Abstract—Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists since the work of Darwin in 1872. It includes both measurement of facial motion and recognition of expression. There are two different ways to analyze facial expressions: one considers facial affect (emotion) and the other facial muscular movements. Many researchers argue that there is a set of basic emotions which were preserved during evolutive process because they allow the adaption of the organisms behavior to distinct daily situations. This paper discusses emotion recognition based on analysis of facial elements. Different feature sets are proposed to represent the characteristics of the human face and their performance is evaluated using Machine Learning techniques. The results indicate that the selected facial features represent a valid approach for emotion identification.

I. INTRODUCTION

Emotion can be defined as a neural impulse which leads a living organism to a certain action. They are relevant mechanisms of social communication and behavior, as well as for living in society. Emotion recognition is important not only to interpret human behaviors and intentions. Emotions strongly influences human cognitive processes, specially those related to decision making and problem solving [10]. In literature, six basic emotions can be found according to [16], these emotions are happiness, sadness, anger, surprise, fear and disgust, which allow human beings to adapt to distinct daily situations. The development of systems or social robots able to interpret and act emotionally raises possibilities to improve human-computer interactions, as well as allows the study and proposal of statements related to human interaction and behavior [38] [3].

Emotions have been described by psychologists in terms of discrete categories, which map different emotions to distinct categories. The most popular example of this description is the six basic emotion categories, specially supported by the cross-cultural studies conducted by Ekman [13], indicating that humans perceive certain basic emotions conveyed by facial expressions in the same way, regardless of culture. The main advantage of a category representation is that people use this categorical scheme to describe observed emotional displays in daily life. The labeling scheme based on category is very intuitive and thus matches people’s experience. However, discrete lists of emotions fail to describe the range of emotions that occur in natural communication settings [54].

An alternative to the categorical description is the dimensional description [42], in which emotions are described as points in a multidimensional space, using continuous scales or dimensional bases. An affective state is characterized in terms of a small number of latent dimensions rather than in terms of a small number of discrete emotion categories. In this view, affective states are not discrete and independent of each other. Instead, they are systematically related to one another. These dimensions include evaluation, activation, control, power, etc. The evaluation and activation dimensions are expected to reflect the main aspects of emotion. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive. In contrast to categorical representation, dimensional representation enables to label a wide range of emotions. However, the projection of the high-dimensional emotional states onto a rudimentary 2D space results in loss of information: some emotions become indistinguishable or are placed outside the space (e.g., fear, anger, surprise) [54].

One recent theory of emotions in modern psychology is the appraisal-based approach [45], which can be seen as the extension of the dimensional approach. In this representation, an emotion is described through a set of stimulus evaluation checks, including the novelty, intrinsic pleasantness, goal based significance, coping potential and compatibility with standards. However, translating this scheme into one engineering framework for purposes of automatic emotion recognition remains challenging [46].

Automatic emotion recognition is a multidisciplinary task involving different research fields, including psychology, computer vision, speech analysis and machine learning. Different methods are naturally used by humans to express emotions, such as voice intonation, body movements and facial expressions. However these natural human actions require complex calculations for computers, reason why researchers investigate different approaches and techniques aiming to detect emotions [28]. Even though body movements or voice are used to express them, emotions are more precisely described by facial expressions, without the need to analyze gestures or voice [18] [19] [12].

Facial expression has been a focus of emotion research for over a hundred years [12] and is central to several leading theories of emotion [15] [25] [48]. Facial expression figures in research on almost every aspect of emotion, including psychophysiology [29], perception [1], social processes [24], emotional disorders [26], and others [8]. Cross-cultural psychological research on facial expressions also indicates that there may be a small set of universal facial expressions.
related to the six basic emotions [14].

Most of the existing facial expression recognizers employ various pattern recognition approaches and are based on 2D spatiotemporal facial features. The usually extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, nose) and the location of facial salient points (corners of the eyes and mouth) or appearance features representing the facial texture, including wrinkles, bulges and furrows. Some examples of geometric-feature-based methods are [4], [49] and [37]. For appearance-feature-based methods it is possible to mention [2] and [30]. There are also studies using both geometric and appearance features [31].

In this sense, this paper focuses on emotion recognition based only in facial expression analysis of frontal pictures. More specifically, different feature sets, based on geometrical features, are proposed and analyzed by three Machine Learning (ML) techniques for the task of emotion recognition. The feature sets use information about facial points, angles, distances and also areas to map the human face. The chosen ML techniques are Multilayer Perceptron Networks (MLPs), based on bio-inspired paradigm, Support Vector Machines (SVMs), based on statistical theories and C4.5 algorithm, based on decision tree rules. The obtained results are quite promising and will be presented and discussed along this paper.

II. FACIAL EXPRESSION ANALYSIS

Facial expressions are the facial changes in response to a person’s emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists since the work of Darwin in 1872 [11]. Automatic facial expression analysis can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human-computer interface (HCI).

Facial expression analysis includes both measurement of facial motion and recognition of expression. There are two different ways to analyze facial expressions: one considers facial affect (emotion) while the other considers facial muscular actions [36] [47]. These two streams stem directly from the two major approaches to facial expression measurement in psychological research [7]: message and sign judgement. The aim of the former is to infer what underlies a displayed facial expression, such as emotion or personality, while the aim of the latter is to describe the physical appearance of the expressed behavior, such as facial movement or facial component shape [49].

While message judgement is all about interpretation, sign judgement is agnostic, independent from any interpretation attempt, leaving the inference about the conveyed message to higher order decision making. The most commonly used facial expression descriptors in message judgement approaches are the six basic emotions [16], Plutchik’s emotion wheel [39] and Russell’s circumplex model [43].

The task of automatic facial expression analysis can be divided into three main steps: face detection, facial feature extraction and classification. Face detection is a processing stage to automatically find the face region for the input images or sequences. It can be a detector to detect face for each frame or just detect face in the first frame and then track the face in the remainder of the video sequence. Detecting a face in a complex scene is non-trivial problem. Most of the existing systems for expression analysis require the face to be nearly in frontal view under defined conditions. For these systems, information about the presence of a face and its coarse location in the scene usually has to be given. Still, the exact location, scale and orientation has to be determined by hand or by an automatic tracking system. Head motion, occlusion, changing illumination, presence of hair, glasses or jewelry are examples of possible complications for the system [49].

After the face is detected, it is time to extract and represent the facial changes caused by facial expressions. Several types of perceptual cues to the emotional state are displayed in the face: relative displacements of features (raised eyebrows), textural changes in the skin surface (furrowing the brow), changes in skin hue (blushing) and others. Depending on how the face and its expression are modeled, features have to be designed so as to condense this information or a part of it into a set of numbers to build the base for the classification, and therefore primarily decide about the quality of the final analysis result [47]. In facial feature extraction for expression analysis, there are mainly two types of approaches: geometric feature-based methods and appearance-based methods. The geometric facial features present the shape and locations of facial components (including mouth, eyes, eyebrows and nose). The facial components or facial feature points are extracted to form a feature vector that represents the geometry of the face. The appearance facial features present the appearance (skin texture) changes of the face, such as wrinkles and furrows. The appearance features can be extracted from either the whole-face or specific regions in a face image to generate a feature vector. With appearance-based methods, image filters, such as Gabor wavelets, are applied [47].

To represent facial expressions, a system can use geometric features only [5], appearance features only [21], or hybrid features (both geometric and appearance features) [52]. Research shows that the use of hybrid features can achieve better results for some expressions. To remove the effects of variation in face scale, motion, lighting and other factors, one can first align and normalize the face to a standard face (2D or 3D) manually or automatically [52], and then obtain normalized feature measurements by using a reference image (neutral face).

Facial expression classification is the last stage. The facial changes can be identified as facial action units (muscular movements) [17] or prototypic emotional expressions. Depending on whether the temporal information is used, the recognition approaches are classified as frame-based or
sequence-based. The frame-based recognition approach uses only the current frame with or without a reference image (mainly it is a neutral face image) to recognize the expressions of the frame. The sequence-based recognition approach uses the temporal information of the sequences to recognize the expressions for one or more frames. Many classifiers have been applied to expression recognition such as neural networks (NNs), support vector machines (SVMs), linear discriminant analysis (LDA), K-nearest neighbor, multinomial logistic ridge regression (MLR), hidden Markov models (HMM), tree augmented naive Bayes, and others. Some systems use only a rule-based classification according to the definition of the facial actions [47].

A. Proposed Facial Analysis

To recognize emotions based on facial expressions, four distinct feature sets to represent the characteristics of the face have been proposed, in order to analyze which one best fits the problem. To analyze a face, first a facial detection system is required. To do this task, Face Tracker system, developed by [44], was employed.

Face Tracker uses an optimization strategy for local experts-based deformable model fitting. Deformable model fitting employs a linear approximation to how the shape of a non-rigid object deforms. It registers a parametrized model for an image, fitting the landmarks into consistent locations of the analyzed object. Based on a reference model of the face, composed by 66 landmark points, the algorithm tries to align the elements of a new user face with the landmarks of the existent reference model, to fit a model for the current face. This is a complex problem since it works with high dimensional optimizations, in which different light conditions and image noise can drastically change the current face model.

Face Tracker non-parametric approach is reminiscent of the well known and understood mean-shift mode seeking algorithm [22]. The approach differs, however, from the traditional mean-shift algorithm as it is being applied over all landmarks simultaneously and also imposes a global prior over their joint motion. The resulting fitting algorithm is simple and affords significant improvements in convergence rate and accuracy over existing approaches [44]. It can handle partial occlusions of the face and was developed using mechanisms which reduce the computational complexity of the proposed approach.

The Face Tracker algorithm first calculates the sets of possible locations for each landmark of the reference model, through a search in a rectangular region. Then a non-parametric estimation of mapped responses is made. Once the mentioned estimation is obtained, the following steps are repeated until convergence is reached: one linearized shape model is found and the mean-shift vectors are computed. After that, the update of the point distribution model (PDM) parameters is calculated and applied. This approach is based in PDM proposed by [9]. The goal is to search the PDM parameters that jointly minimizes the misalignment error over all landmarks, regularized appropriately. Figure 1 shows a face mapped by Face Tracker.

![Face Tracker mapped face, based on 66 feature points.](image)

All feature sets proposed in this paper, denoted from now on as $FS_1$, $FS_2$, $FS_3$ and $FS_4$, are based on geometric features. To obtain them, Face Tracker was modified to map only 33 feature points, instead of the 66 originally mapped. Just a small set of original points was chosen based on previous experiments in which the face was mapped only by Face Tracker feature points, considering different subsets of points. Each subset was chosen by selecting a different amount of points to describe the elements (mouth, eyes, eyebrows, chin and nostrils) of the face. So, three distinct subsets were analyzed: the 66 original Face Tracker points, a subset of 33 selected points and a reduced subset with only 26 points.

For these three subsets, experiments were done according to the ones described in Section IV, using the mentioned ML techniques. The obtained results were not satisfactory, but they pointed out that mapping the face relying only in feature points is not sufficient to correctly describe a human face for the purpose of emotion recognition. The best results were at maximum 51% in accuracy, and better for the 33 points subset. For this reason that number of points was adopted for the feature sets presented here, but new facial information was added to improve the facial modeling.

Figure 2 presents the graphical elements of the four proposed feature sets, illustrated in an image of the Radboud Faces Database [27]. As mentioned, they are based on 33 points: eight mapping the mouth, six for each one of the eyes, three for each eyebrow and the chin, two for nostrils and two delimiting the lateral extremities of the face near the eyes. The points are represented by red dots. Also, to model the eyes, the mouth and regions of the face related to emotional muscular movements, eight areas are computed, which can be identified by the geometric regions delimited by the black color line segments.

To obtain $FS_1$ and $FS_2$, distances and angles among the 33 feature points are considered. The main difference of both lay in the distances and angles considered. For the first one,
in all possible combinations of points, the distances and the angles that the line connecting two distinct points makes with the horizontal axis are obtained. It creates a feature set of dimensionality $D_1 = 2 \cdot 33 + 8 + 2 \cdot 528 = 1130$.

For the second one, only a subset of the distances and angles above are calculated. We chose only the distances and angles which map the eyes, the mouth, the upper mouth lip with nostrils and the lower mouth lip with the chin. This subset is able to describe mouth and eye states and also movements of the mouth in relation to the chin or nostrils. Considering that the eight areas can correctly represent the remaining emotional facial movements not mapped by the angles or distances which are not calculated, this feature set can indeed be representative, with the advantage of low dimensionality, which is $D_2 = 2 \cdot 33 + 8 + 2 \cdot 107 = 288$.

The main goal of $FS_2$ is to avoid the calculations of all distances and angles from $FS_1$, which can result in redundancy. This extra information may slow down the ML generalization process and lead to the generation of high complexity and specialized classifiers, resulting in poor performance.

Both feature sets possess information about the positions of the 33 points, distances among them and the angles that the line connecting two points makes with the horizontal axis. The difference is mainly that $FS_2$ is a subset of $FS_1$ since it covers all 33 points and the areas, but only a fraction of the angles and distances from $FS_1$. It is worth noting that these feature sets represent only single frame based features, no information about the relation of these measurements to their values in a frame displaying a neutral expression is encoded.

To capture this information, feature sets $FS_3$ and $FS_4$, based respectively on the single frame based feature sets $FS_1$ and $FS_2$, were created. These feature sets compare the change in feature values between the current frame, which represents an emotion, and the frame in which the neutral expression of the subject is present. It is important to highlight they only represent the differences of the values (points coordinates, distances, angles and areas), not the values itself.

These four proposed sets are able to encode distinct aspects of the face: peculiar aspects of the facial expression during each emotion and dynamics or differences of the face according to its neutral state.

Since in the analyzed pictures the subjects can display different head inclinations, rotation and translation are applied to each of them before their analysis by ML techniques. Rotation aligns the face to the horizontal axis and translation puts it in first quadrant. Also, to correct problems of subjects being at distinct distances from the camera and possessing different face formats, each picture is normalized by its intra-eyes distance. These affine transforms generate standardized feature vectors.

### III. MACHINE LEARNING TECHNIQUES

One of the main goals of Machine Learning (ML) is the development of computational methods able to extract concepts (knowledge) from samples [34]. In general, ML techniques are able to learn how to classify previously unseen data after undergoing a training process. The classification of samples that were not seen in the training phase is named generalization.

Artificial Neural Networks (ANNs) are based on brain behavior and can automatically perform tasks that imitate the brain reasoning and learning from large information sets. In this work, Multilayer Percetron Networks (MLPs) [41], one of the most popular ANN models, are investigated. MLPs present at least one hidden layer, one input layer and one output layer. The hidden layers work as feature extractors: their weights codify features from input patterns, creating a more complex representation of the training data set. There is no rule to specify the number of neurons in the hidden layers. MLPs are usually trained by the back-propagation learning algorithm [23].

SVMs are learning algorithms based on the statistical learning theory, through the principle of Structural Risk Minimization (SRM) [50]. SVMs accomplish a non-linear data analysis in a high dimension space where a maximum margin hyperplane can be built, allowing the separation of positive and negative classes. They present high generalization ability, are robust to high dimensional data and have been successfully applied to the solution of several problems, like pattern recognition and people’s faces detection in images, in which good performance was obtained.

The C4.5 algorithm [40] is a symbolic learning algorithm that generates decision trees from a training data set. It is one of the successors of ID3 algorithm. The C4.5 algorithm uses a greedy approach to progressively grow a decision tree whose leaf nodes represent classes. C4.5 deals with noise data by using a pruning procedure. In this procedure, ramifications of the trained tree that present, according to some criterion, low expressive power, are pruned. This procedure
Aims to simplify the built tree and to reduce its classification error rate.

IV. EXPERIMENTS

All experiments were performed using the Weka simulator [20] and investigated the performance of ML techniques presented in Section III for the task of emotion recognition. One important issue is related to the way the data is partitioned. It has been shown that person-dependent systems, where data from a subject is present in both the training and test sets, outperform person-independent systems [6]. However, the proposed system should be able to recognize emotions of any person, including people who were not encountered previously. To report on person-independent system performance, it is important to partition the data in such a way that there are no data from one subject in both the training and the test sets. In this sense, data was partitioned in 80% of samples used for training and the remaining 20% for tests. Samples from a given subject are not present in both training and test sets.

The ML techniques have different parameters to set while training the classifier. To find the optimal parameters for the problem in terms of classification performance, it is important that the whole parameter optimization process is done independently of the test set. Otherwise, the results will seem overly optimistic [53]. To find optimal parameter values, a separate 3-fold cross validation [34] loop is employed each time a classifier is training when searching for the optimal parameters. The classifier performance is evaluated for different values of each parameter. This process is repeated for each of the three folds, usually resulting in a slightly different value of the parameter for which performance was maximum. The final chosen parameter value is the average over the values found for the three folds. Parameters are assumed independent and there is no specific order in which the various parameters are optimized. In all the reported experiments, optimization for all unknown parameters is performed this way.

The data analyzed is the Radboud Faces Database (RaFD) [27]. It is a freely available face database containing Caucasian and Moroccan face images of both adults and children. All models in the dataset show eight facial expressions with three gaze directions, photographed simultaneously from five different camera angles. The photos were taken in a highly controlled environment and models wore black T-shirts, had no hair on the face and wore no glasses, makeup or jewellery. Expressions are neutral, anger, sadness, fear, disgust, surprise, happiness and contempt. All displayed facial expressions were based on prototypes from the Facial Action Coding System [17]. Each expression was performed with eyes directed straight ahead, averted to the left and averted to the right.

For the performed experiments, only a subset of RaFD database is used. This subset contains all images in which models are in frontal view, with eyes directed straight ahead and expressing all listed emotions except contempt. This resulted in a subset of 67 images, including both adults and children.

A. Results

For the experiments, two new feature sets were also used. They are called FS1₋₃ and FS2₋₄ which are, respectively, a combination of FS1 with FS3, and FS2 with FS4. The objective is to investigate whether a combination of the created feature sets, which map distinct characteristics of a subjects’ expression, can result in better performance for the task of emotion recognition.

FS1, FS2, FS3, FS4, FS1₋₃ and FS2₋₄ were evaluated by ML techniques SVMs, C4.5 and MLPs. The 10-fold cross validation methodology [34] is applied for training the classifiers. To evaluate its generalization capacity, the performance over the test set, which does not have samples present in the training set, was checked.

The best SVM results for each feature set can be seen in Table I. The explored SVM kernels were Linear, Polynomial and Gaussian. Kernel dependent parameters were optimized by a separate 3-fold loop. In general, best results presented high accuracy rates and were obtained with Gaussian kernel.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Kernel</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS1₋₃</td>
<td>Gaussian</td>
<td>80.22 ± 1.05</td>
</tr>
<tr>
<td>FS2₋₄</td>
<td>Gaussian</td>
<td>75.82 ± 1.12</td>
</tr>
<tr>
<td>FS3₋₄</td>
<td>Linear</td>
<td>90.11 ± 1.07</td>
</tr>
<tr>
<td>FS4₋₄</td>
<td>Linear</td>
<td>87.91 ± 1.03</td>
</tr>
<tr>
<td>FS1₋₃</td>
<td>Gaussian</td>
<td>94.51 ± 1.08</td>
</tr>
<tr>
<td>FS2₋₄</td>
<td>Gaussian</td>
<td>90.11 ± 1.09</td>
</tr>
</tbody>
</table>

Tables II and III present, respectively, the best results achieved by C4.5 algorithm and MLP networks. For C4.5 algorithm, optimized parameters were the number of minimum instances per node and the Confidence Factor. Lowering the Confidence Factor decreases the amount of post-pruning. Post-pruning in the C4.5 algorithm is the process of evaluating the decision error (estimated percent of misclassifications) at each decision junction and propagating this error up the tree. At each junction, the algorithm compares the weighted error of each child node versus the misclassification error if the child nodes were deleted and the decision node were assigned the class label of the majority class. The training data misclassifications at each node would not provide a sufficient error estimate - the tree was created from this data so it would not lead to any pruning. Instead, the misclassification error must be understood as an approximation of the actual error based on incomplete data.

For MLP networks, optimized parameters were the Learning Rate, the Momentum and the number of nodes in hidden layers. During experiments, all MLP networks evaluated had only one or two hidden layers.

As for SVMs, the C4.5 results achieved high accuracy rates, however not as good as the previous ones. It is
important to note that for SVMs, the best performance can be seen for feature set $FS_{1-3}$, followed by $FS_{2-4}$ and $FS_3$. For C4.5 algorithm, the best performance was also obtained for the same feature sets, but the best one was for $FS_3$. As for SVMs, $FS_1$ and $FS_2$ presented the worst results.

The performance of MLPs was similar to SVMs and better than the C4.5 ones. For them, $FS_{1-3}$ also produced the best result, and high accuracy rates were obtained for all feature sets.

**TABLE II**

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$FS_1$</td>
<td>71.43 ± 1.04</td>
</tr>
<tr>
<td>$FS_2$</td>
<td>71.43 ± 0.97</td>
</tr>
<tr>
<td>$FS_3$</td>
<td>86.81 ± 1.05</td>
</tr>
<tr>
<td>$FS_4$</td>
<td>79.12 ± 1.11</td>
</tr>
<tr>
<td>$FS_{1-3}$</td>
<td>84.62 ± 1.07</td>
</tr>
<tr>
<td>$FS_{2-4}$</td>
<td>84.62 ± 1.09</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Hidden layers</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$FS_1$</td>
<td>One</td>
<td>81.02 ± 1.29</td>
</tr>
<tr>
<td>$FS_2$</td>
<td>One</td>
<td>80.21 ± 1.17</td>
</tr>
<tr>
<td>$FS_3$</td>
<td>One</td>
<td>85.44 ± 1.19</td>
</tr>
<tr>
<td>$FS_4$</td>
<td>One</td>
<td>84.05 ± 1.25</td>
</tr>
<tr>
<td>$FS_{1-3}$</td>
<td>One</td>
<td>90.03 ± 1.11</td>
</tr>
<tr>
<td>$FS_{2-4}$</td>
<td>One</td>
<td>87.59 ± 1.16</td>
</tr>
</tbody>
</table>

When comparing different algorithms, a statistical test is needed to determine the superiority of a particular technique among others. So, the results presented were used to decide which of the techniques achieved better performance with 95% of certainty. Therefore, the main task is to determine if the difference between the techniques $As$ and $Ap$ is relevant or not, assuming the normal distribution of error rates [51]. For such, the average and the standard deviation of the error rates were calculated according to equations 1 and 2, respectively. The absolute difference of standards deviations was obtained by Equation 3 [35].

$$mean(As - Ap) = mean(As) - mean(Ap)$$ (1)

$$sd(As - Ap) = \sqrt{sd(As)^2 + sd(Ap)^2}$$ (2)

$$t_{calc} = ad(As - Ap) = \frac{mean(As - Ap)}{sd(As - Ap)}$$ (3)

Choosing the initial null hypothesis $H_0 : As = Ap$ and the alternative hypothesis $H_1 : As \neq Ap$. If $ad(As - Ap) > 0$ then $Ap$ is better than $As$; however, if $ad(As - Ap) \geq 2.00$ (boundary of acceptation region) then $Ap$ is better than $As$ with 95% of certainty. On the other hand, if $ad(As - Ap) \leq 0$ then $As$ is better than $Ap$ and if $ad(As - Ap) \leq -2.00$ then $As$ is better than $Ap$ with 95% of certainty. The boundary of acceptation region AR: (-2.00, 2.00) for the experiments are based in the distribution table Student $t$ [33].

According to that, statistical tests were carried out to determine whether any feature set analyzed resulted in better performance for each of the ML techniques investigated. For SVMs and MLPs, with 95% of certainty, $FS_{1-3}$ presented better results than $FS_{2-4}$ and $FS_3$. However, for C4.5 algorithm, $FS_1$ presented better results than $FS_{2-4}$ and $FS_{1-3}$, with 95% of certainty.

The statistical analysis was also used to check whether any of the ML techniques presented better performance for the emotion recognition task. Best found results for the three ML techniques were compared and the analysis demonstrated that, with 95% of certainty, SVMs performed better than C4.5 algorithm and MLP networks.

An analysis of each emotion which was easily identified by ML techniques was also made. Tables IV, V and VI show, respectively, the individual performance for each of the emotions investigated for best ML techniques results, to know, $FS_{1-3}$ for SVMs and MLPs; $FS_3$ for C4.5 algorithm.

**TABLE IV**

Confusion Matrix for SVM $FS_{1-3}$

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Real</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Fear</th>
<th>Neutral</th>
<th>Anger</th>
<th>Surprise</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

**TABLE V**

Confusion Matrix for C4.5 $FS_3$

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Real</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Fear</th>
<th>Neutral</th>
<th>Anger</th>
<th>Surprise</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
</tbody>
</table>

As it can be seen, the emotions Happiness, Neutral and Disgust are the most easily recognizable by the ML techniques investigated. Sadness and Surprise seem to be the most difficult emotions to identify, but this problem may be partially explained because, during the experiments, Face Tracker presented problems in identifying mouth lips downwards, which is an intrinsic characteristic of Sadness, and also the mouth wide open, reducing the mouth amplitude by identifying the tongue as the lower mouth lip, which is an
important characteristic of Surprise. All others emotions had significant high accuracy, except for Fear when analyzed by the C4.5 algorithm, which presented higher error rates, being misclassified mainly as Surprise or Sadness.

According to the results, it is possible to note that the chosen facial characteristics provided to ML techniques the capacity to model the emotion recognition task, since they presented good performance. Also, it seems that information from $FS_1$, which maps the face relying on a greater number of angles and distances, best represents the emotional changes of the face. Another important issue is that combined feature sets, which capture differences of the face according to its neutral state as well as characteristics of the facial expression during each emotion, achieve better ML performance.

V. CONCLUSIONS

The emotion recognition problem was investigated in this paper. For such, geometric features such as the shapes of the facial components (eyes, mouth, nostrils, chin and eyebrows) and the location of facial salient points (corners of the eyes and mouth) were used to propose four distinct feature sets which represent characteristics of the human face.

Three ML techniques (SVMs, C4.5 algorithm and MLPs) were investigated and presented good performance with high hit rates when applied to the RaFD database. Statistical tests also demonstrated that SVMs achieved the best performance among all others ML techniques and that feature sets which combine single frame based characteristics with information of differences of the face according to its neutral state improved even more the experimental results.

All presented feature sets were completely developed and validated for the emotion recognition task. The use of predefined areas of the face in conjunction with angles and distances for constructing feature sets is an important contribution and seemed to be a good alternative. Even though reduced feature set $FS_2$ could not be as efficient as feature set $FS_1$, it achieved very good performance in considerably reduced ML training time. Maybe a more robust vision system could benefit even more from this feature set, but it still needs to be evaluated.

The authors are now acquiring a new emotion database, the Extended Cohn-Kanade Dataset (CK+) [32], to extend the experiments and validate the proposed feature sets with different data related to emotions.

### Table VI

**Confusion Matrix for MLPs $FS_{1-3}$**

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>Happiness</td>
</tr>
<tr>
<td>Happiness</td>
<td>13</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>0</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
</tr>
</tbody>
</table>

### References


