International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:5, 2016

Comparing Emotion Recognition from Voice and Facial Data Using Time Invariant Features

Vesna Kirandziska, Nevena Ackovska, Ana Madevska Bogdanova

***Abstract*—**The problem of emotion recognition is a challenging problem. It is still an open problem from the aspect of both intelligent systems and psychology. In this paper, both voice features and facial features are used for building an emotion recognition system. A Support Vector Machine classifiers are built by using raw data from video recordings. In this paper, the results obtained for the emotion recognition are given, and a discussion about the validity and the expressiveness of different emotions is presented. A comparison between the classifiers build from facial data only, voice data only and from the combination of both data is made here. The need for a better combination of the information from facial expression and voice data is argued.

***Keywords*—**Emotion recognition, facial recognition, signal processing, machine learning.

I. Introduction

EVEN for a human it is difficult to be certain about another human's feelings. People usually guess the current emotion(s) of another human, based on certain characteristics. Building a system that recognizes emotions is a challenging problem. Because, there are many ways to represent emotions, there can be different results from the system for emotion perception.

There are different ways for emotion representation. Some were given in [1]. The categorical emotion representation [2] suggests that there are more than 20 emotion categories and they are all distinct in their name, meaning and expression. Examples include excitement, guilt, and happiness. But, if a system for emotion recognition that has so many emotion categories is to be built, the problem of poor emotion perception may arise. This could occur not only due to the many categories, but because the actual difference in these emotions are not so clear and even more so, because emotions could be ambiguous.

Other emotion representation [2], [3] states that there are two types of emotions: basic and complex emotions. Basic emotions are biologically predefined emotions, while complex emotions are the combination and modification of the basic emotions.

The identification of the basic emotions was the goal of many researchers. Eckman in his research of human facial expression identified six basic emotions [3]. These emotions were: anger, disgust, fear, happiness, sadness and surprise.

More complex categorical emotion representation is given by Plutchik [2]. He represented emotion in so called Plutchic's wheel of emotions, where the graphical representation gives information of how emotions are similar to each other. Two dimensions for emotion activation and evaluation are included in the wheel representation of 32 distinct emotions. In Plutchic's research, other eight emotions are taken as basic.

In this paper, the basic emotions identified by Eckman are used. A system for emotion recognition of these basic emotions is made here. In many other researches, systems for emotion recognition are created. Except in emotion representation, these systems could differ on many other levels: Different data source (sound, visual, brain signals), different features, different algorithms for emotion recognition etc.

In terms of the relationship between the extracted data and the model for recognition, models could be static or dynamic. After they are built, static models would not change (as in [4]) in contrast to dynamic models (as in [5]) which change using some feedback information gathered after recognition. From another side, some algorithms use series of time dependent data from which the recognition is made while others use time invariant data. In this paper, data that are used are time invariant, but the emotion recognition itself is time dependent as it will be explained later.

The main part of the paper is presented in the following three sections. In Section II, data used in this research are presented in more detail. In Section III, the approach for making the model for emotion recognition is given. The analysis of the experimental results for emotion recognition of the model built are given in Section IV. In the end, a conclusion statement for the presented research is given.

II. Emotion Perception Data

The study presented here is part of a research in the area of human-robot interaction [6], [7]. Emotion recognition is vital for enabling more natural interaction between humans and robots. Robots that show or perceive human emotions are called affective robots. These robots find their usage as teachers of children with autism, companions of older people, entertainment robots etc.

For a system that recognizes emotions to be used in human-robot interaction, non-invasive techniques for extracting data are of great importance. For this reason, brain signals are not used in this research. Here both visual and sound signals are used. These can be easily collected in the interaction using a camera. These are the main non-invasive data that can be collected for emotion recognition. Given the data from a camera, many different features can be extracted and there is a

Vesna Kirandziska, Nevena Ackovska and Ana Madevska Bogdanova are with the Faculty of Computer Science and Engineering, Macedonia, The Former Yugoslav Republic of (e-mail: vesna.kirandziska@ finki.ukim.mk, nevena.ackovska@finki.ukim.mk, ana.madevska.bogdanova@finki.ukim.mk).

variety of information that is included in these data.

Many researchers that aim at recognizing emotions use both sound and visual data in their work [4], [5]. There are researches that use only visual features and more precisely features from facial expression [8], [9]. Also, some use only sound features [10], [11]. The specific features that could be used for emotion recognition varies among researchers.

The raw data derived from a camera (or another appropriate sensor) are: digitalized pictures frames (visual signal) and digitalized sound wave (sound signal). Given raw data, valuable features could be extracted and used in the model for emotion recognition. The features that are mostly used are the features from human facial expression.

In [12], a relation between facial features and emotions is shown. In the same research, the well-known Facial Action Coding System (FACS) was presented. This system today is used as a guideline for feature extraction from facial expression by many researches. FACS contains descriptions for 46 action units that give valuable information for the emotion expressed by human. These units are based on human facial mimics on small time intervals. The intervals are short enough to get all important information and long enough to get a complete information set. In [12] for each action unit, the correlations with the emotions are presented. These action units are identified by using a method for extracting facial features. One popular method that is used to extract features of a facial image is the Active Shape Model (ASM) [13]. This method has input parameters that have to be set in order to produce features. Different parameters give different feature values. Some researches use ASM to identify action units defined in FACS. But also, ASM can be used to extract many other data valuables for emotion recognition.

In parallel to explaining the relations between facial features and emotions, researches for identifying relations between human voices with emotions were conducted. One of the most important research with this goal was done by Scherer and it was presented in [14], [15]. Scherer tried to identify valuable sound features that are important in emotion expression. He found that the valuable features can be grouped in three different categories: frequency-based, intensity-based and tempo-based. Yet there are no specific voice features or voice action units found so far. Some researches as in [11] use the fact that there are three categories of sound features valuable for emotion recognition. In that paper, a three-dimensional feature space is used. It is obtained from calculating a single value for each sound category. Other researchers use a union of sets of features that belongs to different categories.

A. Facial Expression Data

In this paper, the facial expression features are calculated based on the important facial points. There are many tools that can be used in order to extract the vital data. In this research, the facial points are extracted using the tool Luxand Face SDK [16]. The number of facial points that are extracted is 65. These points are the one that can be moved with a human facial expression. The others are not affected by human mimics. Each facial point is represented by two coordinates in a two dimensional coordinate system. Beside the extracted points, Luxand Face SDK calculates other values that describe the pose of the face, such as the size of the face, the coordinates of the center of the face and the estimated angle of face rotation in the picture.

These raw data are transformed before being used. Since the facial points for different pictures are represented in different coordinate systems, translation, rotation and scale transformation are done for the complete set of the facial points. As a result, new coordinates for all points are obtained. These new points are coordinates in a system where the center of the face is the origin, and the maximal size (height and width) of the coordinate system is 1. The coordinates are afterwards standardized, that enables the data to contain information about the deviation of the coordinates.

The result from the transformation is the transformed coordinates of the facial points. These are used in the model for emotion recognition. It is important to emphasize that these data are different from the data extracted using FACS. First difference is that these data are extracted from just one picture frame, rather than from several picture frames on given time interval. These features are not given some interpretation like the action units - they are taken just as coordinates. On one hand, the data presented here are human depended since all humans have different facial point coordinates. On the other hand, only mathematical operations for transformation are used to create the data used for the emotion recognition.

B. Voice Data

The voice features that are used in this paper are extracted using PRAAT [17]. Some of the voice features used are based on the sound signal intensity, some on the human pitch and speech tempo. Actually, all three categories of important voice features presented in [14] are included here. It should be remembered that these were proven to give valuable information for human emotions.

The voice features are calculated as global features on small sound segments (all sound segments have equal length). For each sound segment, 33 voice features were extracted. Statistical measures like mean, maximal, minimal values and standard deviations are calculated from the local features for the intensity or pitch of the data. Data from the spectral and formant analysis of the sound are also extracted and the jitter and shimmer measures are derived. In our previous researches [6], [18], [19] investigation about valuable voice features was done and these were used here.

In order to improve the sound features, they are calculated only from the speech containing portion of the sound segment. Thus, a deeper segmentation of the sound pitches or intensity is calculated. The tempo based features that are concerned with both the speech containing portion and the non-containing speech portion, are calculated on the whole sound segment.

In the final stage, before creating the model, the extracted voice and facial point data are preprocessed. The non-complete values that resulted from the software for extracting

these features were excluded from the features database. These non-complete values may be a result of noise in the data. Also, the software for extraction of facial points' features could not recognize a face on the picture that results in missing values for some or all facial points.

III. Emotion Recognition

In the work presented in this paper, the data from the well-known emotional data-base eNTREFACE [20] were used. This database contains around 20 videos from 44 subjects. The videos with duration of around 20 seconds are annotated with one of the six basic emotions: anger, happiness, sadness, surprise, fear and disgust. All subjects were asked to act some emotion and pronounce several sentences using the same emotion.

These data are used to build a model for emotion recognition. Since the emotions are represented with six different classes, a classifier for recognition of the specific emotion should be created. A classifier is built for each subject separately, since the visual information that is extracted is highly dependent on a specific subject. To build a universal classifier, different data that are not dependent on a human face and voice should be extracted.

Each video for each subject was divided into segments that last 1 second. For each segment, voice and facial expression features are extracted as explained in Section II. The facial expression features are extracted from one separate picture taken from the segment. The voice data are calculated from the whole segment.

For each subject, there are around 200 data instances. This set of instances was divided in two separate sets: train data set and test data set. For making the model of emotion recognition the train data set were used. The test data set are used for the model evaluation.

The database in this stage contains 98 features (65 facial expression features and 33 voice features). The number of features is relatively high, compared to the number of instances of features extracted for each subject separately. As a consequence, a technique for lowering the dimensionality of the feature space or feature selection should be used. The standard algorithms for feature selection were used: Best set selection, Forward selection and Backward selection. However, the results for the precision of the models obtained with the features selected by these methods were very poor. The reason is the great correlation among the data extracted for the visual data, more precisely from the facial points. To eliminate this huge linear correlation among the used data in this research (especially visual data), a technique that transforms the given data into the data in a lower dimension space is taken into consideration. Here, experimental results are obtained with the techniques: Principal component analysis (PCA) and Linear discriminant analysis (LDA). The LDA uses the information about the annotation of the instances with an emotion class, while the PCA does not. Clearly the transformations done with LDA showed better results. In this research, LDA was used to lower the feature space. However, LDA was not used on all features together. LDA was used for each separate feature category. All features were divided into smaller categories, not only by their source (voice or facial data). For example, one category from the visual data contains all facial points of the left eye. Another category contains information from the nose, and another contains all information from the facial points that describe a human mouth. Similar categories are made for the voice features. For each data category, LDA is performed separately. In total 14 categories were used: eight for facial features and six for the voice features. Using this approach, the transformed data are expected to contain separate features for the subject's nose, subject's mouth, subject's pitch, subject's tempo, etc. The positive fact is that these characteristics are not calculated by some formula or deeper analyses on the features extracted from the visual and sound data, but rather a generic approach was used to transform the raw data into new data that contains information for each category of facial expression or voice data. The drawback of this approach is the fact that this new data cannot be easily interpreted.

Using the explained procedure and applying the LDA, the new facial expression data contains 62 features (40 facial expression features and 20 voice features). The emotion classifier model that was built used these obtained and transformed data. In the Section IV, the model and the experimental results are presented.

IV. Results

The process for building an emotion classifier starts with the data extraction. It is followed by the data preprocessing and transformation. In the end, the model for emotion classifier is built. R [21] was used to build the classification models.

The emotion perception classifier has the task to classify the facial expression and visual data given to one of the six basic emotions: anger, happiness, disgust, sadness, surprise and fear. For each subject, a separate classifier is created based on the training data. The classifier was validated based on the test data.

Various algorithms for classification were used in the experimental phase, in order to find the algorithm that creates the most precise model for our data. The classification algorithms that were used are: K-nearest neighbors (KNN), Decision trees, Neural Network (NN) and Support vector machines (SVM). The best results were presented with SVM. Only these results are shown here. The results presented on Figs. 1 and 2 represent the precision and recall of the emotion classifiers calculated separately for each basic emotion (an-anger, di- disgust, fe- fear, ha-happiness, sa- sadness, su – surprise).

Separate classifiers are made by using the two different data sources (facial expression and voice data). Also one classifier is made using both facial expression and voice features.

The precision and recall of the classifiers were taken from the test data. Test data include all instances of a selected test set of the videos given in the database. The emotion recognition is first made for each instance of the test data and then, using an appropriate algorithm, the emotion recognition

is done based on each video in the test data set. The algorithm is based on the probability of the classification made for each instance extracted for the given video. For each video in the test data, the most probable emotion is recognized by the classification model. This way of emotion recognition calculated on a larger time interval is based on the assumption that emotions are too complex to be recognizing from data taken at smaller time intervals. In fact, in great part of the picture frames, not much information can be concluded for a person's emotion. In the calculation for the most probable emotion, the ordering of the instances is not taken in account. The assumption is that the order of change of human mimics and human speech features are different when different sentences are spoken. Given this assumption, the time ordering of instances is irrelevant for the emotion recognition of the videos.

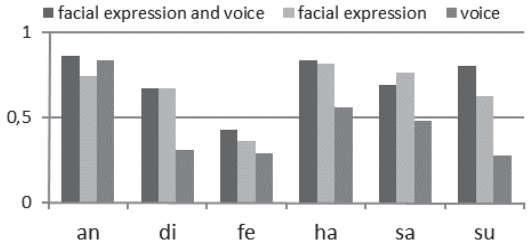

Fig. 1 Emotion precision

The precision and recall of the classification models given at Figs. 1 and 2 are the average precision and recall taken for all test videos for all subject in the database. As expected, the classifier that uses more data, the one based on both facial expression and voice features presents the best results. In this case combining data from different sources contributes to improvement of the classification model. The average precision is 71.4%. Between the classifiers built from just one source, the classification model based on facial expression data shows a higher precision (66.2%) from the model based on voice features (46%). This low precision of the voice feature classifier can be the result of: poor data or non valuable voice features. Sound data usually have much noise, and here no algorithm for noise cancelation was used. Leaving the noise behind, the data itself could have the valuable information since the recorded videos were done by acting emotions. The question whether it is easier to act a facial expression or a voice can be set.

If the collected data are good, then one possibility is that the voice features extracted in this research do not give much information about emotions. Even though [14] found that sound signals are descriptive for emotions, there is no research so far that gives a list of valuable features.

Analyzing the precision and recall separately for the facial expression classifier and for the voice feature classifier give us more information about the properties of the classifiers and a deeper understanding of the classifiers crated using different data sources.

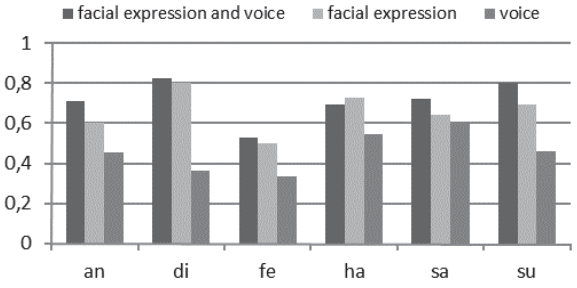

Fig. 2 Emotion recall

Form Fig. 1, one can notice that for some emotions there is a significant representation in the voice, and other emotions are better represented by facial gestures. The classifier that uses facial expression features has a highest precision for happiness (81.3%), sadness (76.1%) and anger (74.4%). One interpretation is that facial expression data show valuable information for all these emotions. On the other side, the classifier that uses voice features classifies best the anger emotion. This suggests that in the extracted voice features there is valuable information for recognizing anger. The problem for the anger emotion is that many videos (more than 50%) are falsely classified as anger (Fig. 3) so the features are not exclusively connected to the anger emotion. All other emotions except anger are classified with a very poor precision of less than 50%. One reason for the poor classification is that the proper features are not extracted. Another reason is that there are no voice features that are valuable for all other emotions classification. This should be further investigated.

The emotion that has the least precision and recall in all classifiers is fear. Using the data in this research, this emotion is classified with a precision around 40% when using both sources and even less when just one data source is used. The classifier from facial expression features has an unsatisfactory precision for disgust and surprise, also.

From the aspect of the recall of the emotions (Fig. 3), the disgust emotion has the highest value from the classifier built from the facial features data. This suggests that there are some facial expression features that are only correlated with the disgust's emotion. Similarly, the next most distinguishable emotion from facial expression features is surprise. In contrast, the sadness emotion has the highest recall from the classifier build based on voice features.

Figs. 2 and 3 present that the precision and recall of the classifiers based on two different sources have different values for different emotions.

Some emotions are better classified from the voice data, others from the visual facial expression data. This may suggest that combining both data sources: facial expression features and voice features, a better classifier could be done. The classifier done here that combines these features gives same improvement for the emotion recognition, but even better improvement was expected. An alternative approach would be to account for the importance of each feature (or feature category) in the calculations for each separate emotion.

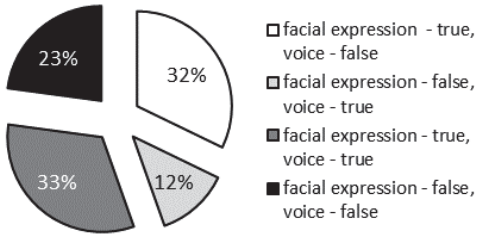

Fig. 3 Categorization of test instances based on their classification from the facial expression and voice classifier

Fig. 3 presents the portions of instances that were true or false classified from the model based on the two separate data sources: facial expression and voice data. On Fig. 3, it is shown that in only 12% of the instances both, the facial features and voice features classifier failed to recognize the correct emotions. A greater part (33%) of the instances were classified correctly by both classifiers, while 32% were classified correctly only by the facial feature classifier. The goal for creating a new classifier that combines the different sources is to lower the percentage of instances that are correctly classified from one data source, but incorrectly from the other. This research will be done in the future work. Here, it is shown that there is a great possibility to build a better classifier that combines facial expression and voice data sources.

V. Conclusion

In the paper presented here, the problem of emotion recognition was investigated. Both facial expression and voice features were used in order to detect a human emotion. The data that were used here differs from all other approaches in other researches for emotion perception since only some mathematical operations and LDA was used to transform data. No additional information for the model of emotion perception was used. The features were only categorized by type, and then using LDA they were projected to a lower dimension space.

In the approach described above, a classifier that includes both facial expression and visual features for emotion perception had a precision of over 71.4%. In this paper, also a comparison of the classification models built only from facial expression data and only from voice data was given by comparing the precision and recall for each emotion.

The results suggested that another combination of the facial expression and voice features could give a much higher precision than the precision gained in the classifier that uses both data sources built here (71.4%). The combination should consider the importance of each feature (feature group) for the recognition of different basic emotions.

With this research the results from combining different data sources for emotion recognition are analyzed and the next steps for further research are given.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius" University.

References

[1] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion representation, analysis and synthesis in continuous space: A survey”, Automatic Face & Gesture Recognition and Workshops, Santa Barbara, 2011, pp. 872-834.

[2] R. Plutchik, “The nature of Emotions”. American Scientist, vo. 89, July-August, 2001, pp. 344-350.

[3] P. Ekman, W.V. Friesen and P. Ellsworth, “Emotion in the human face: Guidelines for research and an integration of findings”, New York: Pergamon Press, 1972.

[4] A. Metallinou, S. Lee and N. Sarayanan, “Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice”, Tenth IEEE International Symposium on Multimedia, Berkeley, CA, 2008, pp. 250-257.

[5] D. Chen, D. Jiang, I. Ravyse and H. Sahli, “Audio-Visual Emotion Recognition Based on a DBN Model with Constrained Asynchrony”, Fifth International Conference on Image and Graphics, 2009, pp. 912-916.

[6] V. Kirandziska and N. Ackovska, “Human-robot interaction based on human emotions extracted from speech”, In Proc. Of the TELFOR, Belgrade, Serbia, 2012, pp. 1381-1384.

[7] V. Kirandziska and N. Ackovska, “Effects and usage of emotion aware robots that perceive human voice”, IADIS Multi Conference Computer Science and Information Systems, Prague, Czech Republic, 2013.

[8] L. Malatesta, J. Murray, A. Raouzaiou, A. Hiolle, L. Cañamero and K. Karpouzis, “Emotion Modeling and Facial Affect Recognition in Human-Computer and Human-Robot Interaction”, Image, Video and Multimedia Systems Lab, National Technical University of Athens, and Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, 2009.

[9] Y. Miyakoshi and S. Kato, “Facial emotion detection considering partial occlusion of face using Bayesian network”, IEEE Symposium on Computer and Informatics, 2011, pp. 96-101.

[10] T. Vogt E. Andr´e and J. Wagner, “Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization”, Affect and Emotion in HCI, Springer-Verlag Berlin Heidelberg. LNCS 4868, 2008, pp. 75–91.

[11] O. Kwon, K. Chan, J. Hao and T. Lee, “Emotion Recognition by Speech Signals”. Proc. of Eurospeech, Geneva, September, 2003, pp. 125-128.

[12] P. Ekman, W.V. Friesen and J.C. Hager, “Facial Action Coding System Investigator’s Guide”, 2002.

[13] Ko K., Sim K.: Development of the Facial Feature Extraction and Emotion Recognition Method based on ASM and Bayesian Network. FUZZ-IEEE, Korea (2009)

[14] K. R. Scherer, “Vocal affect expression: A review and a model for future research”, Psychological Bulletin, vol. 99, 1986, pp.143-165.

[15] K.R. Scherer, R. Klaus, R. Banse, H.G. Wallbott and T. Goldbeck, “Vocal Cues in Emotion Encoding and Decoding. Motivation and Emotion”, 1991, pp. 123-148.

[16] Luxand Inc. *Luxand SDK*. Online. https://www.luxand.com/facesdk/. (accessed 2015).

[17] P. Boersma and Weenink, “PRAAT: doing photetics by computer” (Version 5.1.05). 2009. http://www.praat.org/ (accessed 2015).

[18] V. Kirandziska and N. Ackovska, “Sound features used in emotion classification”, The 9th International Conference for Informatics and Information Technology, Bitola, Macedonia, 2012, pp. 91-95.

[19] V. Kirandziska and N. Ackovska, “Finding Important Sound Features for Emotion Evaluation Classification”, IEEE Region 8 Conference EuroCon, Zagreb, Croatia, 2013.

[20] M. Kotsia, et al. “The enterface’05 audio-visual emotion database”, 2006.

[21] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2008. URL http://www.R-project.org (accessed 2015).