applications, such as security [13], human-computer interaction [14], driver safety [15], and health care [16].

We propose a method for recognizing emotions from facial expressions using different criteria found in the literature. Using facial features extracted with publicly available facial landmarks, action unit detection tools, and emotional video databases, we show that the proposed method for categorizing emotional photograms allows a valid set of emotionally labeled photograms that can then be used for emotion recognition. We show that different basic classifiers can be used to adequately classify the emotional photograms obtained from the photogram set. Finally, we show that, without any kind of preprocessing of specific photogram selection, using photogram emotional classification classifiers allows video sequences to be emotionally classified with majority voting strategies, and that outcomes improve when combining different classifiers.

The remainder of this paper is organized as follows. The next section presents studies found in the literature on affective computing and how to recognize emotions using artificial intelligence techniques. The method used to automatically extract and categorize relevant emotional photograms from video recordings is presented next, together with the procedures followed to train photogram-based emotion classifiers, apply them to emotionally classify videos, and apply majority voting strategies to classify those videos. Subsequently, a study was conducted to validate the proposed approach, and the results are presented in detail. After a discussion of the results obtained, several conclusions are drawn, and future work is proposed.

II. RELATED WORK

There are several theories about emotions that come from psychology that may be employed in affective computing. One of the most commonly used classifications is categorical [17]. According to this theory, there is a discrete number of emotions that depend on the reference being considered, and the objective is to identify the most appropriate category during the interaction. Without considering that the number of discrete values is not unanimous, some authors have studied emotions that could be applied regardless of culture [3]. According to their research, six emotions may be considered universal (these are known as the Big Six emotions): Sadness, fear, joy, anger, surprise, and disgust.

Furthermore, the number of emotional resource datasets or databases open for research purposes is increasing. A thorough review can be found in [18]. The categorical emotional databases found in the literature mainly use the Big Six emotions as emotional categories. [19] reviewed databases with emotional facial expressions, as shown in Table S1. In their work, aimed at researchers from different research fields (perceptual and cognitive sciences, affective computing, and computer vision), Kaulard et al. found that facial expressions are an important channel to show both emotional and conversational expressions [19]. In the literature, several databases employ emotional categories [20]–[22]. There are databases that store emotional audio and videos [22]–[28]. Popular publicly available emotional facial expression databases, including those with pictures alone or video clips, include the Cohn-Kanade database (CK+) [20], the Japanese Female Facial Expression (JAFFE) [29], the multimedia understanding facial expression database (MUG) [30], [31], the Indian Spontaneous Expression Database (ISED) [32], Radboud Faces Database (RaFD) [33], Oulu-CASIA facial expression database [34], AffectNet [35], and the CMU multi-pose, illumination, and expression face database (Multi-PIE) [36], MMI [37], and AFEW [38].

Most datasets were recorded in laboratory settings. Therefore, there are studies that reveal challenges in unconstrained real-world environments, such as lighting variation, head pose, and subject-dependence, which may not be resolved by only analyzing images/videos in the FER system [39].

Regarding emotion recognition based on facial expressions, Ekman and Friesen developed a facial action coding system (FACS) [2] to describe facial expressions using action units (AUs). Of the 44 FACS AUs that they defined, 30 AUs were anatomically related to the contractions of specific facial muscles, 12 of which corresponded to the upper face, and 18 to the lower face. AU scans occur either singly or in combination [12]. Using FACS, human encoders may manually encode all facial expressions using these 30 AUs. The AU combinations defined in the FACS may describe the emotional labels. Due to its descriptive ability, FACS has emerged as a facial behavior measurement criterion in several fields, including computer-based vision [20]. In general, it has a reliability between good and excellent, owing to the occurrence, intensity, and timing of the individual AUs and more global measurements corresponding to particular emotion combinations [40]. FACS has been validated in several studies, and its utility has been demonstrated in a wide range of studies with infants and adults in North America, Europe, and Asia. A good introduction to FACS-related literature can be found in [41]. The Emotion Facial Action Coding System (EMFACS-7) was later proposed based on FACS to detect which basic emotions have corresponding prototypical facial expressions [42]. In everyday life, however, such prototypical expressions occur relatively infrequently. Instead, emotion is more often communicated by subtle changes in one or
a limited number of discrete facial features, such as tightening of the lips in anger or obliquely lowering the lip corners in sadness [43]. In [44], the psychometric evaluation of FACS was summarized, explaining validity, stability, utility, and reliability (interobserver agreement), which includes occurrence and temporal precision, intensity, and aggregates.

Studies have also been conducted to empirically investigate the facial action unit configurations used by actors to convey specific emotions in short-affect bursts and to examine the extent to which observers can infer a person’s emotions from facial expression configurations. The 13 emotions selected to be enacted (Surprise, Fear, Anger, Disgust, Contempt, Sadness, Boredom, Relief, Interest, Enjoyment, Happiness, Pride, and Amusement), and recognition of actors’ intentions by human judges ranged from 35.0% for surprise to 62.1% for enjoyment [45].

Other studies have presented automatic FER based on FACS, such as Bartlett et al. [46] and Baltrusaitis et al. [47]. Automatic recognition is performed in several steps, starting from face and eye detection, including facial landmarks, head pose, and eye gaze. Next, AU estimation is performed using classifiers such as support vector machines (SVMs) [48].

A review of FER based on visual information can be found in [18]. The author presents, on the one hand conventional FER approaches consisting of three steps, namely, face and facial component detection, feature extraction, and expression classification. On the other hand, he presented deep-learning-based FER approaches using deep networks, and even a hybrid deep-learning approach combining a convolutional neural network for the spatial features of individual frames and long short-term memory for temporal features of consecutive frames. In his paper, he also presents a brief review of publicly available evaluation metrics and a comparison with benchmark results for a quantitative comparison of FER research.

In general, there are several ways to determine the category or concrete value of emotions in order to recognize them, as reviewed in [18] and [49]. Usually, after preprocessing where faces are detected, geometric transformations are made, images are processed, features are extracted, including action units, and classifiers are used to label input figures or features. The methods to draw up the classifications include support vector machines [48] and k-nearest neighbors [50]. However, there are other deep learning methods, such as convolutional neural networks [51] and recurrent neural networks [52], which reduce the need for preprocessing [53].

In their work, Kahou et al. employed recurrent neural networks in video clips included in the Acted Facial Expressions in the Wild (AFEW) 5.0 dataset [54]. Lakshmi and Palanivel applied a Viola-Jones algorithm to each frame to detect the face and mouth regions. Next, they extracted features, trained, and classified emotions using support vector machines in the Indian spontaneous expression database (ISED) [55]. Abdul-salam et al. studied the utilization of a deep convolutional neural network to ten emotions from the Amsterdam dynamic facial expression set-bath intensity variations (ADFES-BIV) dataset [56]. Hu et al. employed two networks (local and global) and a fusion to determine their performance with AFEW, CK+, and MMI datasets and concluded that integrating both networks achieves a better performance than using them separately [57]. Mahmoud et al. proposed a framework for the fusion of local and global features (geometric and texture features) suitable for real-time applications, even in the presence of occlusions, noise, and illumination changes, and evaluated it with MMI, CK+, and static faces in the wild (SFEW) datasets [58]. Zhou et al. [59] used images from different databases to train a neural network for emotional state detection from facial expressions targeted to learning environments. Sini et al. [60] developed a system that analyzed facial expressions to prevent the occupants of the vehicle from experiencing bad feelings by adapting the vehicle driving style depending on their mood.

Instead of determining emotion categories, there have also been studies aimed at detecting emotion-dimensional values. For example, in [61], a three-stage method was proposed to learn the hierarchical emotion information context for predicting affective dimension values from video sequences. In the first stage, a feed-forward neural network is used to generate a high-level representation of the input features. In the second stage, a bidirectional long short-term memory learns the context information of the feature sequences from the high-level representation and obtains the initial recognition results of the input. In the third and final stages, another bidirectional long short-term neural network learns the context information in an unsupervised manner to correct the initial recognition results and obtain the final results.

In addition, [62] reviewed the existing novel machine and deep learning networks proposed by researchers specifically designed for FER based on static images, presented their merits and demerits, and summarized their approach.

III. METHOD

Efficient facial feature extraction is a crucial step toward accurate facial expression recognition. To encode facial characteristics, the starting point was the FACS system proposed by Ekman and Friesen [3]. The Facial Action Coding System (FACS) [3] describes facial expressions by action units (AUs), 30 of which are anatomically related to the contractions of specific facial muscles and can occur either singly or in combination [12]. As mentioned in the previous section, EMFACS [42], based on FACS, was developed in which a subset of action units related to emotions was considered. Table 1 shows the action units (AUs) and the categorical emotions related to them: Happiness (Ha), Sadness (Sa), Surprise (Su), Fear (Fe), Anger (An), and Disgust (Di). It should also be noted that not only the AUs associated with the emotions have to be considered; the intensity related to each action taking place must also be considered. FACS has a 5 intensity scale in order to delimit the intensity grade of each AU.

Lucey et al. [20] found that using categorical labels as ground truths by themselves is highly unreliable, as these impersonations often vary from the stereotypical definition
outlined by the FACS, which can cause errors in the ground truth data that affect the training of emotion recognition systems. Consequently, they labeled their databases according to FACS-coded emotion labels using a selection process. In addition, they included the Contempt (Co) emotion. Table 2 shows the FACS encoding of emotions by [20].

**Algorithm 1** Key Emotional Photogram Extraction and Parameterization

| Input: Videos with people displaying emotions  |
|------------------------------------------------|
| **Output:** List of parameterized and emotionally labeled photograms |
| for Each video do |
| Action Unit based photogram parameterization |
| end for |
| for Each parameterized video do |
| for Each parameterized photogram do |
| if A photogram is adequately associated with emotion then |
| Label photogram with corresponding emotion |
| Add labeled and parameterized photogram within key emotional photograms |
| end if |
| end for |
| end for |

The method used in this study requires parameterizing all photograms to extract the key emotional frames from videos in a given database. To do so, facial features must first be extracted for all the photograms in all videos. Once done, parameterized photograms must be analyzed according to the facial expression-based emotion recognition evidence found in the literature. Both the FACS and the refined emotional feature selection process proposed by [20] are used to label each photogram with its corresponding emotion if any of the criteria for a given emotion is met. Algorithm 1 depicts the key emotional photogram parameterization and extraction process.

Once the set of key emotional photograms was obtained, in order to validate its usefulness for emotional classification of both images and video clips, two further steps were performed, one for image classification and the other for video clip classification.

The first step checks the validity of the obtained set of emotionally categorized photograms. To this end, different classification algorithms must be used on a set of photograms, and their performance is analyzed. Good performance in different algorithms indicates that the set of photograms is sufficiently suitable for achieving a good classification.

After the photogram emotion recognition classifiers were trained and the set of photograms validated, all photograms in videoclips were parameterized and classified using different classification algorithms for photogram classification. This process is described in Algorithm 2.

**Algorithm 2** Videoclip Photogram Classification

| Input: Emotionally labeled photogram-based classifiers and videoclips to be analyzed  |
|------------------------------------------------|
| **Output:** Videoclip emotion recognition |
| for Each videoclip do |
| for Each classifier do |
| for Each videoclip photogram do |
| Classify photogram |
| end if |
| end for |
| end for |

Once all photograms have been classified with all classification algorithms, one matrix is obtained per video, $A_{mn}$, where $m$ is the number of photograms in the video, and $n$ is the number of classifiers. Each cell in $A$ stores the emotion that classifier $n$ assigns to photogram $m$. To determine the category assigned to a video, three different strategies were used.

**A. STRATEGY 1: SINGLE BASE CLASSIFIER MAJORITY VOTING**

There are $n$ possible outcomes: one per classifier from $1 \leq i \leq n$. The emotion is assigned to an emotion that has
the highest number of votes:

\[
clip\_emotion\_\text{clf}_i = \text{mode}_{1 \leq j \leq m} \left( A_{1,i}, A_{2,i}, \ldots, A_{m,i} \right)
\]

where \( m \) is the number of photograms in each video, and \( clf \) returns the classification of the given classifier for the \( ith \) photogram of the video clip. If there is a tie, then the assigned emotion corresponds to the emotion with fewer instances in the set of emotionally classified photograms. This strategy was applied to all classifiers, and the outcome was the one with the highest accuracy in our study.

**B. STRATEGY 2: MAJORITY VOTING PER PHOTOGRAM**

Majority voting is applied to each photogram from among all the classifiers, with the outcome as the consensus emotion for the photogram. Then, majority voting is applied to all consensus photogram classifications to obtain the emotion for the video:

\[
clip\_emotion = \text{mode}_{1 \leq j \leq n} \left( \text{mode}_{1 \leq j \leq m} \left( A_{1,i}, A_{2,i}, \ldots, A_{n,j} \right) \right)
\]

where \( m \) is the number of photograms in each video, \( n \) is the number of classifiers, and \( A(i, x) \) \((1 \leq x \leq m)\) returns the classification of the \( x \)th classifier for the \( ith \) photogram of the video clip. The statistical mode was used to determine the emotions assigned to the photograms. The same process was applied to all photograms in the sequence. The above-mentioned criterion applies when there are ties.

**C. STRATEGY 3: SIMPLE MAJORITY VOTING**

In this case, majority voting was calculated for all the cells in \( A \). The same criteria apply when there are ties.

\[
clip\_emotion = \text{mode}_{1 \leq j \leq n, 1 \leq j \leq m} \left( A_{i,j} \right)
\]

The emotion assigned to the entire videoclip is determined by the statistical mode of the values of all cells in matrix \( A \).

All of the above strategies were devised by considering all the possible classifiers. As a final refinement in the process, all possible classifier combinations are considered; therefore, different columns of \( A \) will be considered depending on the classifiers selected in a given combination. For \( n \) classifiers, the number of combinations to study is \( 2^n - 1 \).

**IV. STUDY**

In this section, the materials used in this study are described. Next, the facial characteristics considered were specified. Subsequently, the supervised classification techniques used for emotion classification, depending on the emotional categories considered, were presented. Finally, the experimental setup was explained.

**A. MATERIAL**

Two databases with facial expression recordings were employed in this study: MUG [30] and Extended Cohn–Kanade (CK+) [20]. All of them had recordings and metadata to label each recording.

The MUG dataset included 86 subjects (35 women and 51 men). Image sequences begin and end at a neutral state and follow the onset, apex, and offset temporal patterns. For each of the six emotions, the authors recorded a few image sequences of various lengths. Each sequence contained 50–160 images. In addition, for each subject, a short image sequence depicting the neutral state was recorded. Sequences with correct imitation of the expressions were selected as part of the database. This database is publicly accessible [31]. A total of 980 emotionally categorized image sequences with 139,100 photograms were included in the MUG database.

The Extended Cohn-Kanade (CK+) dataset contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age, with a variety of genders and heritage. Each video shows a facial shift from neutral to targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either 640 × 490 or 640 × 480 pixels. From these videos, 327 were labeled with one of seven expression classes (anger, contempt, disgust, fear, happiness, sadness, and surprise) and included 5,876 photograms. The CK+ database is widely regarded as the most extensively used laboratory-controlled facial expression classification database and is used in most facial expression classification methods. Each of the expression sequences reflects the expression from the neutral emotion to the apex of the emotion [20].

**B. FACIAL LANDMARKS AND ACTION UNITS**

To extract facial characteristics, Open Face [47] software was employed. This software enables the identification of faces in video photograms, and can then be used to extract a set of 709 facial features related to the location and rotation of the face, the direction of the look, the location of the parts of the face in 2D and 3D, and whether AUs are present and if they are, their intensity. Figure 1 shows the face detection process made by OpenFace starting from photograms with a photogram displaying anger in the CK+ database, and the association of AUs with the extracted face.

**FIGURE 1.** Face detection and AU association on a CK+ database photogram [20].

A script for automatic processing to associate emotional categories with particular photograms was developed, taking facial characteristics and the emotions represented in each video. The outputs given by OpenFace were analyzed to search for AUs activation related to each emotion in the video photograms.
Photograms that activated the AUs corresponding to the emotion in the video, according to [3] and [20], were collected. As the MUG database included video sequences with neutral emotion, photograms with complete absence of emotion-related AUs were categorized as neutral in the tagged video sequences. As there were no whole image sequences specifically displaying neutral emotion at all times, no neutral photograms were categorized for CK+, as no specific photogram selection was intended to be performed on the databases, even though image sequences start with a neutral pose. Table 3 shows the amounts of emotionally labeled photograms for both MUG and CK+ using Ekman FACS alone. Table 4 shows the amount of emotionally labeled frames for each dataset and emotion using both Ekman’s FACS [3] and the qualifying criteria by Lucey et al. [20]. In this case, it emotionally categorized 33,187 photograms in the MUG dataset and 3,561 in the CK+ dataset, which corresponded to 23.85% and 60.60% of the overall photograms in the MUG and CK+ databases.

### TABLE 3. Amount of emotionally categorized photograms in each dataset after parameterization (Ekman FACS alone).

| Dataset | Sa | Fe | Ha | An | Su | Di | Co |
|---------|----|----|----|----|----|----|----|
| MUG     | 537| 0  | 6,922 | 155 | 2,795 | 326 |
| CK+     | 36 | 0  | 949 | 74 | 180 | 11 |

### TABLE 4. Amount of emotionally categorized photograms in each dataset after parameterization (Ekman and Lucey).

| Dataset | Sa | Fe | Ha | An | Su | Di | Co | Ne |
|---------|----|----|----|----|----|----|----|----|
| MUG     | 548| 800| 7,838| 4,663| 9,177| 8,627| -  | 1,534|
| CK+     | 36 | 108| 1,079| 768 | 964 | 424| 182| -  |

### C. SUPERVISED CLASSIFICATION ALGORITHMS

Ten supervised classification algorithms were selected for data analysis. The selection was made to include algorithms with different classification paradigms, including rule-based, tree-based, distance-based, probabilistic, and function-based algorithms. Only basic algorithms with their default configurations were selected; that is, no hyperparameter tuning strategies were used.

The algorithms employed are briefly described next.

1) **BAYESIAN NETWORKS**

A Bayesian network [63] or directed acyclic graphical model is a probabilistic model that represents a set of aleatory variables and their conditioned independencies by means of a directed acyclic graph.

2) **LOGISTIC REGRESSION**

This is also called logit regression or logit model, and in statistics, it is considered a regression model where the dependent variable is categorical [64].

3) **SUPPORT VECTOR MACHINES (SVMs)**

These are a set of associated supervised learning methods employed for classification and regression [48]. Taking the input data as two vector sets in an n-dimensional space, an SVM builds a hyperplane in this space to maximize the margin between both sets.

4) **k-NEAREST NEIGHBORS (k-NN)**

This is a classifier algorithm of the nearest neighbor based on cases [50]. To classify a new sample, it takes a simple distance measure to find the nearest training instance to the test instance by employing the same training instance set to predict it.

5) **REPEATED INCREMENTAL PRUNING TO PRODUCE ERROR REDUCTION (RIPPER)**

The rule-based learning system presented in [65] creates rules by means of a repeated growing process (to adjust training data) and pruned (to prevent overfitting). RIPPER manages multiple classes, ordering them from the least to the most prevalent, and treating each class in order as two different class problems.

6) **ONE RULE**

This is a one-level decision tree that tests only one attribute [66]. The selected attribute yielded a minimum error.

7) **PART**

This uses a divide-and-win strategy to construct a C4.5 tree with partial decision in each iteration and to generate a rule based on the best tree leaf, as part of the decision list [67].

8) **HOEFFDING TREE**

This is an algorithm to induct an incremental decision tree that can learn from massive data transmission, assuming that the distribution does not change with time [68].

9) **C4.5**

This is a classification model based on a C4.5 decision tree [69]. The tree is built at the top bottom, dividing the training set, and starts selecting the best variable in the tree root.

10) **RANDOM FOREST (RF)**

This builds a forest by combining non-pruned trees [70].

The Weka software package [71] was used for the selected algorithms. This package includes automatic learning algorithms for data mining. Table 5 specifies the implementations of the algorithms mentioned in Weka, ordering them depending on their types.

The selected algorithms were first used to evaluate the emotionally categorized photograms obtained from the two databases. Subsequently, to categorize the video sequences, the algorithm outcomes were evaluated for all photograms in
### TABLE 5. Used algorithms with their implementation in Weka.

| Algorithm type | Algorithm | Weka implementation |
|----------------|-----------|---------------------|
| Bayes Functions | Logistic Regression | Logistic |
| Support Vector Machines | SMO | SMO |
| Lazy Rules | k-Nearest Neighbors | IBK |
| RIPPER | JRIP | JRIP |
| One Rule | OneR | OneR |
| PART | PART | PART |
| C4.5 | Hoefding Tree | Hoefding Tree |
| Random Forest | Random Forest | Random Forest |

Each video. The outcome was the emotion with the highest number of photograms of the sequence labeled as the selected emotion. In the event of ties, instead of solving them arbitrarily, the emotion with the smallest number of samples in the emotionally labeled photograms was selected.

Finally, three majority voting strategies were used based on the emotional classifications obtained by each algorithm for each photogram. In the first strategy, the classification results of all algorithms for each photogram were used to determine the emotion of each photogram by majority voting. In the second strategy, as the databases contained more emotional photograms than neutral ones, voting was performed with all the emotional labeling of the photograms of a video sequence by a single classifier. In the third strategy, the photogram labeling of all algorithms is considered when categorizing a video sequence by a majority vote. In the event of ties, the outcomes of the algorithm combined with better results were used. All possible algorithm combinations using the abovementioned strategies were explored to seek further optimization in the results without preprocessing the video sequences or selecting specific frames.

### D. EXPERIMENTAL SETUP

Files with the image sequences from all databases were first parameterized using OpenFace, extracting the set of facial features for each one. Then, a script was executed to select the set of emotionally relevant photograms for each database. Subsequently, the experiments on selected \( n = 10 \) base classifiers were performed using their implementations in Weka using their default configurations. Thereafter, each selected classifier classified all photograms in every image sequence. Using detailed voting strategies, the emotions of image sequences were obtained. Finally, all possible classifier combinations were analyzed, with a total of \( 2^n - 1 = 1,023 \).

Selected algorithms were employed with the default setting values given by Weka and also with 10-fold cross-validation [72] to achieve a validated classification precision.

### V. RESULTS

Table 6 shows the results for emotion recognition in the samples of emotionally categorized photograms for both the MUG and CK+ datasets. For both bases, the accuracy was 10 fold cross-validated for each algorithm. The outcomes for each dataset and classifier are shown together with the median and standard deviation. The best outcomes are highlighted in bold. The results were person-independent in both the cases. The models were trained with seven emotions for MUG (Sadness, Fear, Happiness, Anger, Surprise, Disgust, and Neutral), and seven for CK+ (Sadness, Fear, Happiness, Anger, Surprise, Disgust, and Contempt), following the emotion label in the video sequences for both databases.

As can be seen, for the MUG dataset, the highest accuracy was achieved with k-NN and RF classifiers (99.80%), slightly better with k-NN, whereas for the CK+ dataset, the best accuracy (99.49%) was achieved with k-NN. In both cases, the lowest accuracy was obtained with the OneRule classifier: 46.98% for the MUG and 55.63% for the CK+ dataset.

Table 7 shows the results of classifying video clips for emotion recognition in both MUG and CK+ datasets according to the three defined strategies using trained photogram classifiers. Here again, the accuracy for each dataset and classifier is shown with the addition of the medium and standard deviation for each dataset. The best results, highlighted in bold, were RF for the MUG (70.71%) and the Bayesian

### TABLE 6. Emotion recognition classification outcome in the sets of emotional photograms.

| Classifier    | MUG Accuracy (%) | CK+ Accuracy (%) |
|---------------|------------------|-----------------|
| BayesNet      | 73.91            | 83.46           |
| Logistic      | 99.12            | 99.21           |
| SMO           | 98.51            | 99.41           |
| IBK           | 99.80            | 99.49           |
| JRIP          | 98.13            | 95.39           |
| OneR          | 46.98            | 55.63           |
| PART          | 98.57            | 96.07           |
| Hoefding Tree | 76.11            | 70.29           |
| J48           | 98.27            | 96.07           |
| Random Forest | 99.80            | 99.24           |
| Median        | 88.92            | 89.43           |
| Standard Deviation | 17.78        | 15.13           |

### TABLE 7. Emotion recognition classification for video clips by the different photogram classifiers trained and defined majority voting strategies.

| Classifier    | MUG Accuracy (%) | CK+ Accuracy (%) |
|---------------|------------------|-----------------|
| BayesNet      | 51.53            | 66.36           |
| Logistic      | 61.73            | 60.86           |
| SMO           | 64.08            | 61.77           |
| IBK           | 70.20            | 65.14           |
| JRIP          | 61.94            | 62.70           |
| OneR          | 57.24            | 63.30           |
| PART          | 62.76            | 63.61           |
| Hoefding Tree | 34.80            | 63.61           |
| J48           | 63.98            | 63.61           |
| Random Forest | 70.71            | 63.52           |
| Median        | 59.90            | 63.52           |
| Standard Deviation | 10.44        | 1.57            |
Network for the CK+ (66.36%), while the worst results were the Hoeffding Tree for the MUG (34.80%) and Logistic for the CK+ (60.86%).

Table 8 displays the results obtained for classifying video clips for emotion recognition in both MUG and CK+ datasets according to the three defined strategies with trained photogram classifiers and analyzing all possible photogram classifier combinations. In this case, only the best results achieved are shown for each of the defined strategies. The best result for the MUG database was achieved with a combination of IBK and Random Forest with 72.55%, while 88.37% was achieved for CK+ with a combination of IBK and OneR with Strategy 1 and a combination of Logistic and BayesNet with Strategy 3.

![Table 8](image-url)

**TABLE 8.** Emotion recognition classification for video clips by the best classifier voting combination for each used strategy.

| Dataset | S1 (%) | Comb | S2 (%) | Comb | S3 (%) | Comb |
|---------|--------|------|--------|------|--------|------|
| MUG     | 71.53  | IBK, Random Forest | 72.04 | IBK, Random Forest | 72.55 | IBK, Random Forest |
| CK+     | 88.37  | IBK, OneR | 76.45 | IBK, Hoeffding Tree, BayesNet | 88.37 | Logistic, BayesNet |

**VI. DISCUSSION**

This work intends to provide a method for facial emotion recognition based on facial features extracted from video photograms. Nowadays, face representation parameterizers are mature enough to be considered for facial expression recognition (FER), mainly in static images. Therefore, combining FER with facial emotion recognition literature can lead to proper emotional photogram detection in video sequences. The method provided has allowed valid sets of parameterized photograms to be obtained that include a large amount of database photograms that can be classified adequately, showing high accuracy for the photograms in the sets, even when experiments have been performed with different non-optimized classifiers. When applied to video sequences, individual photogram classification was performed on all photograms in videos, and videos were categorized based on majority voting for their photograms with the different voting strategies used. Calculating the outcomes of all the different classifier combinations improved the results. The results obtained allow the presented approach to be validated as, without any kind of photogram selection or preprocessing, it improved the percentage of identified emotional photograms and correctly classified up to 72.55% and 88.37% of the video sequences with two publicly available and widely used databases.

Although both databases used were intended for facial expression recognition, they followed different approaches when their actors displayed emotions. In CK+ recordings, a single neutral-to-peak emotion transition was observed, while the neutral state – onset – apex – offset temporal pattern was applied in the MUG dataset. When applying the proposed method, this led to a notably higher amount of emotional photogram identification in CK+ compared to MUG (60.60% and 23.85%), as video sequences were shorter and ended in peak emotion in CK+, while MUG displayed larger onset and offset. However, when trained photogram classifiers are applied to video sequences, accuracy improvement compared to detected emotional photograms (displayed in Table 4) is much larger in the MUG database (from 23.85% of photograms automatically labeled as emotional to 70.71% correctly detected emotional videos) in comparison with the CK+ database (from 60.60% to 66.36%). As all the photograms were analyzed, the amount of emotional photograms in the videos with the temporal pattern in MUG influenced the accuracy improvement compared to the neutral – peak approach in CK+. Regarding classifier combinations, the improvement was greater in CK+, as it improved by 22.01% in absolute value (from 66.36% to 88.37%), while the improvement was 1.84% in MUG (from 70.71% to 72.55%). In this case, a proper classifier combination was more effective in CK+, as simple classifier combination strategies were particularly effective in flat search spaces where all the classifiers exhibited similar capabilities [73].

The critical choice of not performing any kind of preprocessing or photogram selection when applying photogram classifiers to video sequences has influenced the accuracy of the achieved results. Preprocessing allows the input images to be modified or redundant photograms to be eliminated. Selecting specific photograms has proved to be effective as long image sequences are reduced to a small number when the transitions between neutrality and emotions are previously known for a given database [74]. In addition, as the CK+ videos did not include neutral sequences, neutral emotions were not included. Although it would add one more classification category, video accuracy is likely to be better because starting neutral photograms would be classified as neutral and emotional photograms around the peak count more in majority voting depending on the neutral state length. Accordingly, majority voting results for videos presented in Table 7 can be considered worst-case scenarios when using photogram-based classifiers as the basis for video classification. Nevertheless, the decisions adopted make this approach valid for any database, independent of the approach used for recording and without prior knowledge of transitions or temporal properties.

With regard to photogram set selection, using categorical labels as ground truth by themselves is highly unreliable, as these impersonations often vary from the stereotypical definition outlined by FACS, which can cause errors in the ground truth data, which affects the training of emotion recognition systems. The selection process followed by Lucey et al. [20] allowed sets of photograms to be selected, which included a significant amount of photograms from used databases that have proven to be easily classifiable.
Regarding the limitations of the presented work regarding photogram identification, in contrast to most recent literature, we used the OpenFace parameterizer to extract facial expressions. Although this is a widely used parameterizer, it is possible that not all action units have been adequately identified. In this regard, it must be noted that OpenFace has been trained using multiple databases, including CK+ [47].

Direct comparisons with other works in the literature cannot be made, as this work is based on majority voting on automatically extracted facial features of all classified photograms in video sequences without any preprocessing or photogram selection. Some limitations are common in most cases, including common performance metrics with which to evaluate new algorithms for both AU and emotion detection or standard protocols for common databases to enable quantitative meta-analyses. The cumulative effect of these factors has made benchmarking various systems very difficult or impossible. In addition, some authors employed a leave-one-out cross-validation strategy on the database, while others chose another random train/test set configuration. Other authors have also reported results on the task of broad emotion detection, even though no validated emotion labels were distributed with the dataset. The combination of these factors makes it very difficult to gauge the current state-of-the-art in the field, as no reliable comparisons have been made [20].

Many recent references found in the literature tackle facial emotion recognition with deep learning, such as static images or videos. Regarding image-based classification, [75] used instances from different databases to perform static image recognition of Big Six emotions, with neutral images from externally used databases. [76] presented a partially connected multilayer perceptron neural network with images from different databases. For preprocessing, images were normalized to specific pixel amounts based on different criteria [77]. In some cases, manually annotated pictures were used, as in [78]. Data augmentation techniques are also used because deep networks easily overfit when a large number of sequences are not used [79]. Overfitting caused by a lack of sufficient training data is also an important issue to consider when working with deep learning models [80].

In most studies on deep learning, the video sequences found in the literature are preprocessed, and prior knowledge about transitions in the databases is used when developing deep learning models. Following evaluation protocols that select frames in image sequences, images for model training are obtained, typically neutral expression frames at the beginning and then a different number of frames near the emotional peak of the sequence [74], [81]–[84]. The original image sequences were also normalized at a specific number of frames to train the deep models [85]. In other cases, only those video clips that have one of the labeled emotions and a neutral frame at the beginning are selected [86].

**VII. CONCLUSION AND FUTURE WORK**

The method detailed in this work has proven that extracting emotional photograms from datasets can be used to train classifiers for photogram-emotion recognition. This method is independent of the temporal patterns in the datasets used. It has proven effective for proper emotion recognition in two publicly available datasets, even when using weak classifiers. The results obtained, although difficult to compare with other studies due to methodological differences when performing studies, may be considered good as videos are analyzed throughout their entire length, and no preprocessing or specific photogram selections were made. The percentage of emotional photograms identified was significant compared to all photograms in the analyzed databases, and even with simple classifiers and different combinations of them, results can clearly be improved upon.

With regard to future work, facial feature information extracted with a detailed method can be used to automatically select photograms that represent transitions between neutral and emotional states, thus improving deep learning method-based facial emotion recognition. Furthermore, classification techniques that are already being used can be changed and adjusted to obtain better results. Moreover, the method can be applied to other databases to obtain larger sets of labeled key emotional photograms that can be used to perform cross-cultural studies.

**REFERENCES**

[1] K. Höök, “Affective computing,” in *The Encyclopedia of Human-Computer Interaction*, 2nd ed., M. Soegaard and R. F. Dam, Eds. Aarhus, Denmark: The Interaction Design Foundation, 2012.

[2] P. J. Lang, “The emotion probe: Studies of motivation and attention,” *Amer. Psychol.*, vol. 50, no. 5, pp. 372–385, 1995.

[3] P. Ekman and W. V. Friesen, *Pictures of Facial Affect*. New York, NY, USA: Basic Books, 1980.

[4] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms,” *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, 2003.

[5] M. Simón, E. Sarasua, B. Gamecho, E. Larraza-Mendiluze, and N. Garay-Vitoria, “DAFIESKU: A system for acquiring mobile physiological data,” *Mobile Inf. Syst.*, vol. 2017, Nov. 2017, Art. no. 7261958.

[6] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Distributed semantic models for affective text analysis,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 11, pp. 2379–2392, Nov. 2013.

[7] I. Leifer, G. J. Burghouts, and L. J. M. Rothkrantz, “Recognizing stress using semantics and modulation of speech and gestures,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 162–175, Apr. 2016.

[8] N. Garay-Vitoria, I. Cearreta, and E. Larraza-Mendiluze, “Application of an ontology-based platform for developing affective interaction systems,” *IEEE Access*, vol. 7, pp. 40503–40515, 2019.

[9] A. Ortiz, “Avatares para la interacción emocional,” Ph.D. dissertation, Dept. Comput. Archit. Technol., Universidad del País Vasco-Euskal Herriko Unibertsitatea, Donostia-San Sebastián, Spain, 2008.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[11] N. C. Ehbner, M. Riediger, and U. Lindenberger, “FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation,” *Behav. Res. Methods*, vol. 42, no. 1, pp. 351–362, 2010.

[12] Y.-L. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[13] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, and A. Ross, “Automated facial expression recognition system,” in *Proc. 43rd Ann. Int. Carnahan Conf. Secur. Technol.*, Oct. 2009, pp. 172–177.

[14] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.
[81] J. Zhao, X. Mao, and J. Zhang, “Learning hierarchical emotion context for continuous dimensional emotion recognition from video sequences,” *IEEE Access*, vol. 7, pp. 62894–62903, 2019.

[82] P. Rodriguez, G. Cucurull, J. Gonzalez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, “Deep pain: Exploiting long short-term memory networks for facial expression classification,” *IEEE Trans. Cybern.*, early access, Feb. 9, 2017, doi: 10.1109/TCYB.2017.2662199.

[83] K. Zhang, Y. Huang, Y. Du, and L. Wang, “Facial expression recognition based on deep evolitional spatial-temporal networks,” *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[84] Y. Tang, X. Zhang, and H. Wang, “Geometric-convolutional feature fusion based on learning propagation for facial expression recognition,” *IEEE Access*, vol. 6, pp. 42532–42540, 2018.

[85] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, “Deep temporal appearance-geometry network for facial expression recognition,” 2015, arXiv:1503.01532. [Online]. Available: http://arxiv.org/abs/1503.01532

[86] A. Majumder, L. Behera, and V. K. Subramanian, “Automatic facial expression recognition system using deep network-based data fusion,” *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 103–114, Jan. 2018.

---

**JUAN MIGUEL LÓPEZ-GIL** received the Ph.D. degree in computer science from the University of the Basque Country (UPV/EHU), in 2007. He was a Visiting Professor at the University of Lleida, from 2007 to 2010, where he was the Academic Director of the master’s program on human–computer interaction. He has been a Lecturer at the UPV/EHU, since 2010, where he is currently an Associate Professor at the Computer Languages and Systems Department. His current research interests include user interface usability and accessibility evaluation, adaptive user interfaces, and affective computing in Galan Research Group (https://galan.ehu.eus/Galan/).

**NESTOR GARAY-VITORIA** received the M.S. degree in informatics and the Ph.D. degree in computer science from the University of the Basque Country (UPV/EHU), Donostia-San Sebastian, in 1990 and 2000, respectively. He was the first Academic Director of the master’s program on assistive technology for personal autonomy given at UPV/EHU. He is interested on computer science education at all educational levels. He is currently an Associate Professor with the Computer Architecture and Technology Department, UPV/EHU. His research interest includes human–computer interaction for special needs at the Egokituz Research Group (http://egokituz.org/en/). He received the Extraordinary Award for his Ph.D. dissertation, in 2002.