Automatic bi-modal emotion recognition system based on fusion of facial expressions and emotion extraction from speech

Dragos Datcu
D.Datcu@tudelft.nl

Leon J.M. Rothkrantz
L.J.M.Rothkrantz@tudelft.nl

Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology, The Netherlands

Abstract

Our software demo package consists of an implementation for an automatic human emotion recognition system. The system is bi-modal and is based on fusing of data regarding facial expressions and emotion that has been extracted from speech signal. We have integrated Viola&Jones face detector (OpenCV), Active Appearance Model – AAM (AAM-API) for extracting the face shape and Support Vector Machines (LibSVM) for the classification of emotion patterns. We have used Optical Flow algorithm for computing the features needed for the classification of facial expressions. Beside the integration of all processing components, the software system accommodates our implementation for the data fusion algorithm. Our C++ implementation has a working frame-rate of about 5fps.

1. Introduction

In the current paper we present the details for the implementation of a bimodal emotion recognition algorithm that incorporates facial expression recognition and emotion extraction from speech. The multimodal emotion data fusion model works at the high, semantic level and that takes account of the dynamics in facial expressions and speech.
In our approach, the emotion recognition algorithm works for the prototypic emotions. We have based our single modality data processing methods on previous works [3][4] we have conducted for the recognition of emotions from human faces and speech.

The recognition is done by using Viola&Jones features and boosting techniques for face detection [4], Active Appearance Model – AAM for the extraction of face shape and Support Vector Machines –SVM [5] for the classification of feature patterns in one of the prototypic facial expressions. For training and testing the systems we have used Cohn-Kanade database [5].

The SVM classification works on inputs generated by determining the Optical Flow on the consecutive images of the user's face area (Figure 1).

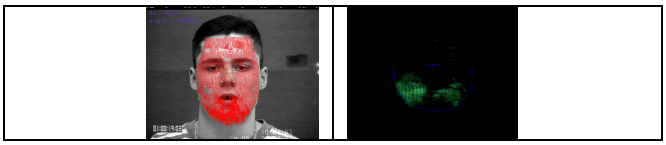

Figure 1. Results of applying optical flow for one frame (left) and whole frame sequence from Cohn-Kanade data set.

In the case of emotion recognition from speech, the analysis is handled separately given the type of segmentation [4].
The parameters used for classification consist of the mean, standard deviation, minimum and maximum of the following acoustic features: *Fundamental frequency* (pitch), *Intensity*, *F1*, *F2*, *F3*, *F4* and *Bandwidth*.

The fusion model aims at determining the most probable emotion of the subject given the emotions determined in the previous frames. A data window contains the current and the previous *n* frames for the analysis. **Figure 2** depicts the Dynamic Bayesian Network - DBN for the emotion data fusion.

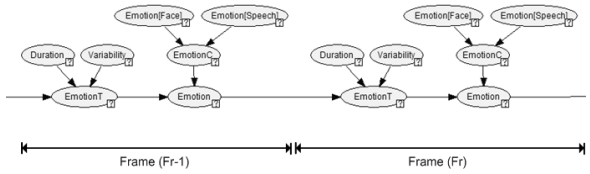

Figure 2: The DBN model for multimodal emotion recognition.

The variable *Duration* represents the stability of last determined emotion in consecutive frames in the analysis window. It has three states: *short*, *normal* and *long*. Each of the three possible situations are assumed to hold differently for each facial expression. Accordingly, it is assumed that for instance an emotion transition is likely to happen from one emotion to another after the former

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

emotion has been shown during a number of consecutive frames.

The variable *Variability* represents the knowledge on the variability of previously determined emotions in the analysis window. The variable *EmotionT* represents the most probable emotion taking into account only the emotion of the subject determined at the previous frame in the analysis window and the probability of showing another emotion. The variable *EmotionC* is the emotion of the subject as it is computed by the facial expression recognition and the emotion extraction from speech at the current frame. The variable *Emotion* is the emotion of the subject at the current frame to be determined.

2. Implementation

Figure 3 shows a snap shot of our software implementation [2] for the bimodal human emotion recognition system. Our system runs on Windows machines. For the detection of faces we have mainly used the implementation of Viola&Jones method from Intel's Open Source Computer Vision Library – OpenCV. We have used AAM-API [6] libraries for the implementation of Active Appearance Models. For the speech processing part we have built Tcl/Tk scripts in combination with Snack Sound Toolkit, a public domain toolkit developed at KTH.

Finally, we have built our facial feature extraction routines, the facial expression recognition system and the emotion extraction from speech in C++ programming language. For the classification component, we have used LIBSVM [1]. On an Intel Core 2 CPU @2.00 GHz, 2.00 GB of RAM our software implementation works at speed of about 5 fps.

The AAM module of our initial facial expression recognition system requires the detection of faces for each frame in the incoming video sequence. We have obtained a considerable improvement in terms of speed by nearly doubling the frame rate with an algorithm that uses information regarding the face shape of one subject at the current frame as initial location for the AAM fitting procedure at the next frame in the video sequence. In this way the face detection is run only at certain time intervals comparing to the case when it is run for all the frames.

The disadvantage of this algorithm is the higher probability of generating faulty results for the extraction of face shape.

Moreover, the faulty such cases attract definitive erroneous results for the rest of the frames in the sequence. This happens because of the possibly sharp moves of the head, rather low speed of the implementation and because of the incapacity of the AAM algorithm to match model face shapes to image face shapes when the translation, rotation or scaling effects present high magnitudes. The case can be overcome by using an effective face tracking algorithm to anticipate the move of the head in the sequence of frames.

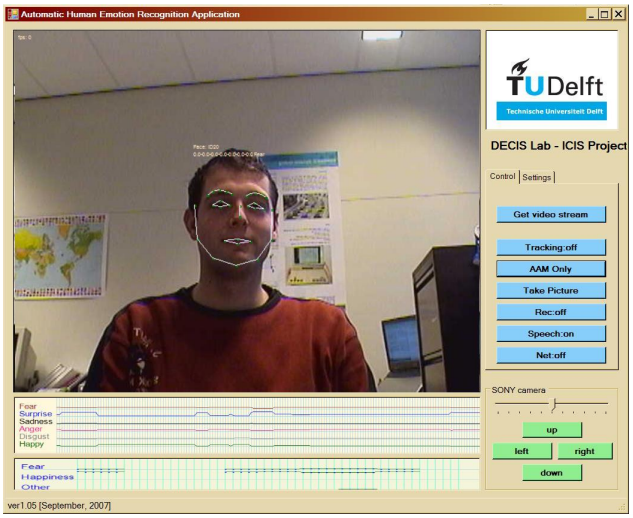

Figure 3. A snap shot of our software implementation of the bimodal emotion recognition algorithm.

References

[1] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[2] D. Datcu, and L. J. M. Rothkrantz, "Multimodal workbench for human emotion recognition", Software Demo at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07, Minneapolis, Minnesota, USA, 2007.

[3] D. Datcu, and L. J. M. Rothkrantz, "The use of Active Appearance Model for facial expression recognition in crisis environments", Proceedings ISCRAM2007, ISBN 9789054874171, pp. 515-524, 2007.

[4] D. Datcu, and L. J. M. Rothkrantz, "The recognition of emotions from speech using GentleBoost Classifier", CompSysTech'06, June 2006.

[5] T. Kanade, J. F. Cohn, Y. Tian, "Comprehensive database for facial expression analysis,". Proc. of the 4th IEEE Int. Con. On Automatic Face and Gestures Reco., France, 2000.

[6] M. B. Stegmann, "The AAM-API: An Open Source Active Appearance Model Implementation", Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, 6th Int. Conference, Montréal, Canada, Springer, pp. 951-952, 2003.

[7] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

[8] P. Viola, and M. Jones, "Robust Real-time Object Detection." Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, 2001.

[9] OpenCV: Open Source Computer Vision Library; http://www.intel.com/technology/computing/opencv/.

[10] Snack Sound Toolkit; http://www.speech.kth.se/snack/.