A DIGITAL PRINTING APPLICATION AS AN EXPRESSION IDENTIFICATION SYSTEM.

Fachrurazi, Ashari Imamuddin
Politeknik Negeri Cilacap, Indonesia

ABSTRACT:

Human Computer Interaction (HCI), a growing research field in science and engineering, aims to provide a natural way for humans to use computers as tools. Humans prefer to interact with each other mainly through speech, but also through facial expressions and gestures, for certain parts of the speech and displays of emotions. The identity, age, gender, and emotional state of a person can be obtained from his face. The impression we receive from the expression reflected on the face affects our interpretation of the spoken word and even our attitude towards the speaker himself. Although emotion recognition is an easy task for humans, it still proves to be a difficult task for computers to recognize user’s emotional state. Advances in this area promise to arm our technological environment by means for more effective interactions with humans, and hopefully the impact of facial expressions on cognition will increase rapidly in the future. Will do. In recent years, the adoption of digital has increased rapidly, and the quality has improved significantly. Digital printing has resulted in fast delivery and needs-based costs. This article describes a sophisticated combination classifier approach, an empirical study of ensembles, stacking, and voting. These three approaches were tested on Nave Bayes (NB), Kernel Naive Bayes (kNB), Neural Network (NN), Auto MultiLayer Perceptron (Auto MLP), and Decision Tree (DT), respectively. The main contribution of this paper is the improvement of the classification accuracy of facial expression recognition tasks. In both persondependent and nonpersondependent experiments we showed that using a combination of these classifier combinations gave significantly better results than using individual classifiers. It has been observed from experiments that the overall voting technique by voting achieves the best classification accuracy.

KEYWORDS: Human Computer Interaction, facial emotion recognition, facial expressions, facial action coding system, classifier combination, facial features, AU-Coded facial expression, CK+ database, digital printing.

1. INTRODUCTION

Human beings can express their emotion through voice, body gestures and facial expression. But the most expressive way a human being displays his/her emotional state is through facial expressions. Facial expressions are the facial changes in response to a human being’s internal emotional states, intentions, or social communications. Face is the primary signal system to show the emotion of a person. Face recognition and automatic analysis of facial expressions are one of the most challenging research areas in the field of Human- Computer Interaction (HCI) and have received a special importance. In the 1990s, there has been growing interest to construct automatic methods of recognizing emotions from facial expressions in images or video. Emotion as a private state is not open to any objective observation or verification. So, the recognition of the emotional state of a person is really a challenging issue. Relativity, categorization and identification of emotion are three crucial factors in emotion analysis. The relativity factor of emotion depends on the person’s facial expression or state of mood whereas other two factors are comprehensible but require high technical affluence of computer intelligence. In recent years, computers and automated processing have had a considerable influence on prepress. The integration of prepress and press, as well as automation in printing and the integration of related processes, have also reached a certain maturity. The use of digital printing applications such as advertising, photos, architectural design etc. and integration of these applications into traditional print markets is rapidly expanding.
Emotion is the realm where thought and physiology are inextricably entwined, and where the self is inseparable from individual perceptions of value and judgment toward others and us. In the last few years, automatic emotion recognition through facial expression analysis has been used in developing various real-life applications such as security systems, computer graphics, interactive computer simulations/designs, psychology and computer vision. In psychology, emotion is often defined as a complex state of feeling that results in physical and psychological changes that influence thought and behavior. Our emotions are composed of three critical components: a subjective component (how we experience the emotion), a physiological component (how our bodies react to the emotion), and an expressive component (how we behave in response to the emotion). These different elements can play a role in the function and purpose of our emotional responses.

There are a lot of words for the message persons get from the face (afraid, terrified, horrified, apprehensive, worried, to mention a few of those related to fear), but few to describe the source of those messages. Human beings do have the terms smile, grin, frown, squint, but there are relatively few such words that identify particular facial configurations, distinctive wrinkle patterns, or temporary shapes of the facial features. Without terms to refer to the face human beings are incapable in comparing or correcting their interpretations of facial expression. According to Ekman and Friesen[1], the face gives more than one kind of signal to convey more than one kind of message. Sometimes the people can’t differentiate the emotion messages from the other messages conveyed by the face. The face provides three types of signals: static (such as skin color), slow (such as permanent wrinkles), and rapid (such as raising the eyebrows). The static signals include many more or less permanent aspects of the face like-skin pigmentation, coloration, the shape of the face, bone structure and the size, shape and location of the facial features (brows, eyes, noise, mouth). The slow signals include changes in the facial appearance which occur gradually with time and in addition to the development of permanent wrinkles, changes in muscle tone, skin texture and even skin, coloration occur with age. The rapid signals are produced by the movements of the facial muscles, resulting in temporary changes in facial appearance, shifts in the location and shape of the facial features and temporary wrinkles. These changes have been occurred on the face for a matter of seconds or fractions of a seconds. All three types of facial signals can be modified or disguised by personal choice, although it is hardest to modify the static and slow signals. So one can be misled, intentionally or accidentally, by rapid, slow or static signals. The face is not just a multi-signal system (rapid, slow, static) but also a multi-message system. In order to describe the emotion, it is referred to transitory feelings such as fear, anger, surprise, happiness etc. When these feelings occur, the facial muscles contract and visible changes are appeared on the face. Scientists have found that accurate judgments of emotion can be made from the rapid facial signals. It is important to note that the emotion messages are not transmitted by either the slow or the static facial signals; however these may affect the implications of an emotion message. Sometimes the facial expression analysis has been confused with emotion analysis in the computer vision domain. For emotion analysis, higher level knowledge is required. For example, although facial expressions can convey emotion, they can also express intention, cognitive processes, physical effort, or other intra- or interpersonal meanings. Computer facial expression analysis systems need to analyze the facial actions regardless of context, culture, gender, and so on. In regard to facial expressions of emotion we should have a higher knowledge on rapid facial signals and their distinctive messages.

II. MOTIVATION AND RELATED WORK

Facial expression is one of the most significant ways for human beings to communicate their emotions and intentions. The face can express sooner than people verbalize or even realize their feelings. Since the mid of 1980s a remarkable novelty has been brought to build computer system to understand and use the natural form of human communication. There are lot of praiseworthy researches have been carried out in this Human-Computer Interaction (HCI) field during last decades. In recent years, the developments in HCI have abetted the user to interact with the computer in novel ways beyond the traditional boundaries of the keyboard and mouse. This emerging field includes areas, such as computer science, engineering, psychology and neuroscience. As facial expressions provide important clues about emotions, several
approaches have been envisaged to classify human affective states. Facial expressions are visually observable, conversational, and interactive signals that regulate our interactions with the environment and other human beings in our vicinity [2]. Therefore, behavioural research, bimodal speech processing, videoconferencing, face/visual speech synthesis, affective computing and perceptual man-machine interfaces are those principle driving applications that have lent a special impetus to the research problem of automatic facial expression analysis and produced a great number of interests in this research topic. In 1872, Darwin [3] took attention by firstly demonstrating the universality of facial expressions and their continuity in human beings and animals. He claimed that there are specific inborn emotions, which originated in serviceable associated habits. He also explained that emotional expressions are closely related to survival. In 1971, Ekman and Friesen classified human emotion into six archetypal emotions: surprise, fear, disgust, anger, happiness, and sadness. These prototypic emotional displays are also referred to as basic emotions. Ekman and Friesen also developed the Facial Action Coding System (FACS) for describing facial expressions by action units (AUs). Their work helped to attract the attention of many researchers to analyze facial expressions by means of image and video processing. By tracking facial features and measuring the amount of facial movement, they tried to categorize different facial expressions. Recent works on facial expression analysis and recognition used these universal “basic expressions” or a subset of them. Suwa et al. [4] presented a preliminary investigation on automatic facial expression analysis from an image sequence. Being motivated by some psychological studies,

De Silva et al.[5] carried out experiments on 18 people to recognize emotion using visual and acoustic information separately from an audio-visual database recorded from two subjects. They claimed that sadness and fear emotions are being better identified with audio, and some emotions, e.g., anger and happiness, are better identified with video. Furthermore, Chen et al.[6] cited the use of audio visual information in a multimodal HCI scenario for computers to recognize the user’s emotional expressions. He concluded that the performance of the system increased when both modalities were considered together. The features are taken typically based on local spatial position or displacement of specific points and regions of the face, unlike the approaches based on audio, which use global statistics of the acoustic features. Pantic [7] explored the complete review of recent emotion recognition systems based on facial expressions. He strongly argued to include the essence of emotional intelligence into HCI design in the near future so that the system could be able to recognize a user’s affective states—in order to become more human-like, more effective, and more efficient. Moreover, he provided new techniques for developing the initial phase of an intelligent multimodal HCI—an automatic personalized analyser of a user’s nonverbal affective feedback.

Mase [8] developed an emotion recognition system based on the major directions of specific facial muscles. He was one of the first to introduce image processing techniques to recognize facial expressions. He used optical flow to estimate facial muscle actions which can then be recognized as facial expressions in both a top-down and bottom-up approach. In both approaches, the objective was on computing the motion of facial muscles rather than of facial features. Facial expressions are the result of facial muscle actions which are triggered by the nerve impulses generated by emotions. The muscle actions cause the movement and deformation of facial skin and facial features such as eyes, mouth and nose. As the texture of a fine-grained organ of facial skin has helped in extracting the optical flow, muscle actions have been extracted from external appearance. By the use of K-nearest neighbor for classification, four emotions, namely: happiness, anger, disgust and surprise have been recognized with an accuracy of 80%. Yacoob et al. [9] proposed a similar approach for analyzing and classifying facial expressions from optical flow based on qualitative tracking of principle regions of the face and flow computation at high intensity gradients points. Instead of using facial muscle actions, they constructed a dictionary to convert local directional motions associated with edge of the mouth, eyes and eyebrows into a linguistic, per-frame, mid-level representation. The main goal of this approach was to develop computational methods that relate such motions as cues for action recovery. Face region motion refers to the changes in images of facial features caused by facial actions corresponding to physical feature deformations on the 3-D surface of the face.
Rosenblum et al. [10] also computed optical flow of regions on the face, then applied a radial basis
function network architecture that learned the correlation between facial feature motion patterns and human
emotions. This architecture was specially invented to classify expressions. Besides, Lanitis et al. [11]
invented a flexible shape and appearance model for image coding, person identification, pose recovery and
facial expression recognition. Black et al. [12] developed a mid-level and high-level representation of facial
actions using parametric models to extract the shape and movements of the mouse, eye and eyebrows. This
approach is also recommended in [10] with 89% of accuracy. Tian et al. [13] obtained 96% accuracy using
permanent and transient facial features such as lip, nasolabial furrow and wrinkles. They also used
geometrical models to locate the shapes and appearances of those features. In this study the objective was
to recognize the AU, developed by Ekman and Friesen. Essa et al. [14] introduced a system that quantified
facial movements based on parametric models of independent facial muscle groups and achieved 98% accuracy.
They invented spatial-temporal templates that were used for emotion recognition. In this study, the
face was being modelled by the use of an optical flow method coupled with geometric, physical and
motion-based dynamic models. Matsuno et al. [15] described an approach for recognizing facial
expressions from static images focusing on a pre-computed parameterization of facial expressions. Their
approach plotted a grid over the face and warped it based on the gradient magnitude using a physical
model. In that model, the amount of wrapping was represented by a multi-variate vector which was
different than a learned vector of four facial expressions (i.e., happiness, sadness, anger and surprise).
Sebe et al. recommended a method introducing the Cauchy Naive Bayes classifier which used the Cauchy
distribution as the model distribution for recognizing emotion through facial expressions displayed in video
sequence. Their person-dependent and person-independent experiments showed that the Cauchy
distribution assumption typically produced better results than Gaussian distribution assumption. They used
the simplified model proposed by Tao and Huang which took an explicit 3D wireframe model of the face.
The face model consisted of 16 surface patches embedded in Bezier volumes. The wireframe model and the
12 facial motion measurements were being measured for facial expression recognition. The 12 features
have been used to measure the facial motion in the face model and using these features the 7 basic classes
of facial expression have been defined for classification.

In the last few years, automatic emotion recognition through facial expression analysis has been used
in developing various real life applications such as security systems, computer graphics, interactive
computer simulations/designs, psychology and computer vision. Though many researchers employed
machine learning classifiers (e.g., Cauchy-Gaussian assumption and support vector machine)
individually to recognize emotion state on facial expression analysis, but the proposed method adopt
classifier combination methodology. To the best of our knowledge, classifier combination methods were
not used yet by any researcher in emotion classification task. So, we employed the classifier combination
methodology to enhance the accuracy of emotion classification task.

III. FEATURES SET

Features are basically function or properties of some variable. Feature classification is to classify the
features into classes for our purpose that may be our target class or noisy class. Efficient measurement of
facial expression is necessary to understand the functionality of face-to-face communication. Most of the
studies on automated expression analysis focus on classification of basic expressions (i.e., joy, fear,
sadness, disgust, surprise, and anger) [16]. Suitable features selection is necessary to accomplish this
classification task. We have used action unit, landmark, intensity and their combination as features in our
experiment.

Action Unit

Action Units (AUs) are the fundamental actions of individual muscles or groups of muscles. AUs
represent the muscular activity that produces facial appearance changes defined in Facial Coding System
by Ekman and Friesen. The reason behind in using the term AU is that more than one action have been
separated from what most anatomists described as one muscle.

Landmark and Intensity: Landmark is one of the important features used in this experiment. It has
been used in CK+ database invented by Cohn et al. [17]. The similarity normalized shape, denoted by $s_{n}$, refers to the 68 vertex points for both the x and y coordinates. The 68 vertices produces a raw 136 dimensional feature vector. These points are the vertex locations after removing all the rigid geometric variation (translation, rotation and scale), relative to the base shape. The similarity normalized shape $s_{n}$ can be obtained by synthesizing a shape instance of $s$, that ignores the similarity parameters $p$. FACS provides a description of all possible and visually detectable facial variations in terms of 44 Action Units (AUs). Usually, a trained human FACS coder identifies the occurrence of an action unit and codes its intensity in a given facial image. Although the FACS coding is a precise tool for studying facial expressions, it is labor intensive. Therefore, automating the FACS coding and measuring the intensity of AUs would make it easier and widely accessible as a research tool in behavioral science. Generally intensities of AUs are measured from absent to maximal appearance using a six-point intensity metric (i.e., 0 to 5).

IV. DATASET

The Cohn-Kanade AU-Coded Facial Expression Database is publicly available from Carnegie Mellon University. CK database was invented for the purpose of promoting research into automatically detecting individual facial expressions in 2000. Since then, the CK database has become one of the most widely preferred test-beds for algorithm development and evaluation. During this period, three following limitations have been clearly seen:

- While AU codes are well validated, emotion labels are not, as they refer to what was requested rather than what was actually performed,
- The lack of a common performance metric against which to evaluate new algorithms, and
- Standard protocols for common databases have not emerged.

The cumulative effect of these factors has made benchmarking various systems very difficult or impossible. This is highlighted in the use of the CK database [24], which is among the most widely preferred corpus for developing and evaluating algorithms for facial expression analysis. It contains image sequences of facial expressions from men and women of varying ethnic backgrounds. Each subject (resolution 640 X 480) was instructed to perform a series of 23 facial displays that include single action units and combinations of action units. A total of 504 sequences are available for distribution in this database. After development of Extended Cohn-Kanade database (CK+) the number of sequences and the number of subjects are increased by 22% and 27% respectively. Emotion expressions included happy, surprise, anger, disgust, fear, and sadness.

In the present work, CK+ image database used as corpus. In CK+ database each set consists of sequences of image of a subject (i.e., man/woman). During person dependent experiment, each of the sets was used separately. However in person independent experiment, the image database was divided into two parts: the first part, which was used as training, contains the 70% of the entire dataset and the second part, which was used as testing, contains 30% of the entire dataset.

V. PREPARING TRAINING MODELS

Many researchers investigated the technique of combining the predictions of multiple classifiers to produce a single classifier (Breiman, 1996c; Clemen, 1989; Perrone, 1993; Wolpert, 1992). The resulting classifier is generally more accurate than any of the individual classifiers making up the ensemble. Both theoretical (Hansen and Salamon, 1990; Krogh and Vedelsby, 1995) and empirical (Hashem, 1997; Opitz and Shavlik, 1996a, 1996b) researches have been carried out successfully. Two sets of experimentation were carried out: person dependent and person independent facial emotion recognition. Thus two different sets of training models were prepared and tested. Altogether sixty three classifier combination models, namely ensemble,
stacking and voting were prepared for the identification of person independent facial emotions. During experimentations, the features were added in incremental fashion. First AU feature was used to train model. Then intensity and landmark features were added gradually to train. The training models preparation for person dependent and person independent task is same except the size of the training data. So the following four sub- sections describe the training models preparation for both cases.

Models using Individual Classifiers : Initially the five classifiers- NB, k-NB, NN, auto MLP and DT were trained individually with the gradually incremented feature set, i.e., first they were trained with AU feature only; then they were trained with AU and intensity features; and finally they were trained with AU, intensity and landmark feature set.

Models using Classifiers Combination : This section describes the empirical study of the state-of-the-art classifier combination approaches, namely ensemble, stacking and voting. Each of these three approaches was tested with Naïve Bayes (NB), Kernel Naïve Bayes (k-NB), Neural Network (NN), auto Multi-Layer Perceptron (auto MLP) and Decision Tree (DT). The previous work [11][14][19] on facial emotion identification used these classifiers independently. In this work, the five classifiers were used to establish the effect of combining models. The following sections report the experimentations and obtained results.

Training Ensemble Models : Two popular methods for creating accurate ensembles are- bagging (Breiman, 1996c) and boosting (Freund and Schapire, 1996; Schapire, 1990). These methods rely on resampling techniques to obtain different training sets for each of the classifiers. Bagging approach was applied separately to five base learners, namely NB, k-NB, NN, auto MLP and DT. Initially the size (number of iteration) of the each base learner is set to 2. Then the experiments were performed with gradually increased size (size > 2).

It has been observed that initially the classification accuracy is increased with increasing in size parameter, but after a certain size value, the accuracy is almost stable. Cross validation technique was used to set the suitable size for base learners. Table-5.1 reported the suitable size of the base learners, i.e., NB, k-NB, NN, auto MLP and DT. It can be noticed from Table 5.1 that bagging using NN and auto MLP classifiers require less iterations than that of other classifiers, i.e., NB, kNB and DT.

1. MODEL TESTING AND RESULTS

Evaluation : The actual impacts of the proposals or the predicted results are interpreted by the evaluation. The proposed facial emotion model was evaluated using standard precision, recall and F-measure matrices.

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

\[
\text{Precision} (P) = \frac{TP}{TP + FP}
\]

where, TP = true positive, FP = false positive
Recall in the classification context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

\[
\text{Recall (R)} = \frac{TP}{TP + FN}
\]

where, \( TP = \text{true positive} \), \( FN = \text{false negative} \)

In a classification task, a precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C but it states nothing about the number of items from class C that were not labeled correctly; whereas, a recall of 1.0 means that every item from class C was labeled as belonging to class C but it states nothing about how many other items were incorrectly also labeled as belonging to class C).

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. In this context, F-measure is the harmonic mean of precision and recall that combines precision and recall.

\[
\text{Precision (P)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

**Testing Models:** As discussed in the training section, the training of models is categorized into two sets of experiments, namely person dependent test and person independent test. The following subsections detail the outcome obtained for those experiments.

**Person dependent Testing:** In Person dependent testing, we used the leave-one-out strategy, i.e., one sequence of each test emotion was left out, and the rest were used as the training sequences. Table-5.4 shows the recognition rate for each person for the five classifiers, and the total recognition rate averaged over the five people. Notice that the fifth person has the worst recognition rate. The fact that subject-5 was poorly classified can be attributed to the inaccurate tracking result and lack of sufficient variability in displaying the emotions. It can be noticed that the any of the five classifiers does not significantly decrease the recognition rate (and improves it in some cases), even though the input is unsegmented continuous sequence. Among five classifiers, auto MLP and NN performs almost the same with high recognition rate. The performance of DT is slightly better than that of kNB.

**Conclusion and future work**

In this work, the proposed methodology for recognizing emotions through facial expressions displayed in video sequences using the state-of-the-art classifier combination approaches, namely ensemble, stacking and voting.

The main contributions of this work are-

- introduction of classifier combination methodologies in the emotion recognition on facial expressions task.
- enhancement of classification accuracy of emotion recognition on facial expressions task.
- Classification accuracy is increased for both the person dependent and independent emotion identification.

Overall voting technique with majority voting achieved best classification accuracy.

The emotion recognition from just the facial expressions is probably not accurate enough. Therefore, other measurements probably have to be employed to interact the emotional state of a human with a computer properly. This work is just another step on the way toward achieving the goal of building more effective computers that can serve us better. For future research, we shall focus on facial expressions and body gestures in individual framework as well as multimodal framework because body movements and gestures have recently started attracting the attention of the HCI community. We will have an approach of skin color segmentation on HSV (Hue, saturation and value) space for facial and body feature extraction to recognise emotion.

One of the future directions of this work may be incorporated the color model (e.g., HSV) into
emotion recognition on facial expressions task. The image may be represented to a color model. The color values may be used to compute AU, intensity etc. This may be applied to process facial expression of images in digital printing. Moreover, the integration of multiple modalities such as voice analysis and context would be expected to improve the recognition rates and eventually improve the computer's understanding of human emotional states. This research will be continued to find better methods to fuse audio-visual information that model the dynamics of facial expressions and speech.

REFERENCES