Brunel University London

Automatic Affective Dimension Recognition from Naturalistic Facial Expressions Based on Wavelet Filtering and PLS Regression

Yona Falinie A. Gaus, **Hongying Meng**
Asim Jan, Fan Zhang, Saeed Turabzadeh

Outline of the presentation

1. Emotion recognition from facial expression
2. Focus of this work
3. Proposed System
   - Image Feature Extraction
     - Edge Orientation Histogram (EOH) /Local Binary Pattern (LBP)/Local Phase Quantization (LPQ)
   - Audio Feature Extraction
     - MFCC
   - Wavelet Filtering
     - Haar Wavelet Transform
   - Machine Learning
     - Partial Least Square (PLS) Regression
   - Filtering on Decision Label
   - Decision Fusion
4. Experimental Results
5. Performance Comparison
   - Development
   - Testing
   - State-of-the-art
6. Conclusion
7. Future Research
8. References

Emotion recognition from facial expression

Human face provides an essential, spontaneous channel for the communication of mental states. In addition, facial expressions directly communicate feelings, cognitive mental states, and attitude towards other people.

In the affective computing , various studies only emphasize on **acted or stereotypical facial expression** while analysing the affective state.

Naturalistic expressions however, presents a big challenge, since the dynamic of these expressions is more complex, leading to a larger variability in the way affect is expressed

Therefore, the focus of this paper is to utilize this property in the naturalistic expressions in an efficient way and build a better automatic affective dimension recognition system.

Emotion recognition from facial expression

- An important challenge is to create systems that can continuously (i.e. over time) monitor and classify affective expressions into either discrete affective states or continuous affective dimensions [12].
- The initial approaches treated the videos as sequences of independent facial expression frames and aimed at improving the classification performances for each independent expression at frame level [1]
- Recent years, the research work in affective computing show significant progress, with a great support from naturalistic expression datasets and competitions [2],[3],[1],[4].

Emotion recognition from facial expression

- Meng and Berthouze [5] proposed a multi-stage automatic affective expression recognition system to use HMMs to take into account this temporal relationship and finalize the classification process. The system achieved the best performance on the audio data of AVEC2011 dataset
- Savran et. al. [6] use temporal statistics of texture descriptors extracted from facial videos, a combination of various acoustic features, and lexical features to create regression based affect estimators for each modality.
- At AVEC2014 affect recognition sub-challenge, the temporal relations in naturalistic expressions was used to boost the performance in decision level filtering [7] [8].
- Inspired by [7] and [8], we will investigate how to use this temporal relations in the feature space further. We designed a wavelet transform based digital filtering technique on feature vector to remove their high frequency component and then integrate it in our affective dimension recognition system

Focus of this work

- To build a system that can comprehensively model the variation from naturalistic facial expression and vocal cues.

- To automatically classify the scale of each Arousal, Dominance, and Valence from video and audio database of *AVEC 2014

***All experiments is tested on the fourth international Audio/Visual Emotion Recognition Challenge (AVEC 2014) dataset and compared to other state-of-the art methods in the affect recognition sub-challenge [4]**

Proposed system

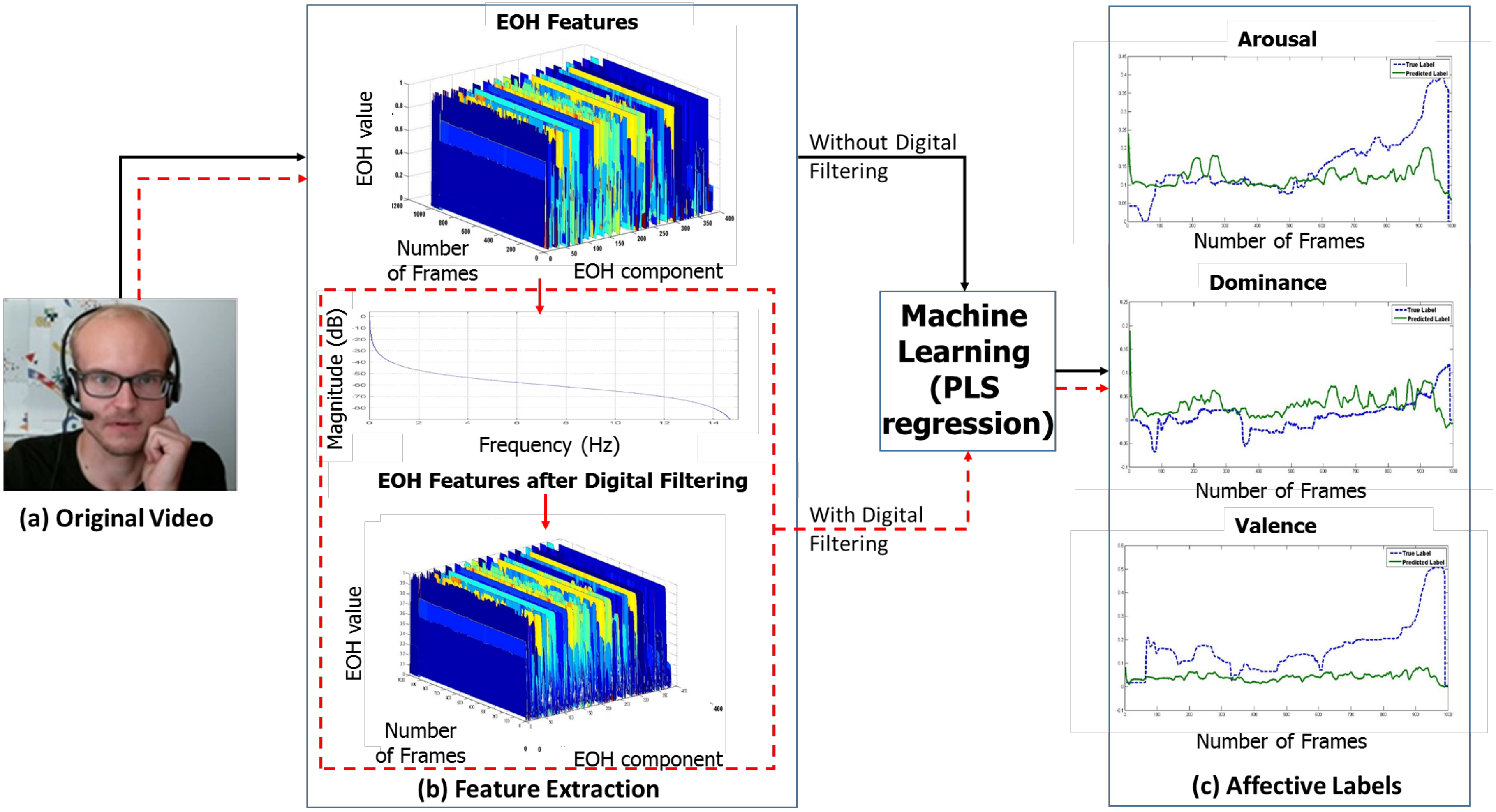
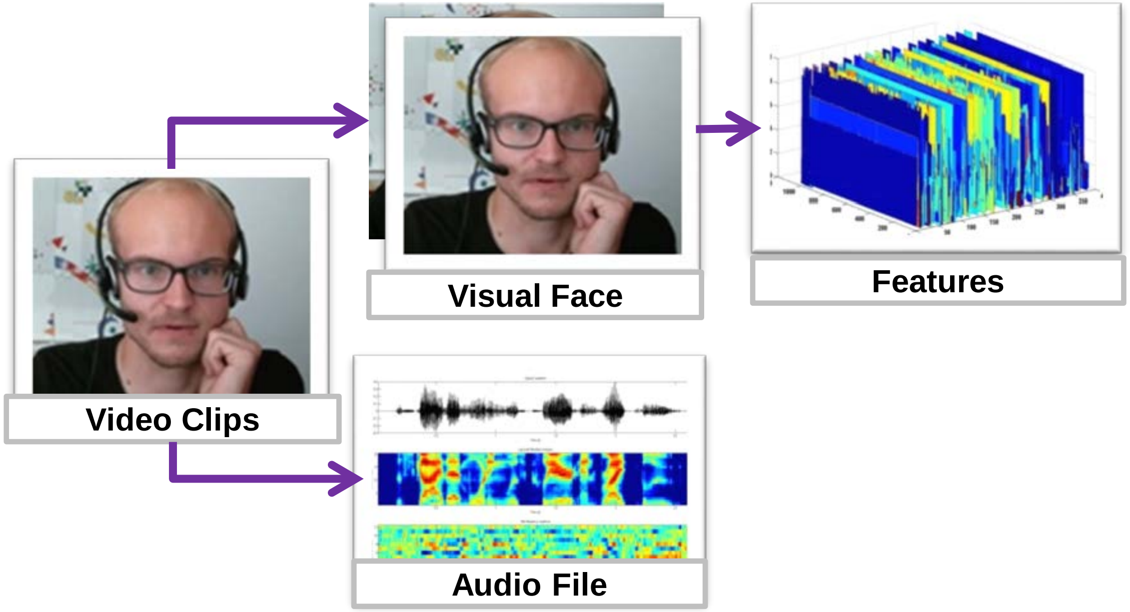


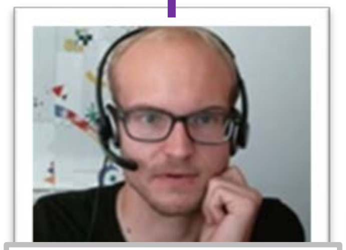
Image Feature Extraction



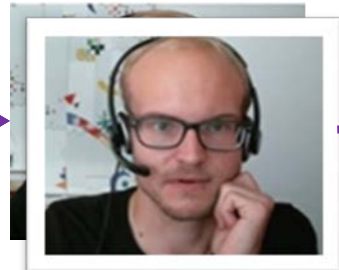
- For each video clips, we deal with the video and audio channel separately.

- For image feature extraction, three dynamic feature are extracted respectively.

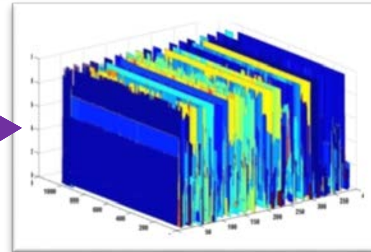
Image Feature Extraction (EOH)



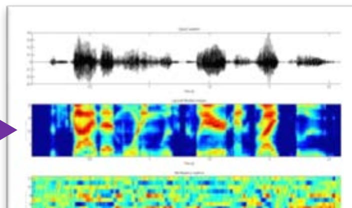
Video Clips



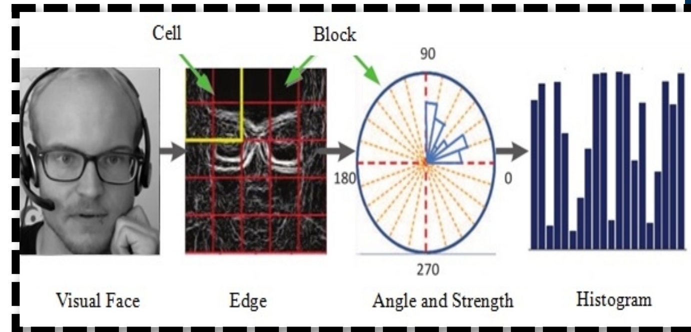
Visual Face



Features



Audio File



EOH Features

- EOH (Edge Orientation Histogram), known as simple version of HOG, captures the edge or the local shape information of an image

- For each video clips, we deal with the video and audio channel separately.

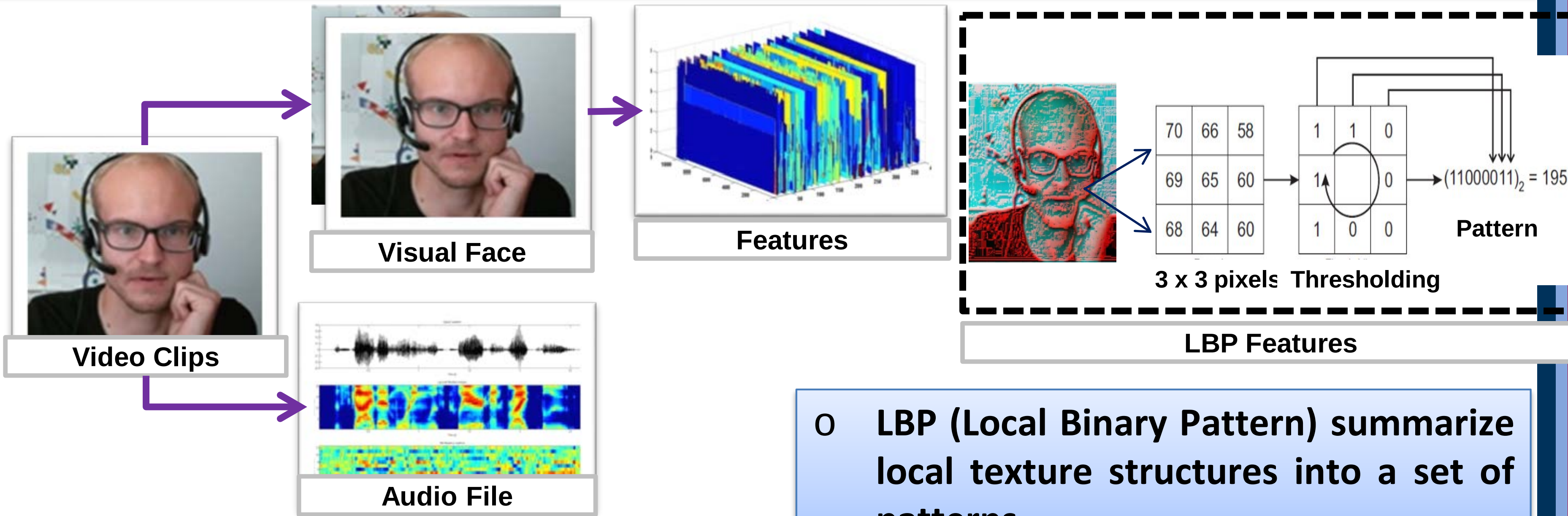
- For image feature extraction, three dynamic feature are extracted respectively.

Edge is captured using Sobel edge detection

Angle and intensity is calculated and arranged into polar coordinate system

Finally histogram of each block is normalized and concatenated into a feature vector

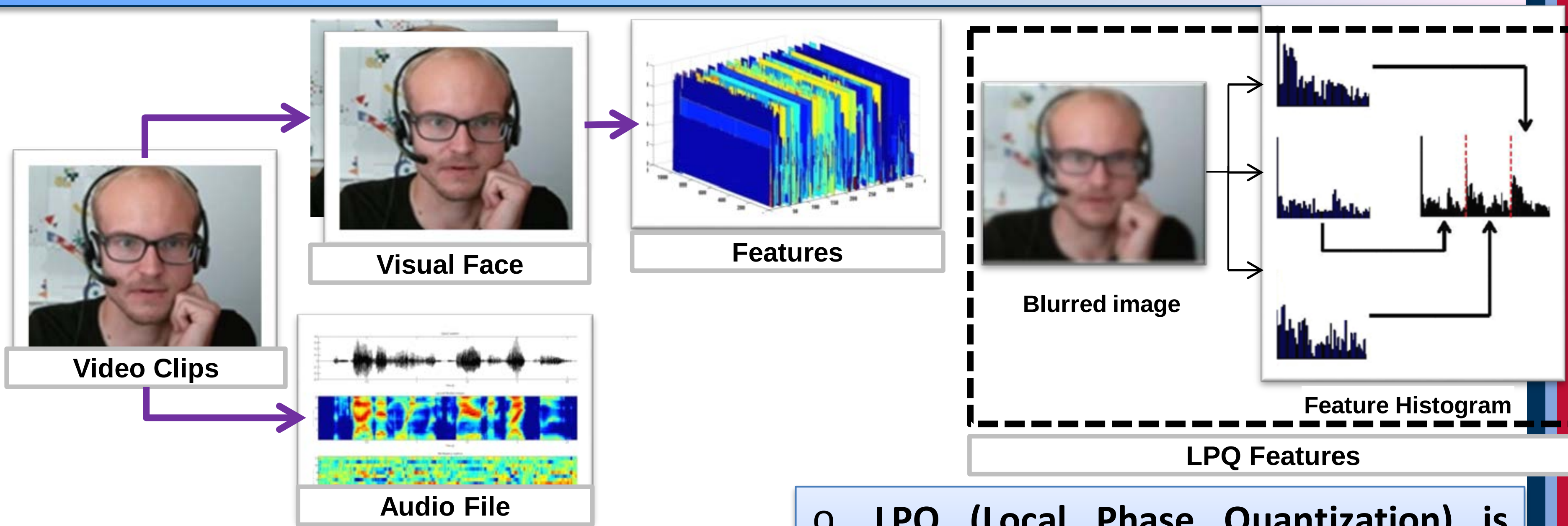
Image Feature Extraction (LBP)



- For each video clips, we deal with the video and audio channel separately.
- For image feature extraction, three dynamic feature are extracted respectively.

- LBP (Local Binary Pattern) summarize local texture structures into a set of patterns
- image pixels by thresholding the 3 x 3 neighborhood with the center value and considering the result as a binary number.

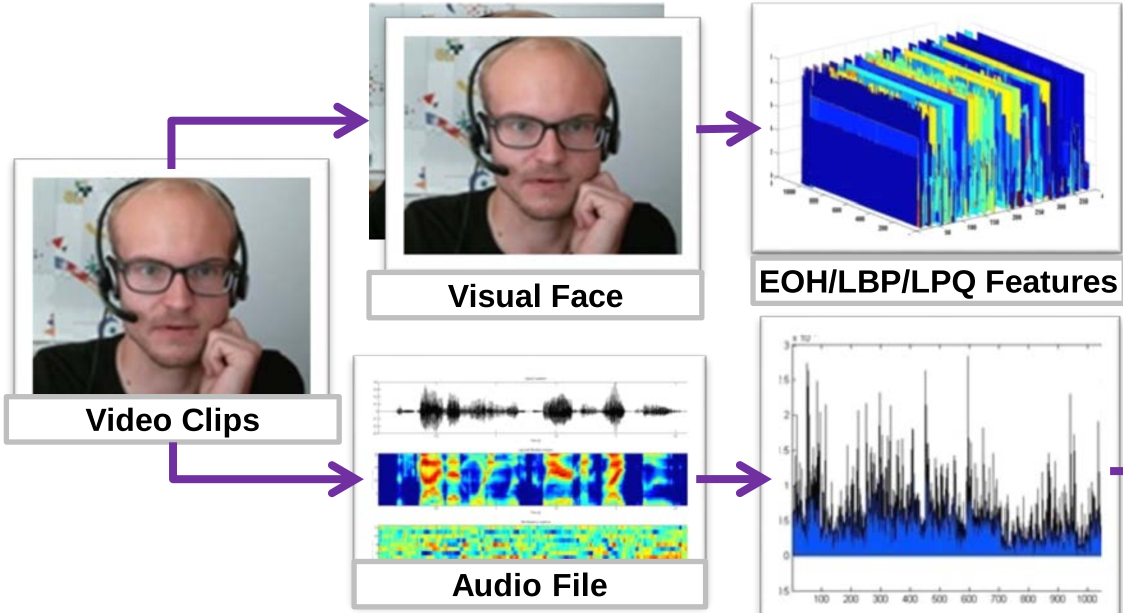
Image Feature Extraction (LPQ)



- For each video clips, we deal with the video and audio channel separately.
- For image feature extraction, three dynamic feature are extracted respectively.

- LPQ (Local Phase Quantization) is based on the blur invariance property of the Fourier phase spectrum
- The LPQ operator is applied to texture identification by computing it locally at every pixel location and presenting the resulting codes as a histogram.

Audio Feature Extraction

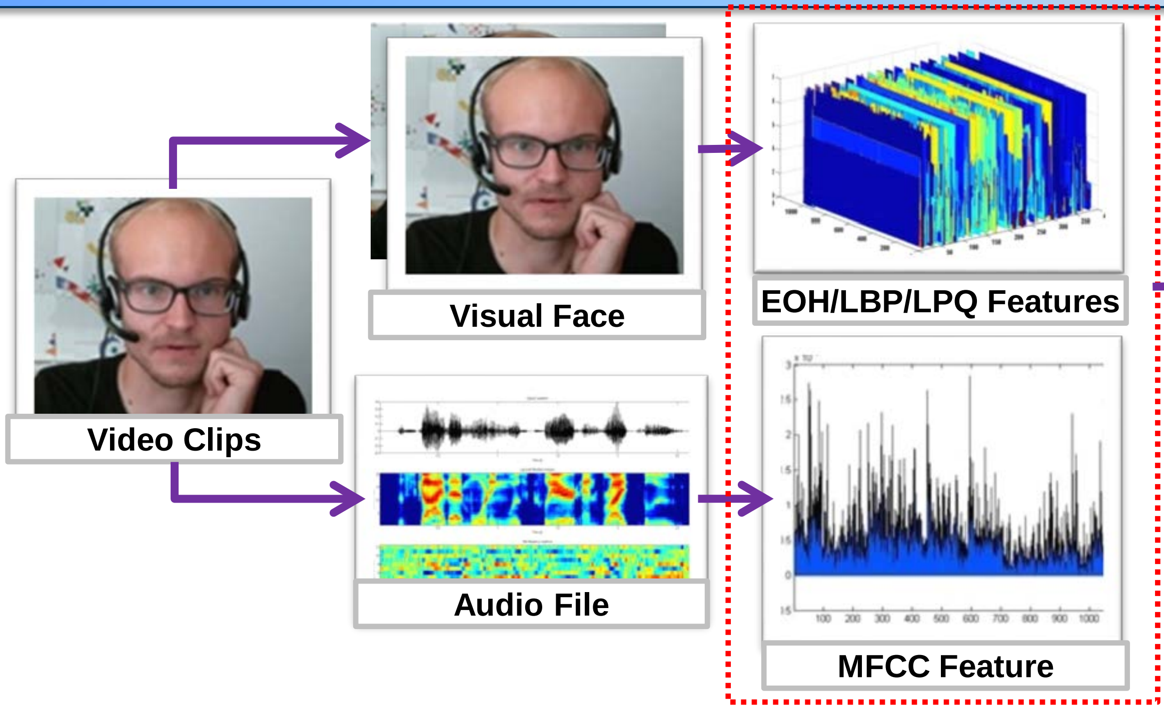


- For each MFCC, we fully utilized ‘long’, ’short’ and ‘valid segmented’ baseline acoustic feature in AVEC 2014 audio database [24]

- For each video clips, we deal with the video and audio channel separately.

- For audio feature extraction, mel-frequency cepstral coefficient (MFCC) are chosen as representation for audio clip

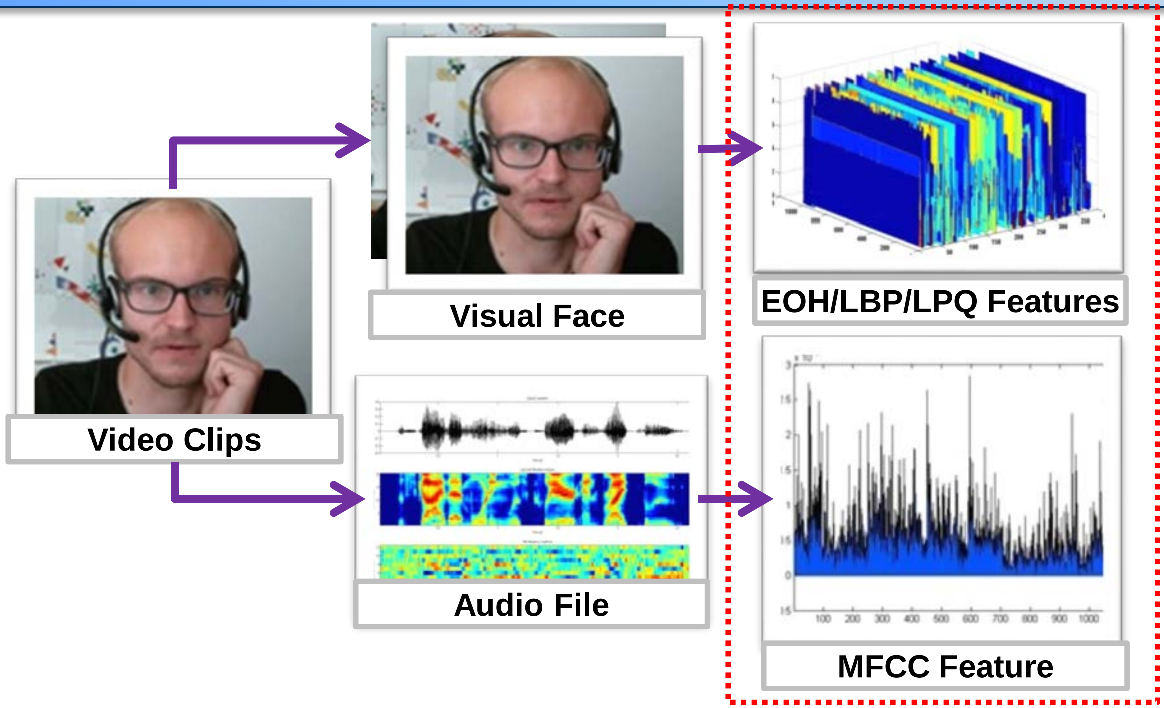
Wavelet Filtering

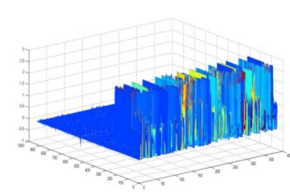


- Haar Wavelet transform is used to remove the irrelevant noise, by compressing high frequency component in the feature space without distortion.

- In this paper, we attempt to design a wavelet transform based digital filtering technique on each features to remove their high frequency component and then integrate it in our affective dimension recognition system

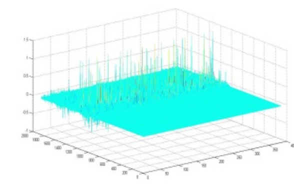
Wavelet Filtering (Haar Wavelet Transform)





Low Frequency Part

$s_k = \frac{x_{2k-1}+x_{2k}}{2},\ k = 1,2,\dots,{}^{N}/_{2}$



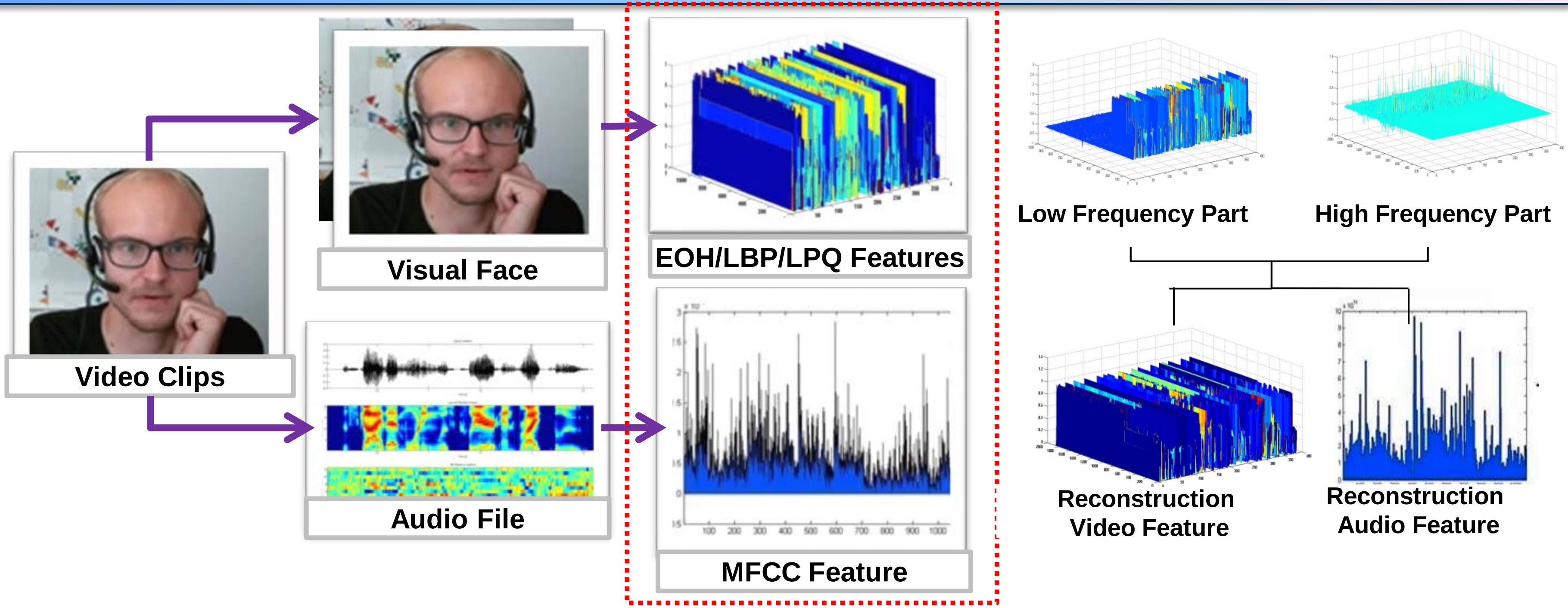
High Frequency Part

$d_k = \frac{x_{2k-1}-x_{2k}}{2},\ k = 1,2,\dots,{}^{N}/_{2}$

- **For a signal , x it can be decomposed into two parts , s and d with the length of *N/2* each based on Haar wavelet transform**

- **s_k is called approximation of the signal which represents the low frequency part of the signal, while d_k is called details of the signal that represents high frequency part of the signal**

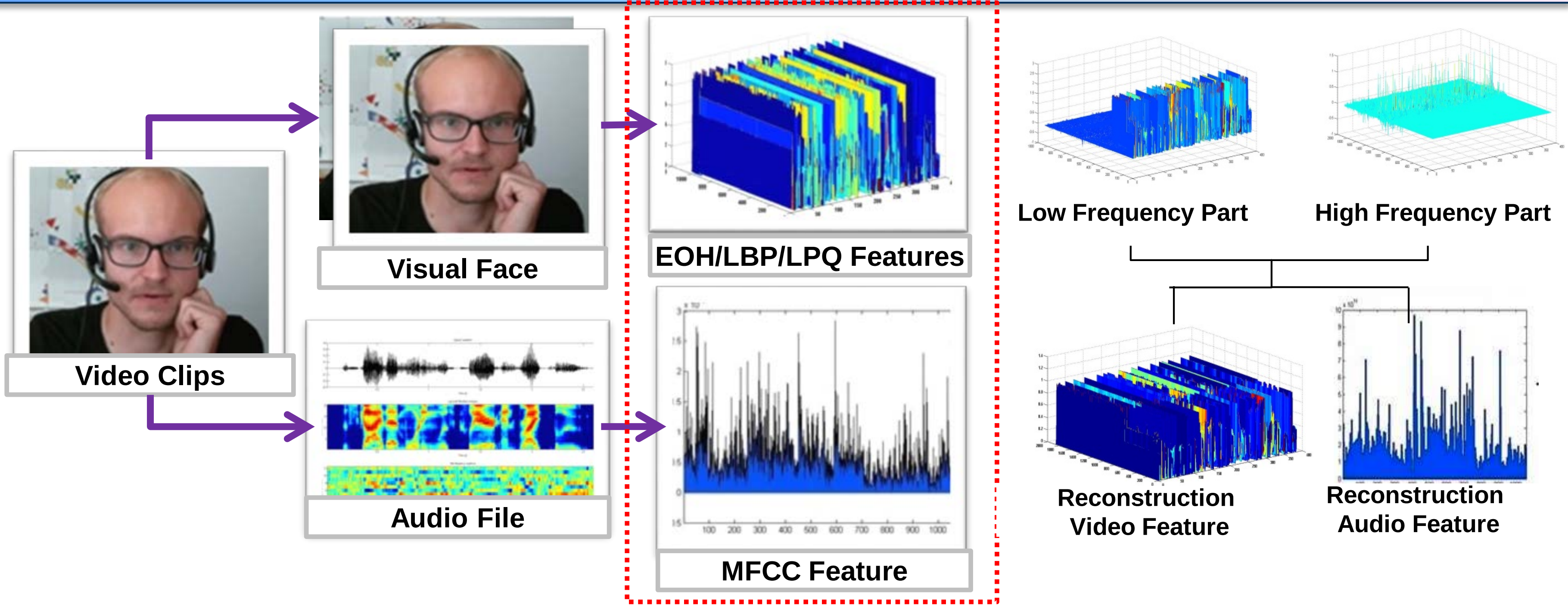
Wavelet Filtering (Haar Wavelet Transform)



- To remove high frequency component , low frequency part s_k will be kept and high frequency part d_k will be replaced by zero

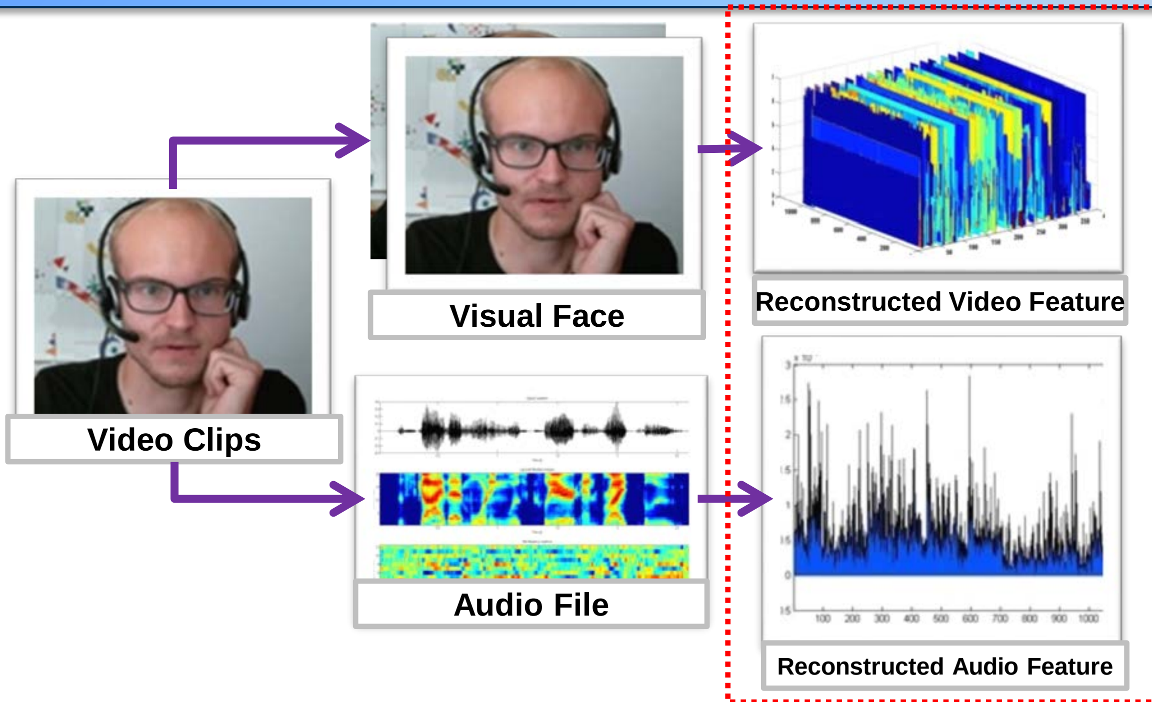
- In this way, the reconstructed signal will lose its high frequency components.

Wavelet Filtering (Haar Wavelet Transform)



- The final reconstructed features will be generated that is smoother along the frame line
- For affective dimension prediction, smooth and simple feature are matching the slow change property of the real affective dimensions

Machine Learning

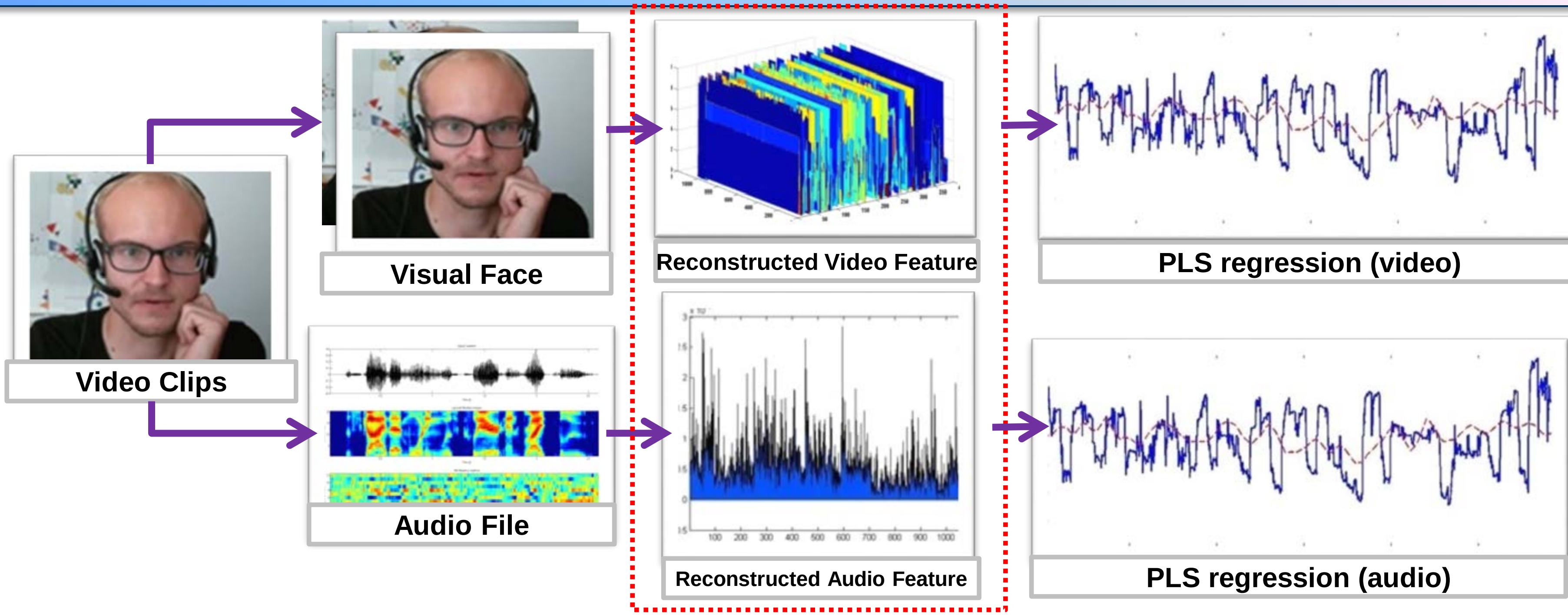


- In affect recognition, the main task was to classify the scale of arousal, dominance and valence from video and audio database of AVEC 2014

- The automatic affective dimension recognition system need to comprehensively model the variation of each video and audio features and automatically predict the scale of each arousal, dominance and valence from video and audio database

- From machine learning point of view, it is a regression problem, not a classification problem, on each individual frame in a image sequence because the predicted value are real numbers.

Machine Learning (PLS Regression)

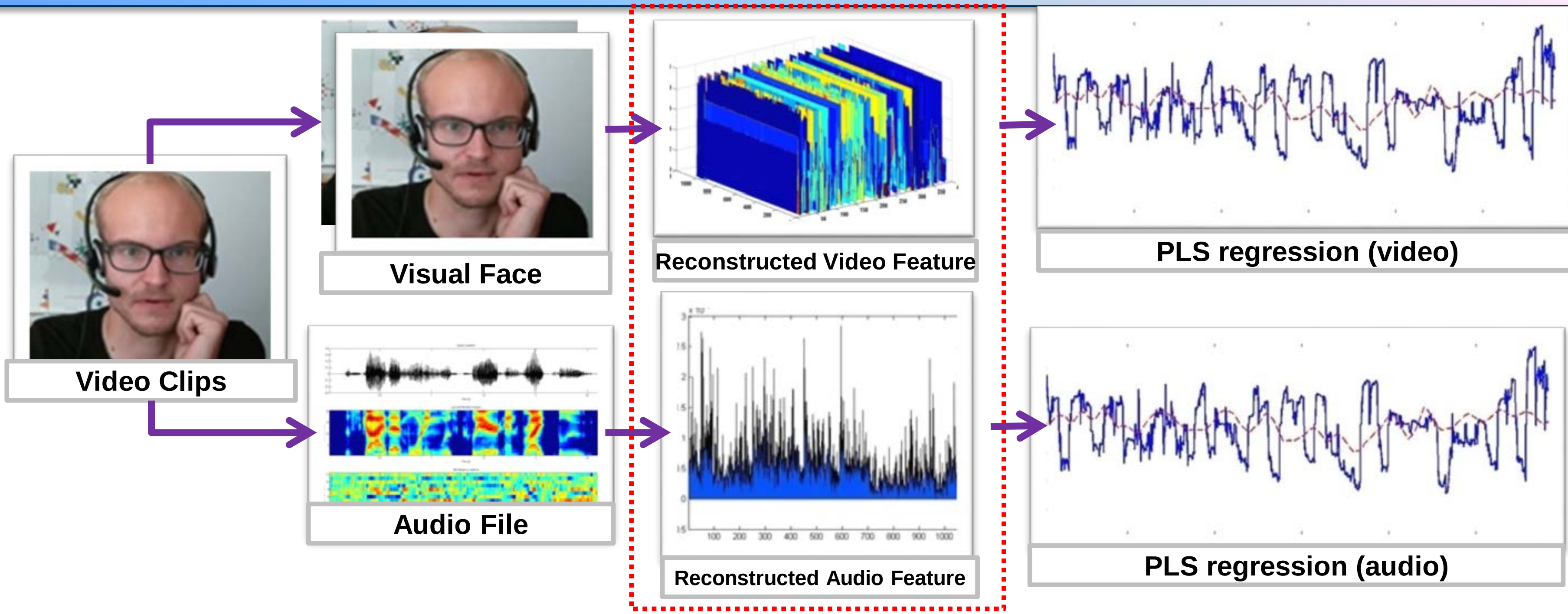


- Partial Least Square (PLS) regression is a statistical algorithm that bears some relation to principal component regression.

- After performing PLS regression for training and testing data, it will give prediction label as the output.

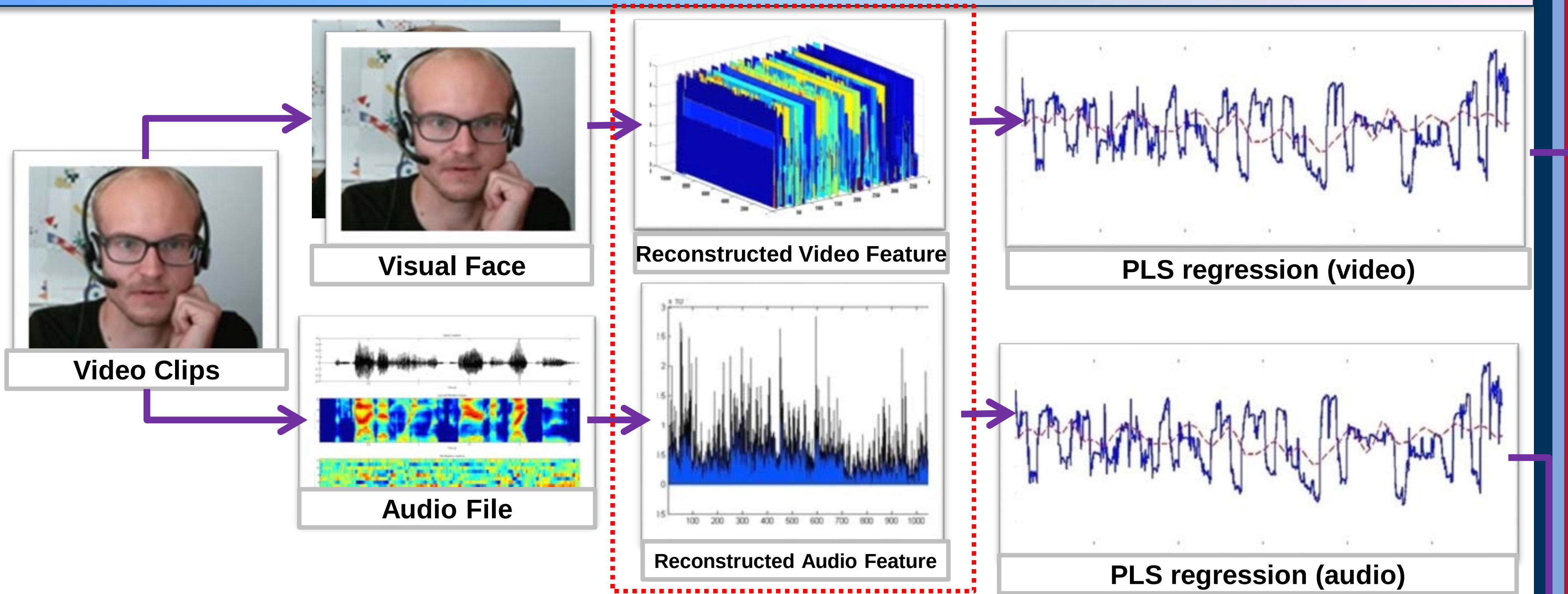
- PLS is being employed towards each features, by building a linear regression model by projecting the response and independent variable to another space

Filtering on Decision Label

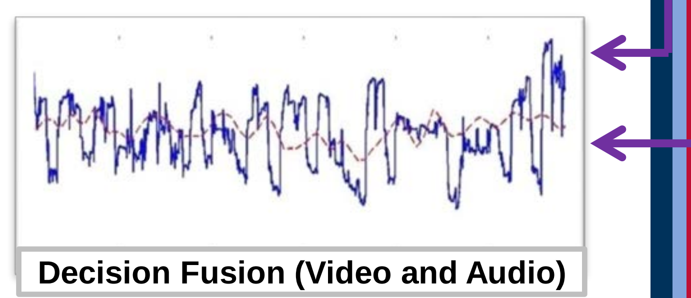


- Since AVEC 2014 requires the prediction of continuous affect labels per frame, we carry out smoothing over the prediction labels using simple low pass filtering. Low pass filtering is carried out on prediction of each development and testing frames to further enhance the results.

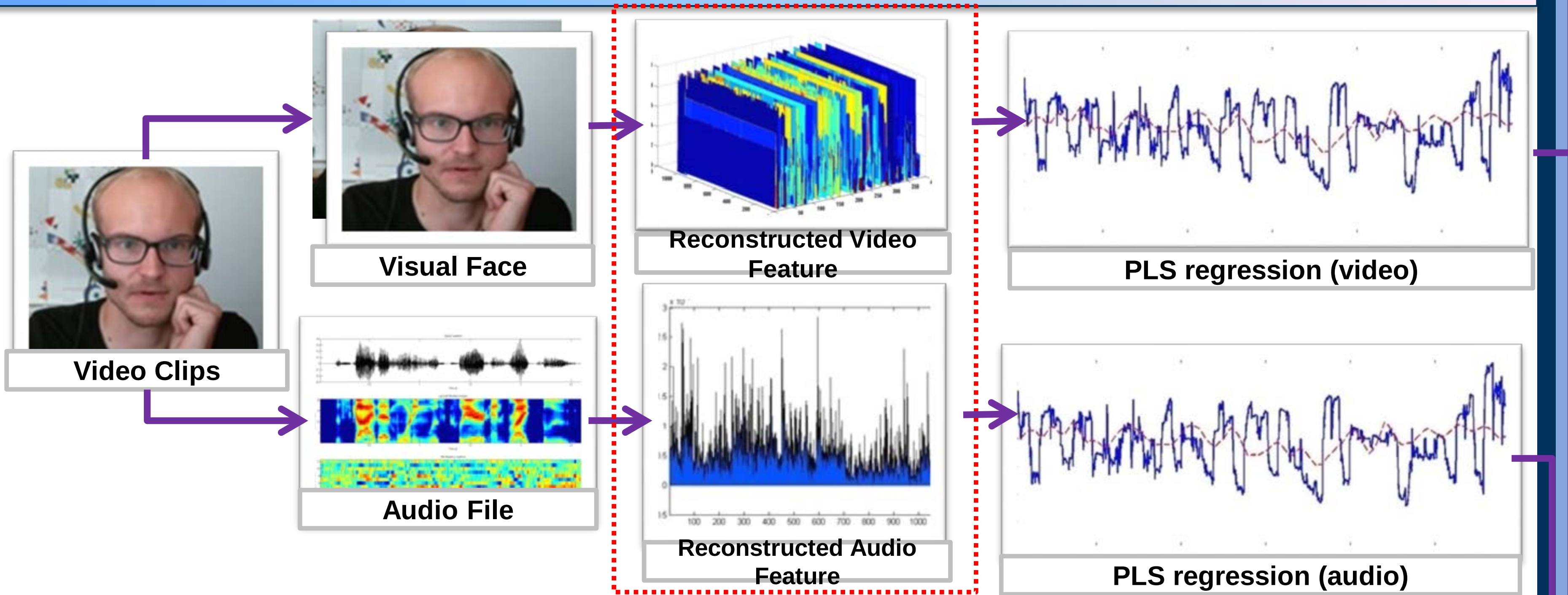
Decision Fusion



- Decision Fusion stage aims to combine multiple decision (video and audio) into a single and consensus one.
- A weighted sum rule is defined to combine the predicted values from each video and audio

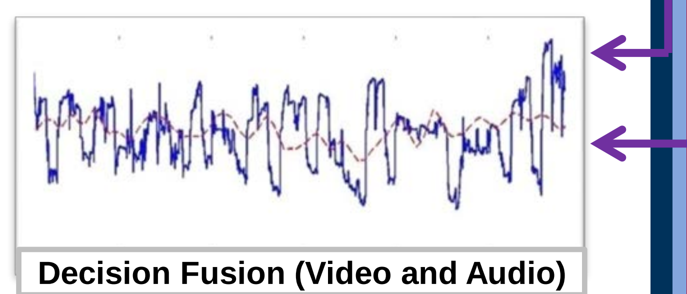


Decision Fusion



- $D_{linear}(\hat{x}) = \sum_{i=1}^{K} \alpha(i) D_i(\hat{x})$

where $\hat{x}$ is a testing sample and $D_i(\hat{x})$ is its i_{th}decision value $(i = 1, 2, \ldots K)$ while $alpha(i)$ is its corresponding weight which should satisfy $\sum_{i=1}^{K} \alpha(i) = 1$



Experimental Results

- Affect recognition challenges concentrates fully on continuous affect recognition of the dimensions of **Arousal, Dominance** and **Valence**
- The label of each dimensions has to be predicted for each frame of the recording.

Evaluates the training sets on the development sets provided by AVEC2014

Evaluation is based on Pearson's correlation coefficient (CORR) and Root Mean Square (RMSE)

Evaluate the training sets on the testing sets provided by AVEC2014

Performance Comparison (Development)

- Table 1 shows the performance of video in different features on each affective dimensions in terms of CORR and RMSE in development datasets.
- Table 2 shows the performance of baseline results taken from AVEC 2014 paper

✓ For each of the feature (EOH, LBP & LPQ) proposed, all the method outperform the baseline results in every dimensions

Table 2: Performance for Video Baseline in Development Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	LGBP_TOP	0.412	-
Dominance		0.319	-
Valence		0.355	-

Table 1: Performance for Video in Development Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	EOH	0.5669	0.0879
	LBP	**0.5868**	0.0956
	LPQ	0.5624	0.0987
Dominance	EOH	0.6021	0.1026
	LBP	0.5135	0.1049
	LPQ	**0.6023**	0.1015
Valence	EOH	**0.5523**	0.0670
	LBP	0.5480	0.0706
	LPQ	0.5211	0.0665

Performance Comparison (Development)

- Table 3 shows the performance of audio in different features on each affective dimensions in terms of CORR and RMSE in development datasets.
- Table 4 shows the performance of baseline results taken from AVEC 2014 paper

✓ For each of the feature (long, short, valid _segmented) proposed, all the method outperform the baseline results in every dimensions

Table 4: Performance for Audio Baseline in Development Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	LLDs+MFCC	0.517	-
Dominance		0.439	-
Valence		0.347	-

Table 3: Performance for Audio in Development Sets

Affect Dimension	Feature	*Audio*	
		CORR	RMSE
Arousal	Long	**0.6136**	0.0992
	Short	0.5911	0.0981
	Vad_seg	0.5954	0.1002
Dominance	Long	0.5866	0.0989
	Short	0.5902	0.0988
	Vad_seg	**0.6054**	0.0987
Valence	Long	0.5773	0.0659
	Short	0.5509	0.0659
	Vad_seg	**0.5798**	0.0661

Performance Comparison (Development)

- Table 5 shows the performance of Fusion (video+audio) in different features on each affective dimensions in terms of CORR and RMSE in development datasets.
- Table 6 shows the performance of baseline results taken from AVEC 2014 paper

✓ We making an attempt to fuse the best performance in audio and video for each affective dimensions, referring from Table 1 and 3

✓ For the video and audio fusion, there is no significant difference in terms of performance. It is because we only use simple fusion rule.

✓ In comparison with baseline results , the proposed method outperform in every modality and dimensions

Table 6: Performance for Fusion (Video+Audio) Baseline in Development Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	LGBP_TOP+ LLDs+MFCC	0.421	-
Dominance		0.348	-
Valence		0.236	-

Table 5: Performance for Fusion (Video+Audio) in Development Sets

Affect Dimension	Feature	*Fusion*	
		CORR	RMSE
Arousal	EOH_Long	0.5668	0.0894
	LBP_Long	**0.5873**	0.0944
	LPQ_Long	0.5165	0.0955
Dominance	EOH_Vad_seg	**0.6021**	0.0988
	LBP_Vad_seg	0.5891	0.1005
	LPQ_Vad_seg	0.5788	0.1011
Valence	EOH_Vad_seg	**0.5525**	0.0654
	LBP_Vad_seg	0.5479	0.0676
	LPQ_Vad_Seg	0.5199	0.0660

Performance Comparison (Testing)

- o Table 7 shows the performance of video in different features on each affective dimensions in terms of CORR and RMSE in testing datasets.
- o Table 8 shows the performance of baseline results taken from AVEC 2014 paper

- ✓ All parameters, such as filter window size, number of componenet in PLS are identical to the previous set experiments on the development sets.
- ✓ For each of the feature (EOH, LBP & LPQ) proposed, all the method outperform the baseline test results in every dimensions

Table 8: Performance for Video Baseline in Testing Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	LGBP_TOP	0.2062	-
Dominance		0.1959	-
Valence		0.1879	-

Table 7: Performance for Video in Testing Sets

Affect Dimension	Feature	*Video*	
		CORR	RMSE
Arousal	EOH	0.5713	0.0921
	LBP	**0.5597**	0.0961
	LPQ	0.5711	0.1017
Dominance	EOH	0.4916	0.1009
	LBP	0.5179	0.0597
	LPQ	**0.4835**	0.0993
Valence	EOH	**0.5032**	0.0570
	LBP	0.5183	0.0597
	LPQ	0.5319	0.0560

Performance Comparison (Testing)

- Table 9 shows the performance of Audio in different features on each affective dimensions in terms of CORR and RMSE in testing datasets.
- Table 10 shows the performance of baseline results taken from AVEC 2014 paper

- ✓ All parameters, such as filter window size, number of components in PLS are identical to the previous set experiments on the development sets.
- ✓ For Audio, only Arousal dimensions of baseline results beat our method. For Dominance and Valence results, each of our method outperformed baseline results.

Table 10: Performance for Audio Baseline in Testing Sets

Affect Dimension	Feature	*Audio*	
		CORR	RMSE
Arousal	LLDs+MFCC	**0.540**	-
Dominance		0.360	-
Valence		0.355	-

Table 9: Performance for Audio in Testing Sets

Affect Dimension	Feature	*Audio*	
		CORR	RMSE
Arousal	Long	0.5277	0.0951
	Short	0.4913	0.0954
	Vad_seg	0.5081	0.0953
Dominance	Long	0.4750	0.0907
	Short	0.4892	0.1797
	Vad_seg	**0.4913**	0.0901
Valence	Long	0.4987	0.0552
	Short	0.4469	0.0553
	Vad_seg	**0.5355**	0.0548

Performance Comparison (Testing)

- Table 11 shows the performance of Audio in different features on each affective dimensions in terms of CORR and RMSE in testing datasets.
- Table 12 shows the performance of baseline results taken from AVEC 2014 paper

- ✓ All parameters, such as filter window size, number of components in PLS are identical to the previous set experiments on the development sets.
- ✓ In comparison with baseline results , the proposed method outperform in every modality and dimensions

Table 12: Performance for Fusion (Video+Audio) Baseline in Testing Sets

Affect Dimension	Feature	*Audio*	
		CORR	RMSE
Arousal	LGBP_TOP+ LLDs+MFCC	0.478	-
Dominance		0.324	-
Valence		0.282	-

Table 11: Performance for Fusion (Video+Audio) in Testing Sets

Affect Dimension	Feature	*Audio*	
		CORR	RMSE
Arousal	EOH_Long	0.5721	0.0950
	LBP_Long	0.5586	0.0935
	LPQ_Long	**0.5760**	0.0968
Dominance	EOH_Vad_seg	0.4913	0.0953
	LBP_Vad_seg	**0.5182**	0.0915
	LPQ_Vad_seg	0.4842	0.0945
Valence	EOH_Vad_seg	0.5030	0.0542
	LBP_Vad_seg	0.5184	0.0559
	LPQ_Vad_Seg	**0.5354**	0.0549

Performance Comparison (state-of-the-art)

The system was trained on the AVEC2014 training set and tested on both development and testing set in comparison with baseline method.

It was also compared with all state-of-the-art methods in the AVEC2014 affect recognition with fairly good performance

Table 13: Performance Comparison with State-of-the-art methods in AVEC2014

Team	Method	CORR	RMSE
Baseline [4]	SVR+Fusion	0.4185	0.2090
Ulm [7]	Subjects+Label Inference	**0.5946**	0.1009
NLPR [9]	Deep Learning+Fusion	0.5499	0.1630
SAIL [10]	Fusion+Temporal Regression	0.5219	0.0831
BU-CMPE [11]	CCA ensemble	0.3932	0.0928
Our method	Wavelet Filtering+PLS+Fusion	0.5432	**0.0810**

NLPR [9], Ulm [7], SAIL [10], and our method achieved better performance than baseline since these four methods use temporal relation in decision label.

However only NLPR [9] and our method took one step further that is investigating temporal relation in feature level;NLPR [9] did it by temporal pooling function in neural network, while ours use wavelet filtering in each EOH, LBP and LPQ features.

Conclusion

- ✓ **Research Contribution**
  - ✓ In this paper, an automatic affective dimension recognition system is proposed based on wavelet filtering and PLS regression for naturalistic facial expression.
  - ✓ Instead of using temporal relation in decision label, Haar Wavelet transform based digital filtering method was used to remove any irrelevant noise in the feature space
  - ✓ The reconstructed features were input to PLS regression and final fusion process was used for combining video and audio modality.

Conclusion

- ✓ **Future research**
  - ✓ **The performance of the proposed system can be enhanced by improving the fusion rule on video and audio modalities.**
  - ✓ **Other wavelet transform filters can be used to compare the performance with Haar Wavelet filters**
  - ✓ **The proposed method can be tested on other naturalistic expressions datasets.**

References

[1]. M. F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge. In *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop, 2013*.

[2]. B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011: The first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2011), 2011*.

[3]. B. Schuller, M. F. Valstar, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge - an introduction. *In ICMI*, pages 361–362, 2012.

[4]. M. F. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 - 3d dimensional affect and depression recognition challenge. In *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop, 2014*.

References

[5]. H. Meng and N. Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics*, 44(3):315–328, 2014.

[6]. A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Com- bining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12*, pages 485–492, New York, NY, USA, 2012. ACM.

[7]. M. Kachele, M. Schels, and F. Schwenker. Inferring depression and affect from application dependent meta knowledge. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 41–48, New York, NY, USA, 2014. ACM.

[8]. R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 33–40, New York, NY, USA, 2014. ACM.

References

[9]. N. Ueda. Optimal linear combination of neural networks for improving classification performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):207–215, 2000.

[10]. L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Multi-scale temporal modeling for dimensional emotion recognition in video. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC ’14, pages 11–18, New York, NY, USA, 2014. ACM.

[11]. H. Kaya, F. Cilli, and A. A. Salah. Ensemble CCA for continuous emotion prediction. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC ’14*, pages 19–26, New York, NY, USA, 2014. ACM.

[12]. E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99)*:1– 1, 2014.

THANK YOU