AUDIO-VISUAL FEATURE-DECISION LEVEL FUSION FOR SPONTANEOUS EMOTION ESTIMATION IN SPEECH CONVERSATIONS

Aya Sayedelahl, Rodrigo Araujo, and Mohamed S. Kamel

Center for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, Canada
{asayedel, raraujo, mkamel}@pami.uwaterloo.ca

ABSTRACT

Recognition of spontaneous human affect has gained a lot of interest recently after a shift in research focus from the traditional six prototypical emotions. In this paper, a combined bi-modal feature-decision fusion approach is proposed to enhance the performance of estimating emotions from spontaneous speech conversations. First, a feature vector consisting of audio information extracted from the whole speech sentence is combined with video features of the individual key frames representing that sentence. Then, the final estimate is calculated by a decision level fusion of predictions from all corresponding frames. The performance is compared with two fusion approaches, the decision level fusion using weighted linear combination, and a simple feature level fusion. The experimental results show improvement in correlation between the emotion estimates and the audio references. In addition, we evaluated the performance of supervised PCA (SPCA) for dimensionality reduction on this audio and video emotion estimation application, which led to better results compared to traditional principle component analysis (PCA).

Index Terms— Emotion Estimation, Fusion, Regression, Audiovisual.

1. INTRODUCTION

Automatic analysis and recognition of human emotional behavior has gained a lot of interest and is considered an important step towards building efficient and more realistic intelligent human-computer interfaces. Computers' ability to perceive and respond to human emotions, which is a non-verbal means of communications, can result in an improved natural interaction between computers and humans [1]. Most of the past research in the field of affective computing focused on uni-modal recognition of prototypical basic emotions mainly from facial expressions [2], and audio cues [3], [4] using acted databases that were captured in controlled predefined lab settings [5]. Although major advances within the field of affective computing research has been achieved, automatic recognition of emotions occurring in natural settings is recent, not fully explored, and is considered a challenging problem [6]. This is due to the nature of real life emotions, which are usually subtle, mixed, and more complex, compared to acted ones. This in turn, increases the difficulty to map the affective human state into a single label or discrete number of classes [7]. Recently, the focus was directed towards recognition of emotions from natural spontaneous expressions in terms of dimensional and continuous description rather than small number of discrete emotion categories [8]. The problem is converted from a classification problem to a regression one. The dimensional representation of emotions are described using the emotion space concept [9] in terms of three dimension primitives; valence (positive to negative), activation (calm to excited), and dominance (weak to strong) dimensions. Given the gradient nature of emotions, a more realistic approach is to represent emotions in a multi-dimensional space. This enables the accurate description of the intensity of emotions in realistic settings [10].

In this paper, we further work with the problem of dimensional spontaneous emotion estimation from speech sentences. We propose a combination of feature and decision level fusion, utilizing both audio and facial cues, to enhance the estimation of continuous labeled emotions in natural spontaneous expressions. Furthermore, we investigate the use of Supervised PCA (SPCA) [15] for dimensionality reduction and compare it with the popular PCA approach.

The rest of the paper is organized as follows. In section 2, a brief description of some of the related work on dimensional emotion recognition from spontaneous expressions with the focus on multimodal emotion recognition is presented. Section 3 presents an overview of the proposed fusion approach. Section 4 describes the pre-processing, feature extraction, and dimensionality reduction used in this work. Section 5 presents the experimental setup, and a brief description of the database used in this study. Results and discussion are presented in section 6 followed by the conclusion and the future work in section 7.

2. RELATED WORK

Most of the recent studies on dimensional recognition of emotions from natural spontaneous expressions concentrated mainly on using audio information [10], [11], [12], and fewer studies have tackled the problem of dimensional emotion recognition in spontaneous expressions using multimodal information [6], [8] , [13].

In [6], audio and facial cues were used to classify spontaneous expressions of realistic human conversation sequence into positive and negative emotions. In [8], audio, facial, and shoulder cues were used for the continuous time prediction of spontaneous emotions in valence and arousal dimensions. While in [13], audio and visual cues were used to recognize emotions in spontaneous expressions in three dimensions, valence, activation, and dominance.

This work in contrast to [6], which treats the problem as binary classification, is a regression problem where the emotions are described in continuous labeled dimensions. In addition, we focus on estimating the underlying intensity of emotion, in valence, activation and dominance dimensions, from the whole sentence in order to detect the overall degree of emotion variation in the conversation rather that continuous prediction of emotions at each time frame [8]. In [13], two separate models are trained, and a decision level fusion is used to fuse the two uni-modal estimates, while in this work, one model is trained and a combination of feature and decision level fusion is employed.

3. FUSION APPROACH

This section presents an overview of the proposed fusion approach and the motivation behind using a combined fusion scheme. Figure 1 shows the block diagram of the proposed emotion estimation system.

First, each sentence is represented by the speech utterance *U* and corresponding number of facial image frames *k= [1, 2,..., N]*, where N is a variable number due to the difference in duration of different sentences. Second, a feature vector is then created which consists of the audio features extracted from the whole utterance concatenated with visual features, extracted from each of the individual visual frames representing that sentence as shown in Figure 1.Third, a frame-level regression model is then trained using the combined audio and visual features for each face image with the audio ground truth labels of the speech sentences used for training and evaluation of the model. Finally, a simple decision aggregation rule is used by averaging the resulting estimates of all image frames corresponding to that sentence for the final emotion estimation of the whole sentence.

The main goal is to design an efficient fusion approach that can deal with the following challenges associated with the emotion recognition problem at hand. First, when dealing with natural spontaneous emotions, the variation in facial expression along the duration of the sentence can be unpredictable. This makes it very challenging to effectively choose one key frame that will provide an accurate representation of the overall emotional state in the sentence, in contrast to acted expressions such as in the case presented in [14] where one key frame with the highest speech amplitude is extracted from the video sequence to represent the emotional state in the sentence. This was based on the observation that, in acted expressions, the human facial features are usually exaggerated at large voice amplitudes. In this work, a variable number of image frames extracted from the video sequence corresponding to each sentence is used to represent the variation in facial expression along that sentence. The facial image frames provided in the database are used as the visual key frames in this work.

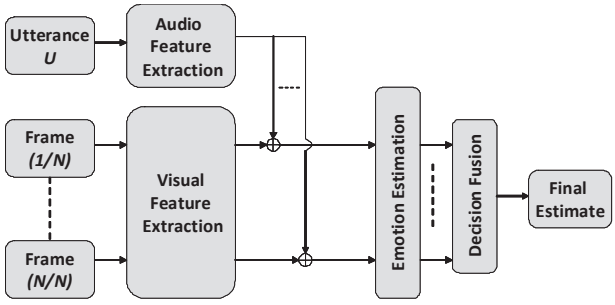

Fig. 1. The block diagram of the proposed emotion estimation system.

Second, when dealing with the fusion at the decision level, one of the issues that arise is how to effectively fuse the continuous outputs of the audio and visual models. In [13] the results of individual uni-modal estimation are fused at the decision level by a weighted linear combination. Although this approach provided enhancement in performance, the adaptive adjustment of weighting factor still poses an extra challenge [13]. In addition, we want to avoid the use of weighted linear combination since different optimal weight combinations can occur when different features, databases, and/or different models are used. Consequently, we investigated the use of a simple fusion rule at the final stage of our fusion approach.

4. FEATURE EXTRACTION AND DIMENSION REDUCTION

In this section, a brief description of the extracted audio and visual features and the dimensionality reduction used in this work are presented.

For audio feature extraction, a set of prosodic and spectral features mainly short time energy, fundamental frequency, and 14 Mel frequency cepstral coefficients (MFCC) are used. First, each input speech utterance is filtered using pre-emphasis filter and then divided into a set of overlapping frames using hamming window of duration

25 ms and 10 ms shifts. The features are extracted from each frame and statistical measures (min, max, median, mean, and variance) are calculated from all the frames to produce a final fixed feature vector of size 70 for each utterance.

For visual feature extraction, Local Binary Patterns (LBP) features are used in this work based on the approach described in [16].The face region in each frame is first detected using the real time face detection approach by Viola and Jones [17].The resulting face regions are then normalized as follows, the coordinates of the two eyes are identified, based on the OpenCV implementation of a Haar-cascade object detector, trained for either a left or a right eye, then the size normalization is done by resizing the image with distance between the eyes of 55 pixels. Afterwards, the whole face image is cropped to the size of 150 x 110, relative to eyes' position. The resulting face images are then divided into 7 x 6 sub-regions. Finally, the LBP descriptors are extracted for each region and then the histograms are mapped into uniform patterns in an (8, 2) neighborhood. The final feature vector is of size 2,478 (7 x 6 x 59) for each face.

When dealing with large number of features, it is beneficial to reduce the dimensionality of the feature vector to reduce the computation complexity during training. Furthermore, larger number of the features does not always guarantee better performance. In this work, two sets of reduced dimension features employing the widely used PCA technique and a recent supervised PCA (SPCA) approach developed in [15] is investigated. Supervised principle component analysis is a dimensionality reduction technique that aims at maximizing the dependency between the projection of data X and the output variable Y through a projection matrix U. The dependency between the projected data and the output variable is measured using the Hilbert-Schmidt independence criterion (HSIC) [18]. Assuming F and G to be two separable reproducing kernel Hilbert spaces (RKHS) and $Z := \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq X, Y$ be a series of n independent observations, an empirical estimate of HSIC is given by:

$$\mathrm{HSIC}(Z, F, G) = (n-1)^{-2} \operatorname{tr}(\mathrm{KHLH}) \tag{1}$$

where K is a kernel of $U^T X$, L is a kernel of Y, and $H = I - n^{-1} e e^T$ (e is a column vector of 1's and n is the dimensionality of the random variables). According to this measure and the work by Barshan [15], the dependence between the linear kernel of XU and a kernel of Y is maximized by maximizing the corresponding HSIC empirical estimate as shown in equation (2)

Fig. 2. Sample images from the VAM-Faces database.

$$\begin{aligned} &\underset{U}{\operatorname{argmax}}\ \operatorname{tr}(U^{\mathrm{T}} \mathrm{XHLHX}^{\mathrm{T}} U) \\ &\text{subject to } U^T U = I \end{aligned} \tag{2}$$

One of the advantages of using SPCA is that, the algorithm can be kernelized which makes it applicable to non-linear dimensionality reduction tasks [15]. In this work we investigated the use of SPCA using the RBF kernel. The dimensional reduction step is performed after the audio and visual features are concatenated which gave better results in our work compared to reducing the dimension before concatenation. This is performed for both the fusion at feature level and our proposed fusion approach.

5. EXPERIMENT

5.1. Database

Since the main focus of the work is the estimation of continuous labeled emotions in spontaneous conversations, the VAM corpus [19] is used for this study. It is an authentic spontaneous audio-visual database from real life conversations, which was recorded from the German TV talk show "Vera am Mittag", in which guests mostly talk about their personal issues in a spontaneous, affective and unscripted manner [19]. Figure 2 shows sample images from the database. The emotions are described with three continuous-valued emotion primitives, namely valence, activation, and dominance. Our fusion approach is tested on a subset of the database containing a total of 234 audio sentences that has been previously used in [13] and their corresponding annotated subset of facial image sequence with a total of 1486 facial images.

5.2. Emotion Estimation

In this work, the emotion estimation from each modality is first investigated alone and then using both audio and visual information. To investigate the effectiveness of our fusion approach, the results are compared to another two fusion approaches, fusion at feature level only, and fusion at the decision level.

For the fusion at feature level, the visual features from the frames representing the speech sentence are combined together by averaging the LBP histograms of the corresponding blocks of each frame, similarly to what was done in [20]. Then, the resulting feature vector is concatenated with the audio feature vector to form the final combined feature vector.

For the decision level fusion, two separate models are trained, one using the audio features from the whole utterance, and the other using visual features from the corresponding frames. We employed the weighted linear combination similar to the fusion used in [13], for the purpose of comparison, to fuse the results.

$$Y_{av} = wY_v + (1-w)Y_a \quad (3)$$

Where Y_{av} is the combined estimation, Y_v is the visual estimated results, Y_a is the audio estimated results, and w is the weighted factor. The best estimation results were chosen after varying the weight factor from 0 (audio only) to 1(video only),

5.3 Experimental Setup

In this work, the emotion estimation is performed using Support vector regression (SVR) [21] with RBF kernel [$k(\mathbf{x},\mathbf{y})=\exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$]. Three separate models are developed to estimate the continuous values of the three emotional dimensions (valence, activation, and dominance). The design parameters of the SVR model, which consists of the penalty parameter C, the loss function ε, and the kernel parameter γ are optimized on the training subset by performing 5-fold cross validation then the best parameters were used to retrain the SVR model using the whole training set. LIBSVM toolbox was used for the implementation of epsilon-SVR model [22].

10-fold cross validation is used to evaluate the performance of the uni-modal as well as the fusion models. The data is split in a way to insure that faces corresponding to a certain sentence are together during either the training or the testing step for each run. For fair comparisons, the cross validation indices have been kept the same during all the experiments done in this work. The main performance measure is taken as the average correlation coefficient of the 10 fold runs, which measures the strength of the relationship between the estimates and the references for each emotion primitive separately. The average mean linear error is also calculated. We performed three experiments, one using the whole feature set and the other two on the reduced dimension feature set by applying PCA and supervised PCA approaches.

For the dimensionality reduction using PCA approach, three experiments were performed using different feature vector dimensions (the components that retained 90%, 95%, and 99% of the variance). It was noticed that the components that retained 90% of the variance gave the best results and adding more features didn't provide any improvement. For SPCA, we evaluated the experiments at different feature dimensions and the number of features that produced the highest correlation was used as the final feature vector dimension.

6. RESULTS AND DISCUSSIONS

In this section, we present the experimental results of our spontaneous emotion estimation problem. Table 1 shows the experimental results and comparisons using the complete set of audio and visual features, while Tables 2 and 3 present the results in the reduced feature space using PCA and SPCA respectively.

Looking at the fusion results, we can see that the fusion at feature level consistently shows the least performance for almost all experiments. This is believed to be due to the inherent problem associated with using the whole feature vector (e.g. the very high dimensionality of the feature vector after concatenating the audio and visual features accompanied by lower sample size). Although better performance is shown after applying dimensionality reduction, the performance of the simple feature level fusion is still the least performing. This is caused by the loss in details associated with the variation of facial emotions along the sentence, which resulted from combining the LBP histograms for all the key frames.

For the fusion at the decision level, the variation in facial expression from the frames corresponding to the speech sentence are preserved and taken into consideration when training the visual model. This is done by using the feature vector of all individual frames. This resulted in an improvement in performance compared to fusion at feature level. However, dealing with this regression problem, another challenge arises when training two separate regression models with different continuous reference labels for each modality, which can affect the final results.

By using a combination of feature and decision fusion proposed in this work, we are able to preserve the variation of facial emotions along the sentence, by taking into consideration all the information from the key frames representing that sentence and at the same time avoiding training two separate models with different ground truth labels. This approach has shown further improvement to the emotion estimation problem at hand.

Using SPCA features was shown to yield higher performance compared to PCA features for all experiments, uni-modal, as well as all fusion schemes along all three dimensions. This clearly demonstrates the positive impact of supervised dimensionality reduction on the system performance.

Comparing the results from all the experiments, we can see that using SPCA features provided the best performance. In addition, our combined feature- decision

Table 1. Average correlation coefficient and (average mean linear error) comparisons using the whole feature set.

All Features	Valence	Activation	Dominance
Audio only	0.62 ± 0.19 (0.12 ± 0.02)	0.79 ± 0.07 (0.17 ± 0.02)	0.77 ± 0.07 (0.14 ± 0.02)
Video only	0.57 ± 0.19 (0.13 ± 0.02)	0.61 ± 0.11 (0.21 ± 0.03)	0.58 ± 0.12 (0.18 ± 0.02)
Feature-level	0.62 ± 0.18 (0.12 ± 0.02)	0.78 ± 0.08 (0.17 ± 0.02)	0.75 ± 0.09 (0.14 ± 0.01)
Decision-level	0.65 ± 0.15 (0.12 ± 0.02)	0.80 ± 0.09 (0.17 ± 0.02)	0.78 ± 0.06 (0.14 ± 0.02)
Proposed Fusion	0.67 ± 0.13 (0.11 ± 0.02)	0.84 ± 0.05 (0.14 ± 0.02)	0.79 ± 0.10 (0.13 ± 0.02)

Table 2. Average correlation coefficient and (Average Mean linear error) comparisons for the PCA features.

PCA	Valence	Activation	Dominance
Audio only	0.61 ± 0.19 (0.12 ± 0.02)	0.77 ± 0.07 (0.18 ± 0.02)	0.77 ± 0.08 (0.14 ± 0.03)
Video only	0.59 ± 0.15 (0.13 ± 0.02)	0.65 ± 0.09 (0.21 ± 0.03)	0.55 ± 0.2 (0.18 ± 0.03)
Feature-level	0.65 ± 0.13 (0.12 ± 0.02)	0.76 ± 0.07 (0.18 ± 0.02)	0.69 ± 0.13 (0.16 ± 0.02)
Decision-level	0.66 ± 0.14 (0.12 ± 0.02)	0.79 ± 0.08 (0.17 ± 0.02)	0.78 ± 0.07 (0.14 ± 0.02)
Proposed Fusion	0.68 ± 0.13 (0.11 ± 0.02)	0.85 ± 0.05 (0.13 ± 0.02)	0.81 ± 0.09 (0.12 ± 0.02)

Table 3. Average correlation coefficient and (Average Mean linear error) comparisons for the SPCA features.

SPCA	Valence	Activation	Dominance
Audio only	0.62 ± 0.18 (0.12 ± 0.02)	0.80 ± 0.05 (0.16 ± 0.02)	0.79 ± 0.06 (0.13 ± 0.02)
Video only	0.67 ± 0.13 (0.11 ± 0.02)	0.73 ± 0.10 (0.18 ± 0.03)	0.66 ± 0.05 (0.16 ± 0.02)
Feature-level	0.70 ± 0.14 (0.10 ± 0.02)	0.81 ± 0.08 (0.16 ± 0.03)	0.75 ± 0.08 (0.14 ± 0.01)
Decision-level	0.68 ± 0.13 (0.12 ± 0.02)	0.82 ± 0.06 (0.16 ± 0.02)	0.80 ± 0.05 (0.13 ± 0.02)
Proposed Fusion	0.74 ± 0.10 (0.09 ± 0.02)	0.86 ± 0.05 (0.13 ± 0.02)	0.82 ± 0.07 (0.12 ± 0.02)

level fusion achieved the best performance in a consistent fashion along all feature combinations used in this work. The results also show that considerable improvement in correlation is achieved compared to the mean linear error reduction. The best average correlation between the emotion estimates and the audio references using our proposed approach achieved an increase of 12%, 6%, and 3% for valence, activation, and dominance respectively , compared to 6%, 2%, 1% increase using the decision level fusion approach, and 8%, 1% for valence and activation respectively using feature level fusion.

From the percentage of improvement shown above in the three dimensions, it can be inferred that incorporating visual information yielded the highest improvement in valence dimension, while the least improvement was achieved in the dominance dimension. This is consistent with the evaluation of raters in [13] which shows that during evaluation, valence was relatively the hardest to detect using audio cues in contrast to activation which was easier to detect using audio cues, while dominance was the hardest to detect from facial cues.

Table 4 presents the comparison between our best achieved fusion results and the fusion results reported in [13] on the same speech samples. In addition, results of the audio modality which was used as the reference, is also shown in the same table. The main purpose from adding the audio reference is to identify the relative improvement for each aspect of the work since different visual features and different dimensionality reduction technique were used. To be able to compare our results with the results reported in [13], and to assure fair comparison, we calculated the total correlation coefficient and total mean linear error. Based on the results shown in table 4, it is shown that our fusion approach achieved an increase in correlation, between the final audiovisual estimates and the audio references, of 11%, 6%, 5% for valence, activation and dominance respectively compared to 17%, 2%, 2% increase in correlation reported in [13]. It is clearly shown that the higher percentage of improvement for the valence dimension in [13] is due to the very low correlation coefficient obtained from the audio features and the final result of 0.70 was obtained from visual information only. It is also shown that in our case the highest improvement in correlation is achieved in the valence dimension. Finally, it is also shown that a considerably higher correlation coefficient of 0.63 is obtained from the transformed audio feature using SPCA in the valence dimension compared 0.53 from the best audio features selected in [13] using SFFS approach.

Table 4. Correlation coefficient (CC) and mean linear error (MLE) comparisons.

All Features	Valence		Activation		Dominance	
	CC	MLE	CC	MLE	CC	MLE
Audio (ours)	0.63	0.12	0.80	0.16	0.78	0.13
Audio [13]	0.53	0.13	0.82	0.16	0.78	0.14
Audiovisual (ours)	**0.74**	**0.10**	**0.86**	**0.12**	**0.83**	0.12
Audiovisual [13]	0.70	0.14	0.84	**0.12**	0.80	**0.09**

7. CONCLUSIONS

In this work, a combined feature decision fusion scheme is proposed to enhance the emotion estimation of continuous values in three dimensional space, valence, activation and dominance, of spontaneous speech conversations. It is shown that the proposed approach provided improvement in performance compared to both fusion at the feature level and the decision level fusion using weighted linear combination approach. In addition, the approach avoids the problem of training two separate regression models with different continuous reference labels for each modality, which poses another challenge on how to effectively fuse the results. Using SPCA for dimensional reduction has also shown considerable improvement in the results, for both uni-modal and all the fusion schemes used in this work, compared to the traditional PCA. For future work, we plan to investigate the performance of our fusion approach on larger datasets and longer conversations. One of the challenges when dealing with larger datasets is the increase in the number of image frames. We will explore the use of clustering approaches for efficient selection of subset of facial frames from the video sequence. The main goal is to further reduce the number of key frames and at the same time still provide best representation of the speech sentence.

8. REFERENCES

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," *Sixth International Conference on Multimodal Interfaces ICMI*, pp. 205–211, 2004.

[2] L. Zhang, and D. Tjondronegoro, "Facial Expression Recognition Using Facial Movement Features," *IEEE Trans. on affective computing*, vol.2, Issue 4, pp. 219-229, 2011.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, vol.4, pp.IV-957-IV-960, 2007.

[4] J. H. Jeon, R. Xia, Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.4940-4943, 2011.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recogn* 44(3), pp. 572–587, 2011.

[6] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang, "Audio–visual spontaneous emotion recognition," *Proceedings of the ICMI 2006 and IJCAI 2007 international conference on Artifical intelligence for human computing*, pp. 72–90, 2007.

[7] H. Gunes, and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.

[8] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. on Affective Computing*, vol. 2, no. 2, 2011.

[9] R. Kehrein, "The prosody of authentic emotions," *Proceedings of Speech Prosody Conf.*, pp. 423–426, 2002.

[10] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," *IEEE International Conference on Multimedia and Expo (ICME)*, pp.737-742, 2010.

[11] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[12] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," *Proc. 9th Interspeech 2008*, pp. 597– 600, 2008.

[13] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-Visual emotion recognition using an emotion recognition space concept," *Proc. 16th European Signal Processing Conf.*, 2008. Available at: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569103398.pdf

[14] Y. Wang, and L. Guan, "Recognizing human emotion from audiovisual signals," *IEEE Transactions on Multimedia*, 10(5): 936-946, 2008.

[15] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised Principal Component Analysis: visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357–1371, 2011.

[16] C. Shan, S. Gong, P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vision Comput*. 27(6), pp.803–816, 2009.

[17] P. Viola and M. Jones, "Robust real-time object detection," *in Proc. of 2nd IEEE Workshop on Statistical and Computational Theories of Vision*, pp. 1–25,2001.

[18] A. Gretton, O. Bousquet, A.J. Smola, B. Scholkopf, "Measuring statistical dependence with Hilbert–Schmidt norms", in Proceedings Algorithmic Learning Theory (ALT), vol. 3734, pp. 63–77,2005

[19] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," *IEEE Int. Conf. on Multimedia& Expo (ICME)*, pp.865-868, 2008.

[20] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, Maja Pantic: "AVEC 2012 – The Continuous Audio/Visual Emotion Challenge"**,** *in Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012*, ACM, Santa Monica, CA,2012.

[21] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, and V. Vapnik, "Support Vector Regression Machines," *Advances in Neural Information Processing Systems,* pp. 155-161, MIT Press, 1996.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines,"*ACM Transaction on Intelligent Systems and Technology* 2:27:1—27:27, 2011. Software is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.