Feature and Decision Level Audio-visual Data Fusion in Emotion Recognition Problem

Maxim Sidorov [1], Evgenii Sopov[2], Ilia Ivanov[2] and Wolfgang Minker[1]
[1]*Institute of Communication Engineering, Ulm University, Ulm, Germany*
[2]*Department of Systems Analysis and Operations Research, Siberian State Aerospace University, Krasnoyarsk, Russia*

Keywords: Emotion Recognition, Speech, Vision, PCA, Neural Network, Human-Computer Interaction (HCI), Feature Level Fusion, Decision Level Fusion.

Abstract: The speech-based emotion recognition problem has already been investigated by many authors, and reasonable results have been achieved. This article focuses on applying audio-visual data fusion approach to emotion recognition. Two state-of-the-art classification algorithms were applied to one audio and three visual feature datasets. Feature level data fusion was applied to build a multimodal emotion classification system, which helped increase emotion classification accuracy by 4% compared to the best accuracy achieved by unimodal systems. The class precisions achieved by applying algorithms on unimodal and multimodal datasets helped to reveal that different data-classifier combinations are good at recognizing certain emotions. These data-classifier combinations were fused on the decision level using several approaches, which still helped increase the accuracy by 3% compared to the best accuracy achieved by feature level fusion.

1 INTRODUCTION

Image classification is a complex machine learning task that has been investigated by many authors. There are numerous problems being solved in this field, and as many practical applications. One such application is the human emotion recognition problem.

This work focuses on researching the idea of audio-visual data fusion in an attempt to improve the emotion recognition rate. First, the audio features are extracted from audio streams, and video features are extracted from video frame sequences. Once extracted, the audio and video feature datasets undergo a dimensionality reduction procedure by applying PCA. The reduced datasets are used to build unimodal emotion recognition systems by applying two state-of-the-art classification algorithms: a support vector classifier trained by a sequential minimal optimization algorithm, and feed-forward neural network.

The key idea of this work is to evaluate and compare the classification rates of audio and visual unimodal systems, and use the most effective of them to perform feature level data fusion in order to check if this helps to increase the emotion recognition rate. Also, more detailed research is performed on emotion class accuracies so as to determine which data-algorithm combination is doing better at predicting each emotion class.

Another aspect of this work is the decision level fusion of the data-algorithm combinations that were the best at predicting each emotion class. Such fusion of the systems that do well on recognizing distinct emotion classes supposedly will result in a higher overall emotion classification rate.

Several video feature extraction algorithms are used: Quantized Local Zernike Moments (QLZM), Local Binary Patterns (LBP) and Local Binary Patterns on Three Orthogonal Planes (LBP-TOP). QLZM and LBP-TOP are state-of-the-art image and video feature extraction algorithms that are designed to grasp all the main features necessary for solving such problems as human face identification, gender recognition, age recognition, and emotion recognition. LBP is a classical algorithm which is used by many researchers in the field of image recognition and pattern classification. This algorithm is rather simple, and provides a good baseline accuracy estimate in image recognition tasks.

The rest of the paper is organized as follows: Significant related work is presented in Section 2, Section 3 provides a description of methodology,

Sidorov M., Sopov E., Ivanov I. and Minker W..
Feature and Decision Level Audio-visual Data Fusion in Emotion Recognition Problem.
DOI: 10.5220/0005527002460251
In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2015), pages 246-251
ISBN: 978-989-758-123-6

which includes feature extraction, dimensionality reduction and classification steps. In Section 4 is presented a description of the audio-visual database that is used in this work, while Section 5 deals with experiments setup and the results achieved. Concluding remarks and plans for future work can be found in Section 6.

2 SIGNIFICANT RELATED WORK

The paper by Rashid et al. (Rashid et al., 2012) explores the problem of human emotion recognition and proposes the solution of combining audio and visual features. First, the audio stream is separated from the video stream. Feature detection and 3D patch extraction are applied to video streams and the dimensionality of video features is reduced by applying PCA. From audio streams prosodic and mel-frequency cepstrum coefficients (MFCC) are extracted. After feature extraction the authors construct separate codebooks for audio and video modalities by applying the K-means algorithm in Euclidean space. Finally, multiclass support vector machine (SVM) classifiers are applied to audio and video data, and decision-level data fusion is performed by applying Bayes sum rule. By building the classifier on audio features the authors received an average accuracy of 67.39%, using video features gave an accuracy of 74.15%, while combining audio and visual features on the decision level improved the accuracy to 80.27%.

Kahou et al. (Kahou et al., 2013) described the approach they used for submission to the 2013 Emotion Recognition in the Wild Challenge. The approach combined multiple deep neural networks including deep convolutional neural networks (CNNs) for analyzing facial expressions in video frames, deep belief net (DBN) to capture audio information, deep autoencoder to model the spatio-temporal information produced by the human actions, and shallow network architecture focused on the extracted features of the mouth of the primary human subject in the scene. The authors used the Toronto Face Dataset, containing 4,178 images labelled with basic emotions and with only fully frontal facing poses, and a dataset harvested from Google image search which consisted of 35,887 images with seven expression classes. All images were turned to grayscale of size 48x48. Several decision-level data integration techniques were used: averaged predictions, SVM and multi-layer perceptron (MLP) aggregation techniques, and random search for weighting models. The best accuracy they achieved on the competition testing set was 41.03%.

In the work by Cruz et al. (Cruz et al., 2012) the concept of modelling the change in features is used, rather than their simple combination. First, the faces are extracted from the original images, and Local Phase Quantization (LPQ) histograms are extracted in each n x n local region. The histograms are concatenated to form a feature vector. The derivative of features is computed by two methods: convolution with the difference of Gaussians (DoG) filter and the difference of feature histograms. A linear SVM is trained to output posterior probabilities. and the changes are modelled with a hidden Markov model. The proposed method was tested on the Audio/Visual Emotion Challenge 2011 dataset, which consists of 63 videos of 13 different individuals, where frontal face videos are taken during an interview where the subject is engaged in a conversation. The authors claim that they increased the classification rate on the data by 13%.

In (Soleymani et al., 2012) the authors exploit the idea of using electroencephalogram, pupillary response and gaze distance to classify the arousal of a subject as either calm, medium aroused, or activated and valence as either unpleasant, neutral, or pleasant. The data consists of 20 video clips with emotional content from movies. The valence classification accuracy achieved is 68.5 %, and the arousal classification accuracy is76.4 %.

Busso et al. (Busso et al., 2004) researched the idea of acoustic and facial expression information fusion. They used a database recorded from an actress reading 258 sentences expressing emotions. Separate classifiers based on acoustic data and facial expressions were built, with classification accuracies of 70.9% and 85%respectively. Facial expression features include 5 areas: forehead, eyebrow, low eye, right and left cheeks. The authors covered two data fusion approaches: decision level and feature level integration. On the feature level, audio and facial expression features were combined to build one classifier, giving 90% accuracy. On the decision level, several criteria were used to combine posterior probabilities of the unimodal systems: maximum – the emotion with the greatest posterior probability in both modalities is selected; average – the posterior probability of each modality is equally weighted and the maximum is selected; product - posterior probabilities are multiplied and the maximum is selected; weight - different weights are applied to the different unimodal systems. The accuracies of decision-level integration bimodal classifiers range

from 84% to 89%, product combining being the most efficient.

3 METHODOLOGY

3.1 Feature Extraction and Dimensionality Reduction

The first step is to extract audio and visual features from raw audio-visual data. Audio features are extracted using openSMILE – open source software for audio and visual feature extraction (Eyben et al., 2010).Video features are extracted using 3 different algorithms:

- Quantized Local Zernike Moments (QLZM) (Sariyanidi et al., 2013);
- Local Binary Patterns (LBP);
- Local Binary Patterns on Three Orthogonal Planes (LBP-TOP).

The QLZM and LBP algorithms extract features from every single video frame in a video sequence, whereas LBP-TOP deals with spatio-temporal space which includes several consecutive frames. The number of such frames is the LBP-TOP parameter and can be changed. Images with their original resolution were used as an input for QLZM algorithm, whereas for LBP and LBP-TOP video frame images were resized from 1280:1024 resolution to a width of 200 pixels, saving the width-height proportion. For the LBP algorithm the following parameters were used: uniform mapping type with 8 sampling points, radius - 1. LBP-TOP parameters: radii along X, Y, and T axis - (1; 1; 1), number of sampling points in XY, XT and YT planes - (8; 8; 8).

The samples were constructed by means of averaging over the whole audio/video sequence, thus 1 audio/video recording has 1 corresponding feature vector.

All data was normalized, and PCA was applied to the datasets for dimensionality reduction. The number of principal components was truncated by using the Kaiser's rule: the principal components whose λ values were less than or equal to 1 were removed from the model, where λ refers to dataset covariance matrix eigen values.

A description of the extracted audio and video samples is presented in table 1.

3.2 Classification

The resulting unimodal sets of audio and visual features were used as an input for 2 classification algorithms: a support vector classifier trained by a sequential minimal optimization algorithm (W-SMO) (Platt, 1998), and feed-forward Neural Network. Rapidminer and R language algorithm implementations were used. The classification algorithms were applied on audio and every video dataset to determine the emotion classification accuracies of unimodal systems. Also, the emotion class accuracies were determined for each data-algorithm combination. The most promising datasets regarding classification accuracy were combined into a single dataset, i.e. feature level data fusion was performed. Also the audio-visual dataset was constructed by combining all available datasets.

Some data-algorithm combinations showed better classification accuracies on distinct emotion classes. These data-algorithm combinations outputs were fused on the decision level by applying several techniques:

- Voting - the class that gained the most votes among base learners is chosen;
- Average class probabilities - the class probability outputs of the base learners are averaged;
- Maximum class probabilities - the maximum class probability is chosen among all base learners;

Table 1: Audio, video and combined datasets description, # of attributes (PCA) - number of attributes after dimensionality reduction, QLZM - Quantised Local Zernike Moments, LBP - Local Binary Patterns, LBP-TOP - Local Binary Patterns on Three Orthogonal Planes; one feature vector per one audio/video file.

Data	# of attributes	# of attributes (PCA)	# of cases
Audio	984	131	480
QLZM	656	36	480
LBP	59	4	480
LBP-TOP	177	10	480
Audio + LBP-TOP	-	140	480
Audio-visual (Audio + QLZM + LBP + LBP-TOP)	-	180	480

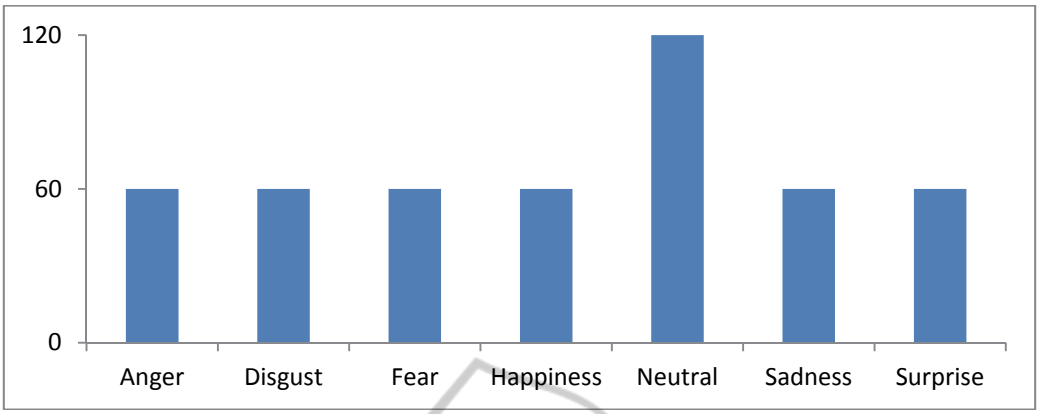

Figure 1: Database emotion classes distribution.

- SVM meta-classifier - an extra SVM meta-learner is trained, the class probability outputs of the base learners are used as input variables, the true class label serves as output variable. The base learners training and meta-learner training is performed on disjoint parts of the training set.

4 DATABASE DESCRIPTION

The SAVEE database (Haq et al., 2009) was used as a data source in this work. The database includes audio-visual information of 4 male speakers reading a pre-defined set of phrases with seven main emotions: anger, disgust, fear, happiness, neutral, sadness and surprise. The overall number of audio-visual recordings is 480. The database includes audio data, audio-visual clips, and sets of split video frames. Fig. 1 shows the database emotion class distribution. As can be observed, there are the same number of video files with all emotions except for neutral state. The small dataset size and the fact that there are only male speakers in the dataset is a drawback, which means that in future work the results presented in this publication should be tried out on a larger dataset, which may improve results.

5 EXPERIMENTS SETUP AND RESULTS

Two classification algorithms were applied to solve the emotion classification problem: W-SMO and Neural Network. The independent speaker classification scheme was used, which means that the data on three speakers was used as a training dataset, and the data on the remaining speaker was used as a test dataset. The classification accuracies were averaged over four different dataset train/test splits (4-fold cross-validation). Neural network parameters: # of hidden layers – 2, # of neurons per hidden layer – (# of attributes + # of classes)/2 + 1, training cycles – 200, learning rate – 0.3. W-SMO parameters: the complexity constant C = 1, normalization is on, tolerance parameter L = 0.001, fit logistic models to SVM outputs is off, the polynomial kernel is used.

Experimental results can be found in table 2. As can be observed from Table 2, audio and LBP-TOP features yield better classification accuracy on both algorithms. This can be explained by the fact that audio and LBP-TOP features grasp more information about human emotions than QLZM and LBP. Combining audio and LBP-TOP features helped improve classification accuracy up to 40.78%.

Table 4 shows class precision for different data-algorithm combinations. As can be observed, all emotions except for sadness can be recognized quite well by using unimodal and multimodal datasets. Anger is best recognized by applying the W-SMO algorithm on the LBP-TOP dataset, happiness is best recognized by applying a neural network on the combined audio-visual dataset etc.

Table 3 shows the classification accuracies achieved by applying decision level fusion techniques on the data-algorithm combinations that achieved best accuracies on distinct emotion classes. Averaging class probability outputs of the base learners improved the classification rate up to 43.48%, which is almost 3% higher compared to the best accuracy achieved by applying feature level fusion.

At the bottom of table 4 the class precision values for decision level fusion systems are presented. It turned out that the best class accuracies are achieved by decision level fusion systems.

Table 2: Emotion classification accuracy (%), speaker independent classification, unimodal systems and feature level fusion.

Data	W-SMO	Neural Network
Audio	36.73	35.42
QLZM	14.58	14.44
LBP	20.35	14.24
LBP-TOP	29.03	26.39
Audio + LBP-TOP	40.78	35.18
Audio + QLZM + LBP + LBP-TOP	27.24	27.35

Table 3: Emotion classification accuracy (%), decision-level fusion over the most effective data-algorithm combinations according to class precisions.

Decision level fusion technique	Accuracy
Voting	41.88
Average class probabilities	43.48
Maximum class probabilities	33.75
SVM meta-classifier	43.12

Table 4: Class precision (%), feature level and decision level fusion of the most effective data-algorithm combinations.

Algorithm	Data	Emotion						
		Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
W-SMO	Audio	41.90	38.31	43.21	35.61	39.95	27.71	30.56
	QLZM	13.31	8.48	4.47	7.50	27.49	2.08	0.00
	LBP	16.07	5.23	0.00	23.44	0.00	0.00	1.58
	LBP-TOP	**63.24**	8.93	12.36	43.85	50.93	4.35	25.00
	Audio + LBP-TOP	62.22	**41.17**	51.82	58.64	69.95	**28.24**	**64.26**
	Audio + QLZM + LBP + LBP-TOP	46.50	12.77	16.67	51.72	**83.33**	6.67	55.56
Neural Network	Audio	57.81	35.62	45.64	42.44	38.99	11.74	19.38
	QLZM	3.40	32.84	20.35	12.50	60.30	1.67	11.36
	LBP	5.58	2.34	0.00	21.70	8.33	0.00	0.00
	LBP-TOP	43.75	7.63	20.77	25.14	19.24	19.71	23.30
	Audio + LBP-TOP	51.34	39.54	**53.95**	58.66	66.40	21.16	56.74
	Audio + QLZM + LBP + LBP-TOP	39.50	21.31	32.86	**61.31**	51.89	3.70	52.78
Voting		**68.65**	38.63	48.66	**65.00**	71.82	19.91	76.39
Average class probabilities over all algorithm/data combinations		50.85	35.52	**71.76**	46.78	66.87	33.93	68.75
Maximum class probabilities over all algorithm/data combinations		57.97	38.92	50.62	55.95	**93.69**	35.96	0.00
SVM meta-classifier		63.89	**41.91**	62.44	58.16	86.74	**38.89**	**90.00**

6 CONCLUSIONS AND FUTURE WORK

While the automatic emotion recognition problem still remains a rather hard task, there are some ways to solve it more effectively. First, the right corpora should be chosen. In this article we tried several audio-visual feature extraction algorithms to test if combining audio-visual features on the feature level would yield better performance. It turned out that for both classification algorithms used in this work that statement holds true. Also, the class precision results made it possible to define data-algorithm combinations that do best at recognizing certain distinct emotions. Fusing those data-algorithm combinations on the decision level improved the classification rate by 3%.

In future the authors plan to perform a broader research of the ideas included in this article, by applying a wider variety of classification algorithms, on a more extensive dataset. The idea of feature selection will be researched as a substitute for using the PCA for dimensionality reduction of audio-visual datasets.

ACKNOWLEDGEMENTS

The authors express their gratitude to Mr. Ashley Whitfield for his efforts to improve the text of this article.

REFERENCES

Haq, S., Jackson, P. J. B. (2009) Speaker-dependent audio-visual emotion recognition. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'09), Norwich, UK, pp.53-58, September 2009*.

Rashid, M., Abu-Bakar, S. A. R., Mokji, M. (2012). Human emotion recognition from videos using spatio-temporal and audio features. *Vis Comput (2013),* 29: 1269-1275.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gulcehre, C., Memisevic, R., Vincent, P., Courville, A., Bengio, Y. (2013). Combining modality specific deep neural networks for emotion recognition in video. *ICMI'13, December 9-13, 2013, Sydney, Australia.*

Cruz, A., Bhanu, B., Thakoor, N. (2012). Facial emotion recognition in continuous video. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), November 11-15, 2012, Tsukuba, Japan.*

Soleymani, M., Pantic, M., Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on affective computing, vol. 3, no. 2, April-June, 2012.*

Eyben, F., Wullmer, M., Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10.2010.*

Sariyanidi, E., Gunes, H., Gokmen, M., Cavallaro, A. (2013). Local Zernike Moment Representation for Facial Affect Recognition. *BMVC'13*.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S. (2004). Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. *University of Southern California, Los Angeles, http://sail.usc.edu.*

Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *TechReport MSR-TR-98-14, Microsoft Research.*