Exploring Multimodal Visual Features for Continuous Affect Recognition

Bo Sun
College of Information Science and Technology,
Beijing Normal University
tosunbo@bnu.edu.cn.

Siming Cao
College of Information Science and Technology,
Beijing Normal University
caosiming@mail.bnu.edu.cn

Liandong Li
College of Information Science and Technology,
Beijing Normal University
bnulee@hotmail.com

Jun He
College of Information Science and Technology,
Beijing Normal University
hejun@bnu.edu.cn

Lejun Yu
College of Information Science and Technology,
Beijing Normal University
yulejun@bnu.edu.cn

ABSTRACT

This paper presents our work in the Emotion Sub-Challenge of the 6th Audio/Visual Emotion Challenge and Workshop (AVEC 2016), whose goal is to explore utilizing audio, visual and physiological signals to continuously predict the value of the emotion dimensions (arousal and valence). As visual features are very important in emotion recognition, we try a variety of handcrafted and deep visual features. For each video clip, besides the baseline features, we extract multi-scale Dense SIFT features (MSDF), and some types of Convolutional neural networks (CNNs) features to recognize the expression phases of the current frame. We train linear Support Vector Regression (SVR) for every kind of features on the RECOLA dataset. Multimodal fusion of these modalities is then performed with a multiple linear regression model. The final Concordance Correlation Coefficient (CCC) we gained on the development set are 0.824 for arousal, and 0.718 for valence; and on the test set are 0.683 for arousal and 0.642 for valence.

Keywords

Continuous Emotion Recognition; CNN; Multimodal Features; SVR; Residual Network

1. INTRODUCTION

Emotion recognition is an important subject in the field of pattern recognition, which is of great interest for human-computer interaction. According to theories in psychology research there are two major emotion computing models [1]: discrete theory and dimensional theory. Discrete theory describes an emotion state as discrete labels such as "surprise", "sad", "happy" etc. The leading study of Ekman and Frisen [2] formed the basis of visual automatic facial expression recognition. Their studies suggested that anger, disgust, fear, happiness, sadness, and surprise are the six basic prototypical facial expressions. This work cannot meet the needs of real life, because it cannot express complex affective states. While dimensional theory considers an emotion state as a point in a continuous space. Russell proposed that each of the basic emotions is a bipolar entity as part of the same emotional continuum. The proposed polar includes arousal and valence [3]. In recent years, recognition of non-acted spontaneous emotions in the continuous dimensional space has attracted researchers' interest. Because it is more suitable to express and understand our complex emotions [4, 5].

AVEC'16, October 16, 2016, Amsterdam, The Netherlands.
 ISBN 978-1-4503-4516-3/16/10…$15.00.
DOI: http://dx.doi.org/10.1145/2988257.2988270

Psychologists show that humans recognize affective states from several modalities, such as facial expression, body gesture, voice, etc [6]. Some researchers advocate that combined multiple modalities will contribute to the recognition accuracy. While most of the existing relative research has focused on single modality, especially from facial expression. To accelerate research in automatic continuous affect recognition from audio, video and physiological data, the Audio/Visual Emotion Challenge and Workshop (AVEC) aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and physiological emotion analysis. The database used for this challenge is RECOLA [8], a multimodal corpus of spontaneous collaborative and affective interactions. [7]

In this paper, we describe our work in the AVEC 2016 challenge. We mainly focus on dimensional emotion recognition from audio, visual and physiology modalities. The Support Vector Regression (SVR) is used for regression prediction. Researches show that video features are more important for emotion recognition [31], so we try a variety of handcrafted and deep visual features. For each video clip, besides the baseline features, we extract multi-scale Dense SIFT features (MSDF), and some types of Convolutional neural networks (CNNs) features, including the deep residual network [27] to recognize the expression phases of the current frame.

The remainder of this work is organized as follows. Section 2 introduces related works in dimensional emotion recognition. Section 3 describes the dataset and features used in AVEC2016 challenge. Section 4 describes the details of the overall methodology chosen for the AVEC 2016 challenge and Section 5 describes the entire experiments we have done and our extensive experimental results. Finally, the conclusion is given in Section 6.

2. Related works

In recent years, recognition of non-acted spontaneous emotions in the continuous dimensional space has attracted researchers' interest. Dimensional theories often argue that discrete emotion categories (such as anger) do not have a specific biological basis, although many promising recognition results have achieved. That is based on the fact that there is no brain region or circuit that is unique to that emotion category [4,5]. Computational models that build on dimensional theories often use the "VAD" theory. VAD model has 3 dimensions: valence, arousal and dominance. V stands for valence, represents the positive and negative characteristics of the individual emotional state; A represents the degree of activation, indicating the individual's physiological activation level; D stands for dominance, which indicates the individual's control of the situation and others. The numerical range of each dimension is -1 to 1, -1 is the lowest value in each dimension, and +1 is the highest value in each dimension.

Recently, deep learning methods have become very popular within the community of computer vision. Studies shows that AlexNet and other CNN models has shown obvious performance for emotion recognition. [32,33] Besides, the Bag of Words (BoW) model based on MSDF has also been widely used in computer vision in recent years. Sikka, K. et al. [12] explored bag of words architectures in the facial expression domain, which has shown remarkable performance on facial recognition.

In the Audio/ Visual Emotion Challenge 2015(AVEC2015), most teams used Long Short-Term Memory (LSTM), which is a recurrent neural network (RNN) architecture (an artificial neural network) proposed in 1997 by Sepp Hochreiter and Jürgen Schmidhuber [19]. Martin Wöllmer et al. [20] proposed a fully automatic audiovisual recognition approach based on LSTM modeling of word-level audio and video features. It has demonstrated that long range context modeling tends to increase accuracies of emotion recognition. Besides LSTM, SVR is also widely used in the AVEC challenges [7].

3. DATASET AND FEATURES

3.1 Dataset

The AVEC 2016 challenge is evaluated on a subset of the RECOLA dataset [8], which was recorded to study socio affective behaviors from multimodal data in the context of remote collaborative work, for the development of computer-mediated communication tools [9]. Spontaneous and naturalistic interactions were collected during they are solving a collaborative task through video conference. Multimodal signals like audio, video, electro-cardiogram (ECG), electro-dermal activity (EDA), heart rate (HR) and its measure of variability (HRV), skin conductance response (SCR) and skin conductance level (SCL), were synchronously recorded from 27 French-speaking subjects. There is only one person in every recording. Emotional dimensions (arousal and valence) are annotated by 6 French speakers in scale [-1, 1] for every 40ms. Gold standard is calculated using a specific normalization technique as reported in [8]. Finally, by stratifying (balancing) on gender and mother tongue, cf. the dataset is equally split into three partitions: train, development and test, with each partition containing 9 different speakers.

3.2 Audio Features

In AVEC 2016 challenge, the 88 baseline audio features computed with openSMILE [25] and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21] configuration file, with a sliding centred window which size depends on the modality. The arithmetic mean and the coefficient of variation are computed on all 42 LLD, which cover the spectral, cepstral, prosodic and voice quality information.

3.3 Visual Features

3.3.1 Face Detection and Alignment

Face detection is a very important step in the whole pipeline, which will directly affect the effectiveness of the visual feature extraction. In order to extract extra video features, like MSDF, and CNN features more accurately, first we follow the face extraction and tracking method of Sikka et al. [12] and Dhall et al. [22]. A mixture of tree structured part model [23] face detector is used to detect the position of face in the first frame of a video. Then use the Intraface toolkit supervised descent method [24] to track facial landmarks of the rest frames in a Parameterized Appearance Model. Finally, 49 landmark points can be used to align faces for expression classification. Through experiments, the first base point is the position of the middle of two eyes; the second one is the position of the central point of mouth. All frames are aligned to this base face through affine transformation and cut to 200 × 200 pixels. There are some frames that have not detected a human face but have a face. To get more face, we follow the face extraction method of Zhu, X., & Ramanan, D. [25]. After the two step of face detection, most of the human face are obtained. For the frames are mis-tracked, all feature sets are interpolated by a piecewise cubic Hermite polynomial to cope with mis-tracked frames.

3.3.2 LGBP-TOP

The local Gabor Binary Patterns [10] is a type of descriptor that is robust to illumination changes and misalignment. It first takes a video frame convolved with a number of Gabor filters. It is followed by the LBP feature extraction through the set of Gabor magnitude response images. The resulting binary patterns are histogrammed and concatenated into a single feature histogram. Like other volume local texture features, in our sense, a video is block-wised to 4×4 to from XY plane. The LGBP-TOP feature with those three spatial frequencies and six Gaussian orientations were then extracted, and a length of 18×4×4×59×3 = 50,976 feature is available. For AVEC 2016 challenge, the 168 reduced dimension LGBP-TOP features are used as appearance visual features. This appearance visual features are obtained by a Principal Component Analysis (PCA) from the 50,976 LGBP-TOP features (99% of variance).

3.3.3 Geometric Feature

The aligned 49 landmarks tracked by the Intraface toolkit are spilt into three regions: left eye (6-10, 26-31), right eye (1-5, 20-25) and mouth (32-49). For every region, we compute the angels between three points and distances between two points. Then the positions of 49 points of this frame and the frame before are concatenated into the vector. At last, the distance of 49 landmarks to the mean central facial position is added to the geometric feature. The feature vector has a length of 71+98×2+49=316 at last.

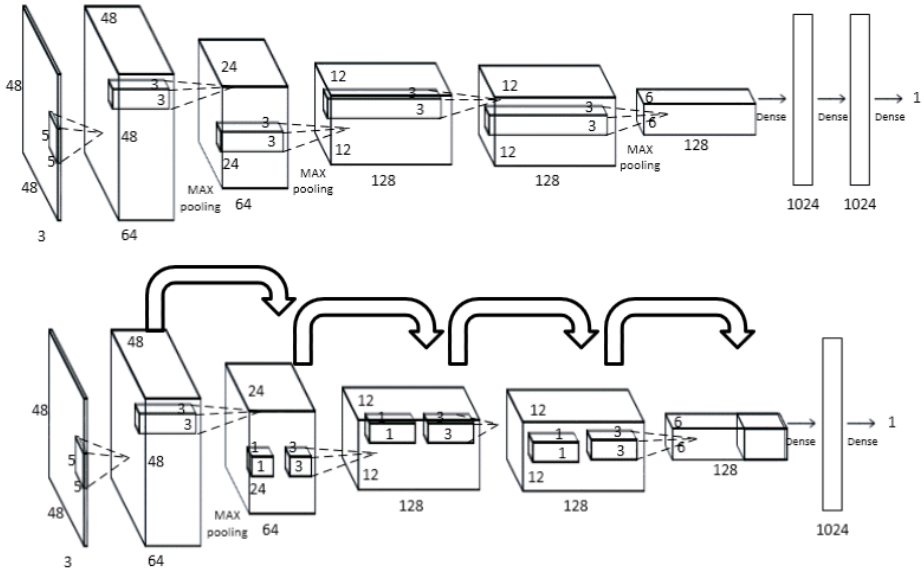

Fig. 1 The architectures of our proposed CNN and ResNet

3.3.4 Multi-scale Dense SIFT

We use the Bag of Words (BoW) model based on multi-scale Dense SIFT features (MSDF) [11] to extract visual features, which has been widely used in computer vision in recent years. Sikka, K. et al. [12] explored bag of words architectures in the facial expression domain, which has shown remarkable performance on facial recognition.

Firstly, multi-scale Dense SIFT [13] features are extracted by setting the width of the SIFT spatial bins to 4, 8, 12 and 16 pixels.

Secondly, we use K-means algorithm to cluster features extracted in the previous step to construct the dictionary. Through experiments all the features are clustered to 800 clustering centers as code-words for the dictionary.

Thirdly, we use Locality-constrained Linear Coding (LLC) [14] to encode the code-words, which can guarantee the sparse and locality of the coded words. Then we use spatial pyramid (SPM) [15] to get spatial information, the layer of SPM was set to 5. In total the geometric set includes 68000 features. We obtained 1458 dimension features by a Principal Component Analysis (PCA) from the 68000 feature (90% of variance).

3.3.5 Deep Visual Features

Inspired by the AlexNet [26] and the Residual Network [27], we design two architectures for emotion recognition.

The AlexNet is a 9-layers deep model, which is designed for ILSVRC-2012 Challenge [28]. The activation function of it is reflected linear unit (ReLU). AlexNet model has 5 convolutional layers and 3 fully connection lays. To avoid over-fitting, the AlexNet uses data enlarge strategy, local normalization and dropout method. In total the geometric set includes 9216 features. We obtained 266 dimension features by a Principal Component Analysis (PCA) from the 9216 feature (95% of variance).

Our proposed CNN is based on the AlexNet. The first convolutional layer filters the input patch with 64 kernels of size 5×5. The second convolutional layer takes as input the response-normalized and max-pooled output of the first convolutional layer and filters it with 64 kernels of size 3×3×64. The third convolutional layer has 128 kernels of size 3×3×64 connected to the (normalized, pooled) outputs of the second convolutional layer. The fourth and fifth convolutional layer both have 128 kernels of size 3×3×128. The third, fourth, and fifth convolutional layers are connected to one another without any pooling or normalization layers. The fully-connected (FC) layers have 1024 neurons each. The ReLU Activations are applied to the output of every convolutional and fully-connected layer. For last layer's output is used as the regression value of the whole network. For feature extraction, we use the last pooling layer as the output, which has 4608 output dimensions. We obtained 205 dimension features by a Principal Component Analysis (PCA) from the 4608 feature (95% of variance).

The whole architecture of our model is shown in Fig. 1. As we use FER dataset [29] for pre-training, all input images are resized to 48×48.

He et.al [27] present the residual learning framework to ease the training of networks that are substantially deeper than those used previously. They explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. They show that the residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset they evaluate residual nets with a depth of up to 152 layers. They had even trained a deep model which is more than 1000 layers. While as there are not enough data, the 1000-layer network achieves not so well as the 152 layers' network. In our experiment, we design a residual network architecture for emotion recognition.

The ResNet we proposed has 4 residual blocks. Each block has a shortcut from input to the output. Two convolutional layers which has 1*1 and 3*3 filter sizes are included in each residual block. Detailed architecture is shown in Fig.1. For feature extraction, we use the last pooling layer as the output, which has 4608 output dimensions. We obtained 663 dimension features by a Principal Component Analysis (PCA) from the 4608 feature (95% of variance). The whole architecture of our proposed ResNet is shown in Fig. 1.

Table 1: The SVR CCC results on the development set

SVR	L2 Loss Primal		L2 Loss Dual		L1 Loss Dual	
Features	Arousal	Valence	Arousal	Valence	Arousal	Valence
Audio	0.796	0.453	0.796	0.459	0.785	0.462
LGBP-TOP	0.483	0.474	0.482	0.474	0.486	0.472
Geometric	0.378	0.612	0.375	0.612	0.401	0.563
ECG	0.265	0.144	0.274	0.159	0.320	0.167
HRHRV	0.365	0.259	0.366	0.264	0.392	0.257
EDA	0.061	0.136	0.063	0.136	0.122	0.234
SCL	0.054	0.150	0.042	0.167	0.116	0.229
SCR	0.046	0.085	0.057	0.063	0.117	0.126
MSDF	0.009	0.370	0.008	0.388	0.008	0.413
CNN	0.152	0.371	0.155	0.373	0.164	0.389
ResNet	0.174	0.341	0.184	0.359	0.215	0.366
AlexNet	0.116	0.444	0.115	0.450	0.127	0.480
Fusion	0.823	0.714	0.824	0.718	0.816	0.723

3.4 Physiological Features

We also use physiological features implemented by [7], including the electro-cardiogram (ECG) signals, the electro-dermal activity (EDA) signals, the heart rate(HR) and its measure of variability (HRV), the skin conductance response (SCR) and the skin conductance level (SCL), with a sliding centered window which size depends on the modality respectively.

The ECG records the electrical activity of the heart. From the ECG signal, we used 19 features including the zero-crossing rate, the four first statistical moments, the normalized length density, the non-stationary index, the spectral entropy, slope, mean frequency plus 6 spectral coefficients, the power in low frequency (LF, 0.04-0.15Hz), high frequency (HF, 0.15-0.4Hz) and the LF/HF power ratio.

The EDA is the property of the human body that causes continuous variation in the electrical characteristics of the skin. From the EDA signal, we used 8 features, including the four first statistical moments from the original time-series and its first order derivate.

The HRHRV is the heart rate and its measure of variability. We used the heart rate (HR) and its measure of variability (HRV) from the filtered (use a zero-delay bandpass filter (3-27Hz)) ECG signal, we used 10 features including the two first statistical moments, the arithmetic mean of rising and falling slope, and the percentage of rising values for each of those two descriptors.

The SCR is the phenomenon that the skin momentarily becomes a better conductor of electricity when either external or internal stimuli occur that are physiologically arousing. We used 8 features, including the four first statistical moments from the original time-series and its first order derivate.

The SCL is directly controlled by the sympathetic nervous system and indicates the activity of the sweat glands in the skin. We used 8 features, including the four first statistical moments from the original time-series and its first order derivate.

4. FUSION REGRESSION

4.1 SVR

In this paper we use Support Vector Regression (SVR) [16] prediction architecture, which is an effective classifier for dimensional emotion recognition. Given a training set of N data points $\{x_k, y_k\}_{k=1}^{N}$, where $x_k \in \mathbb{R}^n$is the kth input pattern and $y_k \in \mathbb{R}$ is the kth output pattern. For a sample (x, y), the traditional regression model is usually based on the difference between the output $f(x)$ of the model and the real output y to calculate the loss. If and only if $f(x)$ and Y are exactly the same, the loss is recorded as zero. While SVR supposed that if and only if the absolute value of the difference between X and Y is greater than ε, then calculate the loss.

Given a training set of N data points $\{x_k, y_k\}_{k=1}^{N}$, where $x_k \in \mathbb{R}^n$is the kth input pattern and $y_k \in \mathbb{R}$ is the kth output pattern, SVR method can be formalized as:

$$\min_{w,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} \ell_\varepsilon(f(x_i) - y_i), \quad (1)$$

The solution of SVR can be expressed as:

$$f(x) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)K(x, x_i) + b\ , \quad (2)$$

We use the LIBLINEAR [17], which is a popular open source machine learning libraries, the L2-regularised L2-loss dual solver and a unit bias was added to the feature vector was chosen, all others parameters were kept to default.

4.2 Multimodal Decision Level Fusion

Multimodal fusion of these modalities is then performed with a multiple linear regression model. The general form of the multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon_i, i = 1,2,\dots,n \quad (3)$$

Where i is the number of explanatory variables, $\beta_j (j = 0,1,2,\dots,n)$ and ε_i are called the regression coefficient. The formula above is also called the general regression function. For our problem, formula can be expressed as

$$P_{multi} = \sum_{i=1}^{N} \beta_i P_i + \varepsilon \quad (4)$$

Where P_i is the unimodal prediction of the modality i, P_{multi} is the fused prediction. $\beta_i (i = 1,2,\dots,N)$ and ε are called the regression coefficient

5. EXPERIMENTS

5.1 Training Regression CNN

We employ the Keras [18] implementation for image feature learning. We tried to directly train CNNs on the AVEC challenge dataset, while the result is not compromising. Though the dataset contains tens of thousands of images, most of them belong to same person, which limits its variety. So we decide to pre-train the CNNs

on other emotion datasets. As there are hardly any datasets for continuous emotion recognition, we choose to use the FER dataset, which is a seven class emotion dataset. First, we use expression images from the FER dataset to pre-train the CNN model using our proposed architecture. The learning rate is set to 0.005. In each iteration, 256 samples are used for stochastic gradient optimization. After 200 epoch's training, our proposed CNN get 67.82% recognition accuracy on the FER validation set. Then we fine-tune the model on the AVEC dataset. The base rmsprop learning rate is set to 0.0001. Images are randomly shifted and rotated to enhance its variety, which is very important for training CNN. However, the finetuned result is not as good as the CNNs trained on the FER. So we directly use the last convolutional layers' output as CNN features.

5.2 Monomial Features Regression Results

Concordance Correlation Coefficient (CCC) [30] is calculated as the evaluation metric for this challenge, which is defined as:

$$\mathrm{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \quad (5)$$

Where ρ is the Pearson correlation coefficient between the two variables x and y, μ_x and μ_y are the means for the two variables x and y, σ_x^2 and σ_y^2 are the variances for the two variables x and y.

Experiments results for each single feature set are shown in Table 1. Twelve feature sets are present. There is one feature set for audio modality; six feature sets (MSDF, proposed CNN, proposed ResNet, AlexNet, Geometric, LGBPTOP) for visual modality; five feature sets (ECG, EDA, SCL, SCR, HRHRV) for physiological modality. We can see that: 1) The best result for arousal dimension is achieved by audio modality, with CCC up to 0.796. 2) The best result for valence dimension is achieved by visual modality, with CCC up to 0.612. 3) The deep visual features are relatively effective on the valence dimension than the arousal dimension. These results are in agreement with previous studies [7].

5.3 Fusion Results

Table 1's last row shows the result of decision level fusion for results from the audio feature, the MSDF feature, the proposed CNN feature, the proposed ResNet feature, the AlexNet feature, the geometric feature, the LGBPTOP, the ECG feature, the EDA feature, the SCL feature, the SCR feature and the HRHRV feature on the development dataset. From this table, we can see that: 1) the L2-regularised L2-loss dual SVR solver performs best. 2) The best result for arousal dimension is 0.824; the best result for valence dimension is 0.723; the best average result for these two dimension is 0.771.

Table 2. The SVR Regression CCC on the development and test datasets for the fusion of all modalities

Datasets	CCC(arousal)	CCC(valence)	Avg
Development	0.824	0.718	0.771
Test	0.683	0.642	0.663

Table 2 shows the best result of decision level fusion for results from the audio feature, the MSDF feature, the proposed CNN feature, the proposed ResNet feature, the geometric feature, the LGBPTOP, the ECG feature, the EDA feature, the SCL feature, the SCR feature and the HRHRV feature on the development and test dataset.

Table 3. Performance comparison between the proposed approach and baseline in AVEC 2016 testing set.

Datasets	CCC(arousal)	CCC(valence)	Avg
Our approach	0.683	0.642	0.663
Baseline	0.682	0.638	0.660

Table 3 shows the comparison between our best result of decision level fusion and baseline for results from the audio feature, the MSDF feature, the proposed CNN feature, the proposed ResNet feature, the geometric feature, the LGBPTOP, the ECG feature, the EDA feature, the SCL feature, the SCR feature and the HRHRV feature on the test dataset. Compared with the baseline [7], the proposed approach performs better.

6. CONCLUSIONS

This paper presents our work in the Emotion Sub-Challenge of the 6th Audio/Visual Emotion Challenge and Workshop (AVEC 2016) Due to some research shows that video features are more important for emotion recognition, we try a variety of manual and depth visual features. Besides the baseline features, we proposed extra the MSDF, and CNN features to recognize the expression phases of the current frame. The method is evaluated on the RECOLA dataset and gain very promising achievement on the test set. In the future, we will focus on improving predictions by extracting more powerful modality features and try other machine learning methods to further improve the recognition performance.

7. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 61501035 and KJZXCJ2016042), the Fundamental Research Funds for the Central Universities of China (2014KJJCA15) and the National Education Science Twelfth Five-Year Plan Key Issues of the Ministry of Education (DCA140229). The authors would like to thank the organizers of AVEC 2016.

8. REFERENCES

[1] Stacy Marsella. Computationally Modeling Human Emotion. *Communications of the ACM*, Vol. 57 No. 12, Pages 56-67, 2015.

[2] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions From Facial Clues*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[3] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] J. R. Fontaine, K. R. Scherer, E. B. Roesch, P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, 18(12), 1050-1057,2007.

[5] C. Breazeal, "Emotion and sociable humanoid robots" [J]. *International Journal of Human-Computer Studies,* 2003,59 (1): 119-155.

[6] H. Gunes, M. Piccardi, Automatic temporal segment detection and affect recognition from face and body display, *IEEE Trans. Syst. Man Cybern. B Cybern.* 39 (1) (2009).

[7] M.Valstar, J.Gratch, B.Schuller, F.Ringeval, D.Lalanne, M.Torres Torres, S.Scherer, G.Stratou, R.Cowie, M.Pantic .

AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge, *AVEC Workshop*, 2016.

[8] F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions", in Proc. of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space(EmoSPACE), IEEE International Conference on Face & Gestures 2013.

[9] F. Ringeval, A. Sonderegger, B. Noris, A. Billard, J. Sauer,and D. Lalanne. On the influence of emotional feedback on emotion awareness and gaze behavior. In *Proc. of ACII*, pages 448–453, Geneva, Switzerland, 2013. IEEE Computer Society.

[10] Almaev, T. R., & Valstar, M. F. (2013, September). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 356-361). IEEE.

[11] Sikka, K., Wu, T., Susskind, J., & Bartlett, M. (2012, January). Exploring bag of words architectures in the facial expression domain. In Computer Vision–ECCV 2012. Workshops and Demonstrations (pp. 250-259). Springer Berlin Heidelberg.

[12] Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., & Bartlett, M. (2013, December). Multiple kernel learning for emotion recognition in the wild. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 517-524). ACM.

[13] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010, June). Locality-constrained linear coding for image classification. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3360-3367). IEEE.

[14] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol. 2, pp. 2169-2178). IEEE.

[15] Vedaldi, A., & Fulkerson, B. (2010, October). VLFeat: An open and portable library of computer vision algorithms. In Proceedings of the international conference on Multimedia (pp. 1469-1472). ACM

[16] Cortes C, Vapnik V. Support-Vector Networks. [J]. Machine Learning, 1995, 20(3):273-297.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, June 2008.

[18] F. Chollet, Keras, https://github/fchollet/keras, GitHub Repository (2015)

[19] Sepp Hochreiter and Jürgen Schmidhuber (1997)."Long short-term memory". Neural Computation 9 (8): 1735–1780.

[20] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, LSTM Modeling of continuous emotions in an audiovisual affect recognition framework, Image and Vision Computing, 2012.

[21] F. Eyben et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 2015. to appear.

[22] Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognitioninthewildchallenge2014: Baseline, data and protocol. In: Proceeding softhe16thinternationalconferenceonmultimodal interaction. ACM, pp 461–466

[23] Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: Computer vision and pattern recognition (CVPR), 2012 IEEE conference on IEEE, pp 2879–2886

[24] Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: Computer vision and pattern recognition (CVPR), IEEE conference on IEEE, pp 532–539

[25] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. Of ACM MM*, pages 835–838, Barcelona, Spain, October 2013.

[26] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[27] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. Computer Science, 2015.

[28] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S.,Ma, S., & Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 1-42.

[29] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In Neural Information Processing, pages 117–124. Springer, 2013.

[30] L. Li. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989.

[31] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," Psychol. Bull., vol. 11, no. 2, pp. 256–274, 1992.

[32] Sun, Bo, et al. "Combining Multimodal Features within a Fusion Network for Emotion Recognition in the Wild." Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015.

[33] Bo Sun, Liandong Li, Guoyan Zhou, Jun He, "Facial expression recognition in the wild based on multimodal texture features," J. Electron. Imaging 25(6), 061407 (2016), doi: 10.1117/1.JEI.25.6.061407.