MUSIC-ELICITED EMOTION IDENTIFICATION USING OPTICAL FLOW ANALYSIS OF HUMAN FACE

V. V. Kniaz [a] *, Z. N. Smirnova [b]

[a] St. Res. Institute of Aviation Systems (GOSNIIAS), Moscow, Russia – vl.kniaz@gosniias.ru
[b] Gnessin Russian Academy of Music, Moscow, Russia – art.vivus@gmail.com

Commission V, WG V/5

KEY WORDS: optical flow, musical psychology, emotion identification, human face recognition

ABSTRACT:

Human emotion identification from image sequences is highly demanded nowadays. The range of possible applications can vary from an automatic smile shutter function of consumer grade digital cameras to Biofied Building technologies, which enables communication between building space and residents. The highly perceptual nature of human emotions leads to the complexity of their classification and identification. The main question arises from the subjective quality of emotional classification of events that elicit human emotions. A variety of methods for formal classification of emotions were developed in musical psychology. This work is focused on identification of human emotions evoked by musical pieces using human face tracking and optical flow analysis.

Facial feature tracking algorithm used for facial feature speed and position estimation is presented. Facial features were extracted from each image sequence using human face tracking with local binary patterns (LBP) features. Accurate relative speeds of facial features were estimated using optical flow analysis. Obtained relative positions and speeds were used as the output facial emotion vector. The algorithm was tested using original software and recorded image sequences. The proposed technique proves to give a robust identification of human emotions elicited by musical pieces. The estimated models could be used for human emotion identification from image sequences in such fields as emotion based musical background or mood dependent radio.

1. INTRODUCTION

Emotions are one of the most complex aspects of human behaviour. They deeply influence the way humans perceive the world in everyday life. The problem of human emotion monitoring, widely known as emotional intelligence (EI), is a rapidly developing field of modern psychology. Nowadays due to wide use of imaging devices and the rapid progress in image processing techniques vision based methods for emotion monitoring are becoming one of the main tools of EI. Moreover EI provides a basis for a range of brilliant new technologies based on human behaviour such as Biofied Building (Kita, 2015), which enables communication between building space and residents.

1.1 Emotions monitoring using facial expressions

The first systematic framework for emotional intelligence is commonly credited to Salovey and Mayer who define EI as 'the ability to monitor one's own and others feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions' (Salovey, 1990a). The authors outline two main kinds of expression of emotion: verbal and nonverbal. Nonverbal expression of emotion, being more instinctive and deeply rooted in human feelings, provides the way to express human mood and state of mind via facial expressions and gestures.

Though more complex in interpretation nonverbal expression of emotions could often provide more accurate information about human mood and state of mind, especially in case when verbal expression is unavailable (e.g. children nonverbal communication, cinema viewers, etc.). Being one of the most prominent indicators of emotions, human's face have attracted interests of many researchers for more than a century since Darwin's now classic study of facial expression (Darwin, 1872). However due to a highly dynamical nature of facial expressions their stable recording and monitoring could be done only by means of video recording system.

1.2 Computer vision based human emotion identification

Though first attempts to record human emotions by means of photography could be found as early as 1975 (Boucher, 1975), the rapid development of facial expression monitoring had begun only when consumer grade digital cameras became available in 1990s. The development MPEG-4 Face and Body Animation (FBA) International Standard (ISO/IEC 14496-1) in 1999 was a major keystone in the field of facial emotion modeling and classification. This standard defines 84 facial feature points (FP) on neutral face, that provide spatial references to general elements of the human's face (eyes, nose, lips, etc.) In addition MPEG-4 FBA defines Face Animation Parameters (FAP) including visemes, expressions, head rotation and low-level parameters to move all parts of the face. They could be used to either represent recorded facial expressions or to create facial expressions using deformable 3D-model of a face. The facial animation parameters are based on the study of minimal perceptible actions (MPA) and are closely related to muscle actions (Pandzic, 2002).

The problem of emotion identification using facial features and computer vision could be decomposed into three main steps:

* Corresponding author

1. Face detection and tracking
2. Estimation of the facial animation parameters – relative position and speeds of facial feature points in spatial space
3. Transformation from the facial animation parameters-space to emotion-space using emotion model

Usually the third step presents most of the problems due to dependence on the formal model of human emotions. Human emotions are highly subjective and nonlinear in nature. It is quite hard to describe them even using informal language. Thus the creation of an accurate model representing the transformation from facial animation parameters-space to emotion-space is an ill-defined problem. The design of such model requires the selection of a representative emotion-space and high-quality training data.

The base facial feature point set called Facial Action Coding System (FACS) proposed by Ekman et. al in (Ekman, 1978) has given a rise to series of publications on human emotion monitoring. The general approach is to collect training data by recording facial expressions a volunteer. Some specially selected images, video or dialogues are used to elicit a known set of emotions during the recording session. Esau et. al proposed to use a fuzzy emotion model to recognize facial expression based represented in FACS (Esau, 19. The training data was obtained by telling volunteers to express a single emotion. The proposed model showed 87% accuracy in recognition of 3D emotion space with axes representing anger, happiness and surprise.

Speech Analysis and Interpretation Laboratory (SAIL) developed an interactive emotional dyadic motion capture database (IEMOCAP) with twelve hours of training data. This database was recorded from ten actors in dyadic sessions with markers representing MPEG-4 facial feature points, which provide detailed information about their facial expression during scripted and spontaneous spoken communication scenarios (Busso, 2008). The actors played specially selected scripts to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state).

1.3 Music-elicited emotion identification

The use of scripted scenarios requires subjective labeling of emotions for model estimation. Moreover verbal communication may lead to rapid changes in emotions, which are quite hard to describe using simple models. This difficulty could be overcome by using music to elicit the desired emotions. Music being smooth and continuous in nature is composed to be emotionally expressive. Thus specially selected musical pieces could be used to evoke a broad range of continuously changing emotions.

In this paper we present the results of music-elicited emotion identification using optical-flow analysis of human face. A set of specially selected songs from '1000 Songs for Emotional Analysis of Music' dataset (Soleymani, 2013) and custom selected classical music pieces were used to record training data. We propose a modified face tracker based on LBP features for initial face position estimation and optical flow for accurate sub-pixel tracking of face center.

The same method is used for accurate tracking of facial feature points and estimation of their facial animation parameters. Both linear and nonlinear autoregressive exogenous models (ARX) were used to relate the extracted facial animation parameters with the music labeled in emotion-space. The estimated models were used to reconstruct elicited emotions in evaluation data.
The second part of this paper presents the music-elicited emotion model. In the third part data collection method is presented. Part four refers to modified LBP-optical flow face tracker. The fifth part presents the results of model estimation.

2. MUSICAL-ELICTED EMOTION MODEL

A number of state-space emotion representations were developed in musical psychology. A good balance between the flexibility and simplicity of the model provides valence-arousal emotion representation.

2.1 Valence-arousal emotion-space

One of the most well known emotion-space is so called 'Circumplex model' or valence-arousal (V-A) emotion representation proposed in (Russel, 1980). Being easy to use for both researcher and volunteer valence-arousal model represent all emotions in two-dimensional space. The first axes called valence represent the human's attitude to some object that could be positive or negative. The second axes called arousal indicates the strength of the emotion. Typical emotion-space vectors positions for affective words are shown in figure 1.

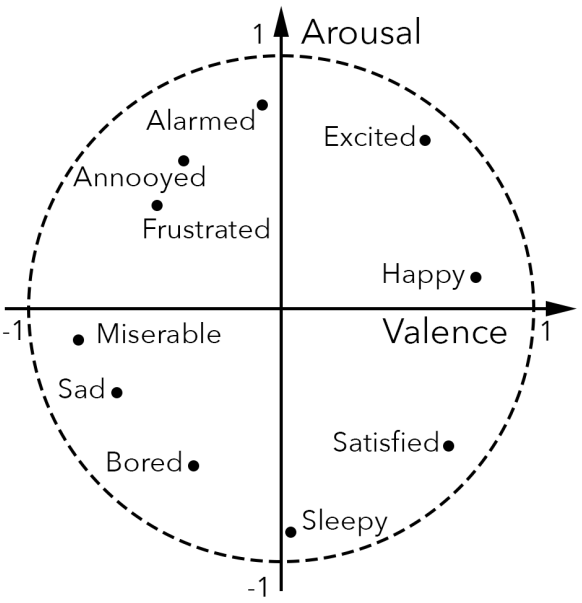

Figure 1. Valence-arousal emotion-space

2.2 The facial animation parameters-space

A number of features from original facial feature points set presented in MPEG-4 FBA were used to extract emotional expressions from images of human face. The selected feature points were used to estimate facial animation parameters that represent facial expressions (table 1).

An original two-dimensional state space was used to represent facial expressions:

$$\boldsymbol{Y}(t) = \begin{bmatrix} v_h \\ v_l \end{bmatrix} \quad (1)$$

$$v_h = |\varphi| + |\psi|$$

$$v_l = |v_{24}| + |v_{25}|$$

Here v_h – modulus of sum of head yaw ψ and roll φ speeds, v_l – modulus of sum of lip corners speeds v_{24}, v_{25}. Allowable yaw and roll angles for training data are restricted to ±30°. Therefore corresponding yaw and roll speeds can be calculated by dividing pixel speeds of feature point #9.12 (nose bone) by visible head radius. The configuration of the used facial feature points is presented on figure 2.

FAP #	FAP name	FP #	FP name
49	Head yaw	9.12	Middle lower edge of nose bone
50	Head roll	9.12	Middle lower edge of nose bone
6	Horizontal displacement of left lip corner	2.4	Left corner of inner lip contour
7	Horizontal displacement of left lip corner	2.5	Right corner of inner lip contour

Table 1. Facial animation parameters and corresponding facial feature points used for representations of facial expression

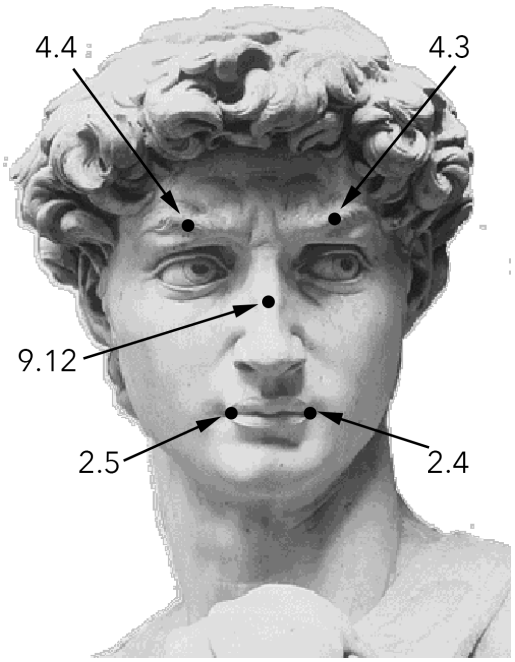

Figure 2. Configuration of the used facial feature points

2.3 Musical-elicited facial expression dynamical model

As the valence-arousal emotion representation can provide a smooth and continuous representation of emotional flow a linear time invariant (LTI) model of facial expressions elicited by emotions could be used:

$$\dot{\boldsymbol{X}}(t) = \boldsymbol{A}\boldsymbol{X}(t) + \boldsymbol{B}\boldsymbol{U}(t) \quad (2)$$
$$\boldsymbol{Y}(t) = \boldsymbol{C}\boldsymbol{X}(t) + \boldsymbol{D}\boldsymbol{U}(t)$$

Here output vector $\boldsymbol{Y}(t)$ is given by equation (1), input vector $\boldsymbol{U}(t)$ is composed of valence value v and arousal value a:

$$\boldsymbol{U}(t) = \begin{bmatrix} a \\ v \end{bmatrix} \quad (3)$$

However due to highly nonlinear nature of human emotions only rough similarity of the LTI model output with the real emotions could be expected. To estimate more accurate emotion model a nonlinear autoregressive exogenous models (NARX) with wavenet network was used. The original model structure proposed in (Zhang, 1997) was used.

3. DATA COLLECTION METHOD

Training and evaluation data was recorded from ten volunteers listening to musical pieces. A number of specially selected musical pieces with various emotional features were used. To record image sequences and label the musical pieces using V-A emotion-space an original emotion capture software was developed.

3.1 Emotion capture software

To benefit from distributed data collection from many volunteers a mobile platform was selected for software development. The following functions were implemented in the software:

- Video recording at 30FPS synchronized with the music playback
- Valence and arousal recording using user interface (dynamic slider) at 30FPS synchronized with the music playback
- Database keeping

The user interface of ‘EmotionGrabber’ software is presented on figure 3.

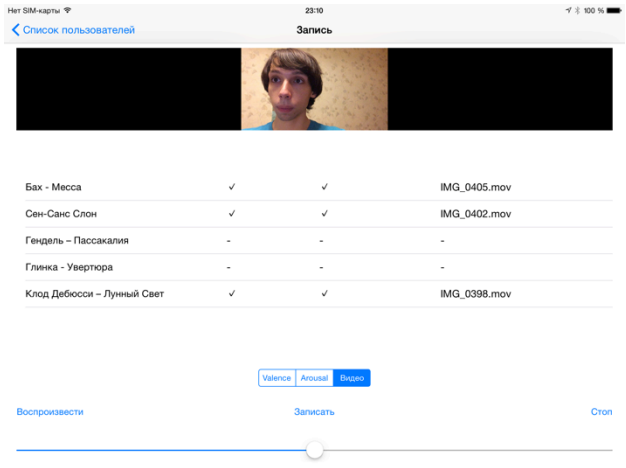

Figure 3. ‘EmotionGrabber’ software user interface

3.2 Musical pieces selection

Music is composed to elicit a variety of emotions and often the composer state the proposed emotion in the name of the musical piece. A number of classical music pieces were carefully selected to provide a variety of emotions. The most prominent emotional fragment of each musical piece with the length of 45s was extracted and uploaded to 'EmotionGrabber' software. The list of musical pieces, their emotional labels, and averaged valence and arousal labeled by volunteers are presented in table 2.

3.3 Data recording workflow

Each volunteer has listened to musical pieces three times. During the first listening session the video recording was done. During the second and the thirds session arousal and valence labelling was performed. To avoid loss of concentration on the music a five-minute pause was done between after every third recording session. First fifteen seconds of recording were discarded, as they are distorted due to initial concentration on 'EmotionGrabber' software user interface.

#	Musical piece	Composer	Emotion Label	V	A
1	Passacaille (No.6), Suite in G minor, HWV 432	Handel G. F.	Sad	-0.8	-0.3
2	Mass in B minor, BWV 232	Bach J. S.	Miserable	-0.9	-0.1
3	Lacrimosa (Requiem)	Mozart W.A.	Sad	-0.8	-0.2
4	Mephisto Waltz No. 1, S.514	Liszt F.	Alarmed	-0.1	0.8
5	Overture from opera 'Ruslan and Lyudmila'	Glinka M. I.	Excited	0.6	0.6
6	The Poem of Ecstasy, Op. 54	Scriabin A. N.	Aroused	0.1	0.9
7	Isle of the Dead, Op. 29	Rachmaninoff S. V.	Tense	-0.1	0.8
8	The Carnival of the Animals, Part V 'The elephant'	Saint-Saëns C.	Happy	0.8	0.2
9	Clair de Lune	Debussy C.	Serene	0.3	-0.5
10	Danse sacrale from ballet 'The Rite of Spring'	Stravinsky I. F.	Afraid	-0.3	0.7

Table 2. List of musical pieces

4. DATA PROCESSING

To process the recorded data a modified LPB-optical flow face tracker algorithm was proposed. The algorithm was implemented in original software for off-line processing of the recorded data.

4.1 Modified LBP-optical flow face tracker

LBP feature trackers are commonly used to track human face in image sequences. However the accuracy of such tracker is usually not good enough to track subtle sub-pixel motions. The LBP-tracker also tends to have random false face detections, which should be detected and filtered.

An optical flow based tracking of human face could be used to detect sub-pixel motion of the face. However due to momentary errors in optical flow calculation such trackers usually have an accumulated error. A modified LBP-optical flow face tracker is proposed to overcome drawbacks of both trackers. The LBP-optical flow trackers uses LPB features to detect the face position on the first frame. After that an optical flow is calculated for the head region for each consequent frame. Non-zero regions of optical flow are averaged across the face region to obtain accurate sub-pixel movement of the frame. The calculated movement is compared with the difference LBP-tracker estimations of head position. If the difference is smaller than maximum allowable head movement between two frames, the optical flow field estimated movement is fused with the movement estimated by LBP-tracker. The fusion between two motion estimates is done using given ratio *k*. The algorithm of the proposed tracker is summarized in figure 4. The algorithm was implemented using optical flow estimation method proposed in (Farnebäck, 2003).

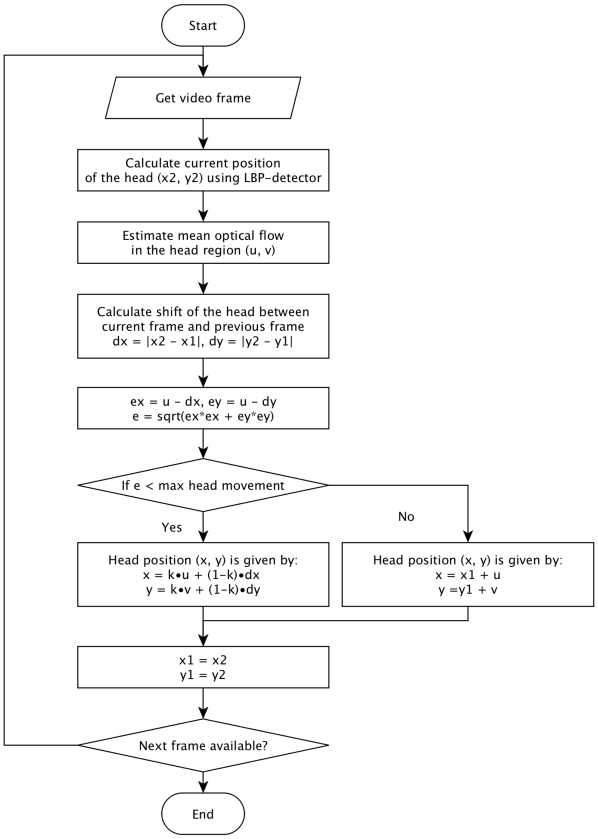

Figure 4. LBP-optical flow face tracker algorithm

4.2 Facial feature points trackers

A general structure of facial feature points' locations was taken from MPEG-4 FBA standard. To correct the locations of abstract facial feature points to their locations on particular face an LBP-classifier for each facial feature point was generated. Real locations of the feature points on the first frame were found by search in the local windows centered at general feature point location. After that modified LBP-optical flow tracker was used to track position of each feature relative to the head's position on consequent frames.

The process of facial feature points tracking is presented on figure 5.a. Position of feature point #9.12 (nose bone) estimated using basic optical flow tracker is marked with an ellipse. Due to accumulated error in optical flow estimation the tracked feature position has moved away from the actual position (up and toward left eye). Position of feature point #9.12 estimated using LBP-optical flow tracking is marked with a rectangle. Static error of the feature point position estimated using the proposed tracker doesn't exceed 1 pixel. Positions of feature points #2.5, 2.4 (lip corners) are marked with black circles.

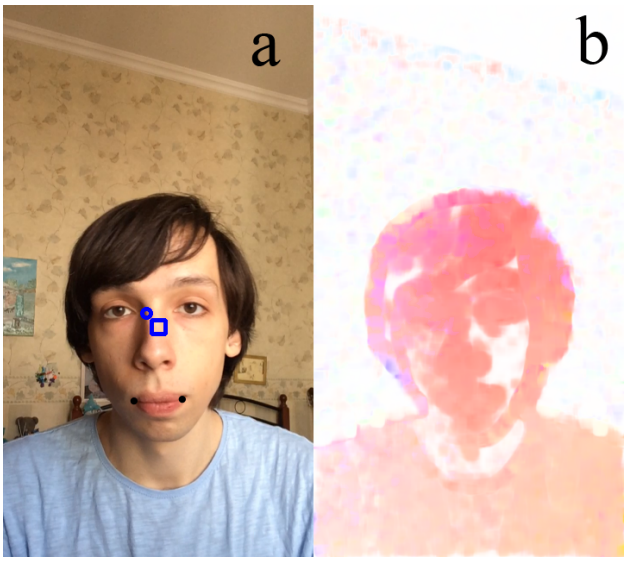

Figure 5. Facial expression extraction using LBP-optical flow tracker. (a) Extracted facial feature points. (b) Corresponding optical flow in Middlebury color-coding.

5. MUSIC-ELICITED EMOTION MODEL ESTIMATION

Two kinds of models were used to represent the connection between labelled emotions and facial expressions LTI models and NARX models.

5.1 LTI model estimation

A state-space model of the system given by (2) was identified from the recorded data using Matlab System Identification Toolbox (Ljung, 2003). The normalized root-mean-square error (NRMSE) of the estimated model was 75%. The comparison of original output data and output of the estimated model is shown figure 7.

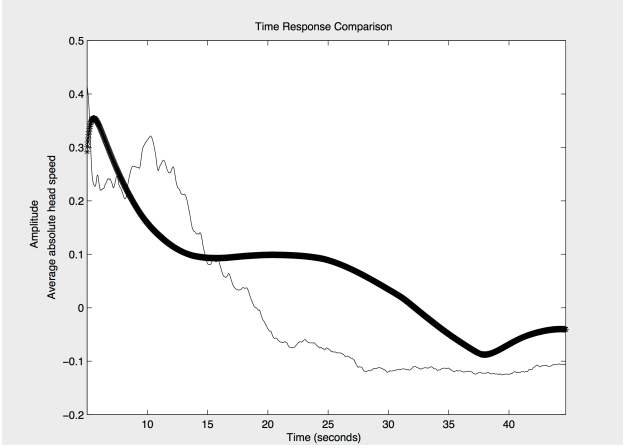

Figure 7. Original output and the output of the LTI model

5.2 NARX model estimation

The NRMSE of the estimated model was 17% (figure 6). The significant reduction of NRMSE of the NARX model could be explained by a better representation of nonlinear nature of emotional behavior by NARX model.

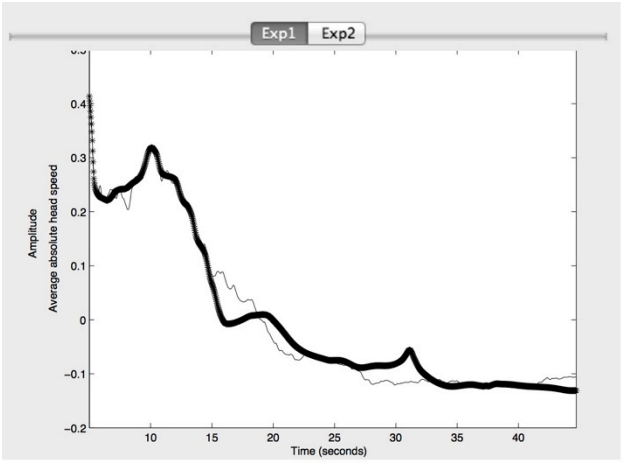

Figure 7. Original output and the output of the NARX model

6. CONCLUSIONS

A method for musical-elicited emotion model estimation was proposed. The model is estimated using the black-box model identification method. The input vector for model identification represents emotion-space, the output vector – facial expression space. The estimated model could be used to reconstruct musical-elicited emotion in valence-arousal emotion-space from recorded image sequence.

The proposed method was tested using specially selected list of classical musical pieces with various emotional content. A training data was recorded from ten volunteers listening to musical pieces. Dedicated software for a mobile tablet computer was designed to automate the recording workflow.

A modified LBP-optical flow optical was proposed to track face and facial features on the recorded video. The proposed tracker

fuses data from optical flow motion estimation and LBP-tracker to achieve robust performance and sub-pixel feature tracking.

Two kinds of model were used to represent the connection between elicited emotions and facial expressions. LTI model failed to reconstruct dynamic nonlinear changes in facial expressions. NARX model proved to be accurate enough to reconstruct emotions on the evaluation data with 17% NRMSE.

REFERENCES

Boucher, J. D., Ekman P. 1975. Facial Areas and Emotional Information. *Journal of Communication*, 25 (2).

Busso C., Bulut M., Lee C., Kazemzadeh A., Mower E., Kim S., Chang J. N., Lee S., Narayanan S. S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42 (4), pp. 335-359.

Darwin, C., 1872/1955. Expression of the Emotions in Man and Animals. *Philosophical Library*, New York.

Ekman P., Friesen. W. V. 1978. *Facial Action Coding System*. Consulting Psychologists Press Inc., California,

Esau, N., Wetzel, E., Kleinjohann, L., Kleinjohann, B. 2007. *Real-Time Facial Expression Recognition Using a Fuzzy Emotion Model*, Fuzzy Systems Conference, FUZZ-IEEE 2007. IEEE International, pp. 1-6

Farnebäck G., 2003. Two-Frame Motion Estimation Based on Polynomial Expansion, *Image Analysis, 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden*, pp. 363-370

Kita S., Mita A., 2015. Emotion Identification Method using RGB information of Human Face, *Proc. SPIE 9435, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2015*

Ljung, L., 2003. *System Identification Toolbox*, The MathWorks, Inc

Pandzic, I., Forchheimer R. 2002. MPEG-4 Facial Animation - The standard, implementations and applications. John Wiley & Sons.

Russell. J. A., 1980. A circumplex model of affect. *Journal Personality Social Psychology*, 39, pp. 1161–1178.

Salovey, P., Mayer, J. D, 1990. Emotional intelligence. *Imagination, Cognition, and Personality*, 9, pp 185-211.

Soleymani, M., Caro M. N, Schmidt E. M., Sha C.-Y., Yang Y.-H. 2013. 1000 Songs for Emotional Analysis of Music, *CrowdMM '13 Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pp. 1-6.

Zhang, Q. 1997. Using wavelet network in nonparametric estimation, *IEEE Transactions on Neural Network*, Vol. 8, No. 2