Smart *Science* Vol. 2, No. 4, pp. 202-208(2014)

http://dx.doi.org/10.6493/SmartSci.2014.262

Automatic Emotional State Detection using Facial Expression Dynamic in Videos

Hongying Meng[1,*], and Di Huang[2]

[1]Department of Electronic and Computer Engineering, Brunel University, London, UK
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
* Corresponding Author / E-mail: hongying.meng@brunel.ac.uk, TEL: +44-1895265496

KEYWORDS : Facial expression, Emotion detection, Motion dynamic, Affective computing

In this paper, an automatic emotion detection system is built for a computer or machine to detect the emotional state from facial expressions in human computer communication. Firstly, dynamic motion features are extracted from facial expression videos and then advanced machine learning methods for classification and regression are used to predict the emotional states. The system is evaluated on two publicly available datasets, i.e. GEMEP_FERA and AVEC2013, and satisfied performances are achieved in comparison with the baseline results provided. With this emotional state detection capability, a machine can read the facial expression of its user automatically. This technique can be integrated into applications such as smart robots, interactive games and smart surveillance systems.

Manuscript received: June 11, 2014 / Accepted: July 01, 2014

1. Introduction

Human face provides an essential, spontaneous channel for the communication of mental states. In addition to functioning as a conversation enhancer, facial expressions directly communicate feelings, cognitive mental states, and attitude toward other people. Automatic emotional state detection from facial expression videos is to identify and understand the emotional state of a person as shown in Fig. 1. It has become an emerging multi-discipline research area that involves computer vision, machine learning, psychology, human computer interface and robotics. Much effort has been dedicated by psychologists to modeling the mapping between facial expressions and emotional states [1]. As smart interactive technology is becoming ubiquitous in our society, and the research community has been addressing the same modeling challenge from a computational perspective called affective computing [2].

Most of the initial work aimed at modeling the mapping between static facial expressions and emotion states [3]. The emotional state can be a discrete set of categories such as the six basic emotions (i.e. happiness, sadness, disgust, anger, surprise and fear) [4]. A typical approach is to model a facial expression as a set of local features. Some of the works that fall within this category have been inspired by Ekman's Facial Expression Coding System (FACS) [1] that codes a facial expression according to patterns of facial muscle activations. This type of approaches is highly dependent on the detection of these local features and usually time-consuming [5]. To overcome the burden and the limitation of these approaches, other methods have been proposed that provide template-models describing the face as a whole. One typical method using this approach is the Active Appearance Models (AAM) [6] that allows for the decoupling of the shape of the face from its appearance. It has been used for emotion recognition from still images and achieved good results [7].

Fig. 1 Can a smart machine recognizes the emotions from these facial expression videos automatically?

Apart from still face images, facial image sequences are also used for facial expression analysis. Not only the nature of the deformation of facial features, but also the relative timing of facial actions as well as their temporal evolution reveal the emotional states. It is clearly that an automated facial expression recognition system can recognize the facial actions, yet modeling their temporal behavior so that various stages of the development of a human emotion can be visually analyzed and dynamically interpreted by the machine. More importantly, it is often the temporal change that provides critical information about what we try to infer and understand in human emotions that possibly link to the facial expressions [8]. Pantic and

Rothkrantz [3] provide an in depth review of studies covering facial point model-ling, feature extraction and facial expression classification approaches.

Whereas most of these studies have focused on acted data set, there is an increasing need to work on more naturalist expressions in order to improve the performance of such a technology in a naturalist setting [9]. Zeng et. al. [10] review the state of the art on multimodal automatic recognition of emotion by combining facial expressions with other modalities such as voice and head pose. Furthermore, there is a need to create algorithms that take into account the temporal information of a facial expression and other part of the body. In recent year, there have been some attempts in this direction that produced interesting performances (e.g. [5, 11-13]).

Emotional state can also be represented in affective dimension space where emotional state recognition is regarded a regression problem to predict values of emotional dimensions. Gunes et. al. [14] provides a good review on the recent progress in this area. Furthermore, other emotional states can also be predicted using emotional state prediction such as pain level [15] and depression [16]. However, due to the complexity of the current methods, few systems have been widely used in real-world applications.

In this paper, we attempt to address this problem by extracting dynamic description of the spatial and dynamical motion of the facial expression and of the upper part of the body (i.e., head pose and shoulder in videos. We also propose an efficient automatic emotion detection system to predict the emotional state of the person. The system was evaluated on the public GEMEP_FERA dataset [17] and AVEC2013 dataset [18].

The rest of the paper is organized as the following. In section 2, the overview of the proposed automatic emotion detection system is described. Then the detailed information on the dynamic feature extraction and prediction is given in section 3 and 4 respectively. The system is evaluated in section 5 on two datasets and the paper is concluded in section 6.

2. Automatic emotion state detection system architecture

Automatic emotion detection system is to capture the emotional state of the user from their facial expression videos by the computer or machine. The overall automatic emotion detection system is shown in Fig. 2. It is machine learning based system that means the system firstly learns from the examples and then works itself. The learning process is called training. The facial expression video clips are input to the system. Then feature extraction stage extracts the facial expression dynamics from the data and makes the feature vectors. These features are fed to prediction part together with the emotion information of the video clips. The prediction part makes the mapping between the feature vector and the emotional information. In the testing phase, the mapping is used by the feature of new input facial expression video and the emotion information of the video is predicted.

3. Feature extractions

It is obvious that it is much easy to judge a person's emotion from the facial expression motions instead of the still images. So the key problem is how to capture these motions from the facial expression videos. In the following, facial dynamics are captured firstly and further analyzed.

3. 1 Facial expression movement capturing

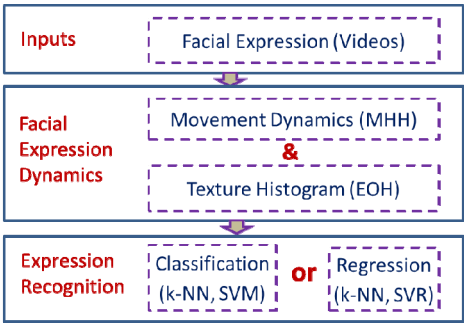

Fig. 2 The overall system for automatic emotion detection from facial expression videos. It mainly includes facial dynamic extraction and expression recognition parts.

There are many methods for motion detection in computer vision. Motion History Histogram (MHH) is a descriptive temporal template motion representation for visual motion recognition. It was originally proposed and applied in human action recognition [19]. It is described in Fig. 3 and the detailed information refers to [20] and [21]. It records the grey scale value changes for each pixel in the video. In comparison with other well-known motion features, such as Motion History Image (MHI) [22], it contains more dynamic information of the pixels and achieves better performance in human action recognition [20]. MHH not only provides rich motion information, but also remains computationally inexpensive [21].

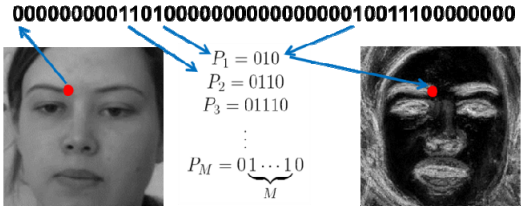

Fig. 3 The process of MHH computing. For each pixel, the change over the frames was coded by '1' if the difference between two consecutive frames is bigger than threshold, otherwise '0'. The counts of each pattern *Pi* (*i=1,...,M*) over all the frames on all the pixels generate *M* MHH images.

Fig. 4 shows the facial expression dynamics of five emotions (Anger, Fear, Joy, Relief, Sadness) in *M*=5 MHH images, which describe expression related cues in 5 levels. It is clear that different emotion has different facial expression dynamics. These motion features comprehensively capture the facial movements. For "Joy",

there are more motions on the face, so the values are high (i.e. bright in the Fig. 3.) while "Sadness" and "Fear", there is less motion from the face, so the values are small.

3. 2 Edge orientation histogram

MHH is further analyzed by the Edge Orientation Histogram (EOH) operator. The EOH is a simple, efficient and powerful operator that captures the texture information of an image. It has been widely used in a variety of vision applications such as hand gesture recognition [23] and object tracking [24].

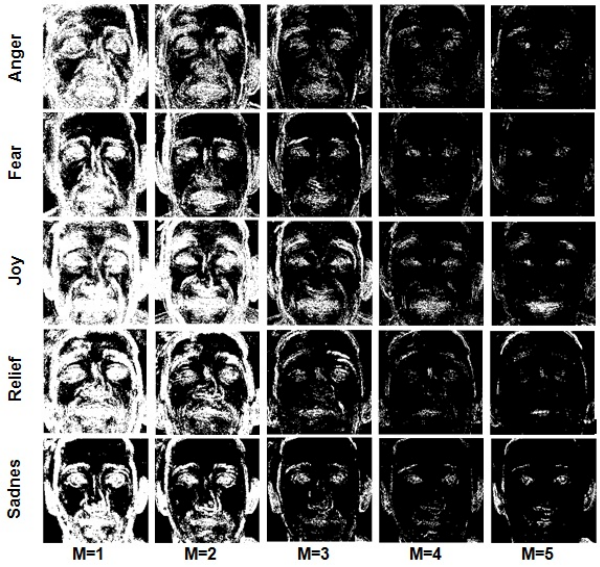

Fig. 4 The dynamics of facial expression videos in M=5 different scales. The first row from a video sample with emotion "Anger" and the followings are from the emotions of "Fear", "Joy", "Relief" and "Sadness" in the GEMEP_FERA dataset.

Fig. 5 shows how EOH feature is extracted. Firstly, the edge image is captured using Sobel edge detection algorithm from each MHH image. Secondly, the angle and intensity of the gradient function on each pixel is calculated and arranged into a polar coordinate system. Finally, the histogram from each block is normalized and concatenated into a feature vector. If the whole image is divided into 4x4 blocks and each polar coordinate system has 24 bins, the feature vector will have 384 components. M MHH images are done separately and then concatenated into one vector. If M=5, the feature vector will have 5x384=1920 components as shown in Fig. 6.

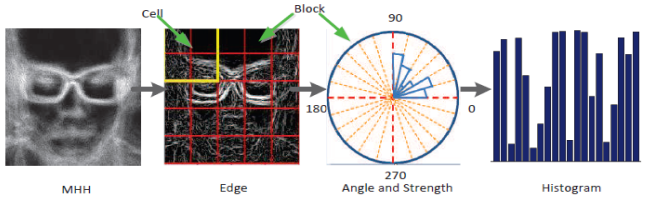

Fig. 5 EOH feature extraction process

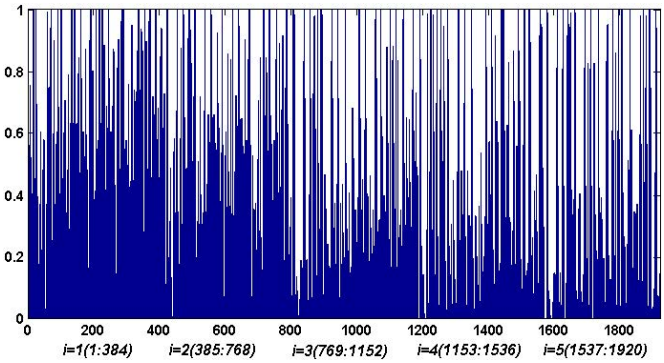

Fig. 6 The MHH_EOH feature is used for prediction. 384 components are generated from each MHH image and totally 1920 components are obtained for one video.

4. Expression prediction

There are two main types of prediction methods. If we ask for the categorization into discrete categories, it is a classification problem. The system produces the category to which the input video belongs. While, it is asked to predict an indicator on how much the emotion is, the predicted value is a real-value indicator. In this case, it is a regression problem. For classification process on videos, there are two evaluation protocols. One is to classify every frame and then use a majority voting scheme to choose the label to assign to the video. Another way is to treat a video as a whole and use a uniform feature for classification. We use the latter in this paper.

4.1 Discrete emotion categorization

For the case of discrete emotions, the prediction task is to identify which emotion the video shows. This is a typical multiclass classification problem. There exist many classification methods that can be used. Here, we introduce two most popular methods: k Nearest Neighbor (k-NN) and Support Vector Machine (SVM).

4.1.1 k-NN classifier

k-NN [25] is a lazy learning method for classifying objects based on the closest training examples in the feature space. Given a sample x, its predicted label $\hat{y}(x)$ can be decided by the majority of the class labels (j =1, 2, …, J) in its k neighbors $N(x) \in (1,2,\cdots N)$ within the N training sample set. i.e.

$$\hat{y}(x) = \arg\max_{j} \sum_{l \in N(x)} I_{\{y_l = j\}} \qquad j = 1,2,\cdots,J \tag{1}$$

where I_A is an index function where it is "1" if "A" is true. Otherwise it is "0".

4.1.2 SVM classifier

Support Vector Machine (SVM) [26] is a very popular and powerful classifier and has achieved excellent performance for pattern recognition task in many applications. SVM classifier constructs a hyper-plane in a high- or infinite-dimensional space to

separate samples from two categories. A good separation is achieved by the hyper-plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The model produced by support vector classification depends only on a subset of the training data (support vectors), because the cost function for building the model does not care about training points that lie beyond the margin.

Since SVM is a binary classifier, it can only classify two classes. For multiple class cases, a winner-take-all approach can be used for the multi-class classification, i.e. deciding the winner between the multiple emotions. Multi-class SVMs are usually implemented by combining several two-class SVMs. In each binary SVM, only one class is labelled as "1" and the others labelled as "-1". The one-versus-all method uses a winner-takes-all strategy. If there are *J* classes, SVM will construct *J* binary classifiers by learning. During the testing process, each classifier will get a confidence coefficient and the class with maximum confidence coefficient will be assigned to this sample.

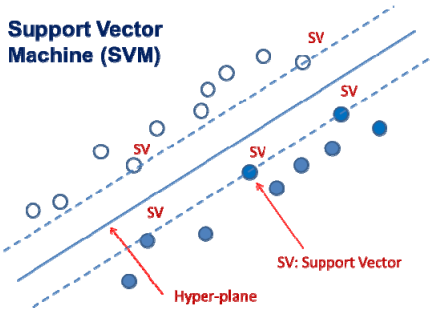

Fig. 7 Support Vector Machine (SVM) is to find the best hyper-plane to separate the samples from two classes.

4.2 Regression methods

In some case, the emotion is not a discrete one. It is a level of the emotional state such as depression. The task is the prediction of the emotional level instead of categorization. In this case, regression methods are needed.

4.2.1 k-NN regression

KNN can also be used as a regression method. Instead of predict the closest label, here, it produce a closest indicator. The equation will be as the following.

$$\hat{y}(x) = \frac{1}{k}\sum_{l=1}^{k} y_l, \qquad l \in N(x) \tag{2}$$

It averages the values of all its k neighbors $N(x) \in (1,2,\cdots N)$ within the training samples.

4.2.2 SVR regression

Similarly, SVM can also be used to solve a regression problem that is called Support Vector Regression (SVR). The SVR Algorithm [27] is very similar to SVM, however it treats a regression problem because the labels are not discrete numbers, but real values. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

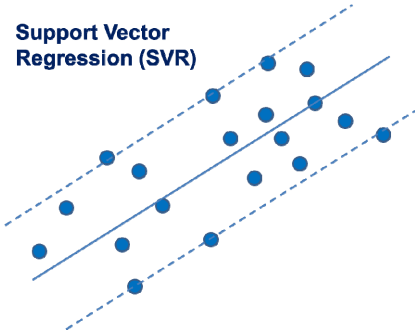

Fig. 8 Support Vector Regression.

4.2.3 PLS regression

The Partial Least Squares (PLS) regression [28] is a statistical algorithm that bears some relation to principal components regression. Instead of finding hyper-planes of minimum variance between the response and independent variables, it builds a linear regression model by projecting the response and independent variables to another common space. Since both the response and independent variables are projected to a new space, the approaches in the PLS family are known as bilinear factor models.

More specifically, PLS tries to seek fundamental relations between two matrices (response and independent variables), i.e. a latent variable way to model the covariance structures in these two spaces. A PLS model aims to search the multidimensional direction in the independent variable space that explains the maximum multidimensional variance direction in the response variable space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multi-collinearity among independent variable values. By contrast, standard regression will fail in these cases.

5. Experimental evaluation

5.1 Discrete emotion prediction

The GEMEP-FERA dataset consists of recordings of 10 actors displaying five types of emotional expressions (anger, fear, joy, relief and sadness) while uttering a meaningless phrase or the word 'Aaah'. There are seven subjects in the training data, and six subjects in the test set, three of which are not present in the training set (person independent partition). The other three are person-specific data. There are totally 155 video clips in the training set and 134 video clips in the testing set.

It is a typical 5-class classification problem and 5-folds cross-validation method was used for the experiments on the training set only. MHH_EOH features were input to k-NN (k=5) classifier and

linear SVM classifier with winner-take-all approach were used for the classification. The experimental results were shown in Table 1 based on the same accuracy measurement F1-score used [17].

From Table 1, it can be seen that both kNN and SVM achieved good results although no optimization was done on the parameters. The SVM classifier achieved almost 60% of F_1-score that is a good performance compared to the baseline results based on features with high computational cost. The confusion matrix of SVM classification is listed on the Table 2.

Table 1 F_1-score accuracy on the GREMP dataset in comparison with the baseline results [17].

Emotion	Baseline	kNN	SVM
Anger	0.89	0.49	0.58
Fear	0.20	0.51	0.59
Joy	0.71	0.52	0.62
Relief	0.46	0.54	0.52
Sadness	0.52	0.68	0.65
Average	0.56	0.55	**0.59**

Table 2 Confusion matrix of SVM method on the GREMP dataset.

	Anger	Fear	Joy	Relief	Sadness
Anger	**15**	3	1	1	2
Fear	6	**15**	1	0	1
Joy	9	4	**28**	9	0
Relief	1	1	1	**16**	8
Sadness	1	7	0	5	**20**

5. 2 Depression level prediction

The second dataset is the AVEC2013 challenge dataset [18]. The proposed approach is evaluated on the Audio/Visual Emotion Challenge (AVEC) 2013 dataset, a subset of the audio-visual depressive language corpus (AViD-Corpus). The dataset contains 340 video clips from 292 subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone in a number of quiet settings. There is only one person in each clip and some subjects feature in more than one clip. All the participants are recorded between one and four times, with an interval of two weeks. 5 subjects appear in 4 recordings, 93 in 3, 66 in 2, and 128 in only one session. The length of these clips is between 20 minutes and 50 minutes with the average of 25 minutes, and the total duration of all clips lasts 240 hours. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years. In the AVEC2013 dataset, the first 50 samples are for training; another 50 for developing; and the left 50 for test.

Fig. 9 Some examples of the AVEC2013 dataset.

In this dataset, depression level was given for each video ranging from 0 to 63. In order to capture the facial dynamic of these videos, MHH was used to capture the motion. For each video, MHH produces 5 ($M = 5$) images of temporal information of each video clip. EOH then operates on each MHH image, leading to a 384-dimensional feature vector, and the total MHH EOH representation concatenates all the 5 EOH features to make a vector of 1920 components. The experimental results were listed in Table 3 and compared to AVEC2013 baseline result on video modality [18].

From Table 3, it can be seen that proposed system outperforms the baseline results on both the development dataset and testing dataset on the video modality.

Table 3 Experimental results on the AVEC2013 dataset in comparison with the baselines.

Partition	Modality	Methods	MAE	RMSE
Development	Video	SVR	7.79	9.36
	Video	PLS	**7.16**	**8.86**
	Video	Baseline	8.74	10.72
Test	Video	PLS	**9.14**	**11.19**
	Video	Baseline	10.88	13.61

6. Conclusion and discussion

In this paper, a dynamic facial expression feature was presented based on the combination of MHH and EOH and then a novel automatic emotional state detection system was proposed using this facial expression dynamic feature and advanced machine learning methods. The system can read the facial expression videos and automatically produce an indicator for the emotional state of the user.

Two dataset were used for the testing. From all the experiments, it is clear that the proposed system achieved better performance than the baselines given by the organizers. For GREMP dataset, the testing labels are not published yet. So it is difficult to be compared with other methods in the FERA2011 challenge [17]. For AVEC2013, there exist some better results on the challenge. However, all of these results are based on both video and audio modalities and more complex algorithms on feature extraction and prediction. These methods are time-consuming and cannot work as quickly as proposed system.

Due to its simplicity, the proposed system has the potentials to be integrated into the real-world applications such as smart robots, interactive games and smart surveillance systems.

ACKNOWLEDGEMENT

This work by Hongying Meng was partially funded by the award of the Brunel Research Initiative and Enterprise Fund (BRIEF) and research exchange major award of UK Royal Academy of Engineering. This work by Di Huang was supported partly by the National Natural Science Foundation of China under Grant 61202237.

REFERENCES

[1] P. Ekman and W.V. Friesen, Facial action coding system: a technique for the measurement of facial movement (Consulting Psychologists Press, Palo Alto, CA, 1978).

[2] R. W. Picard, "Affective computing," M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, http://affect.media.mit.edu/pdfs/95.picard.pdf, Nov. (1995)

[3] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 1424-1445 (2000). DOI: 10.1109/34.895976

[4] P. Ekman, "An argument for basic emotions" *Cognition & Emotion*, **6**, 169-200 (1992) DOI: 10.1080/026999392084 11068

[5] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. "Facial expression recognition from video sequences: temporal and static modeling" *Computer Vision and Image Understanding*, **91**, 160-187 (2003). DOI: 10.1016/S1077-3142(03)00081-X

[6] G. J. Edwards, C. J. Taylor and T. F. Cootes, "Interpreting face images using active appearance models," In Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, pp. 300-305, April (1998) DOI: 10.1109/AFGR.1998.670965

[7] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model" *Signal Processing: Image Communication*, **19**, 723-740 (2004) DOI: 10.1016/j.image.2004.05.009

[8] Y. Zhang and Q. Ji, "Facial expression understanding in image sequences using dynamic and active visual information fusion," In Proceedings of IEEE International Conference on Computer Vision, IEEE, pp. 1297-1304, October (2003) DOI: 10.1109/ICCV.2003.1238640

[9] A. Kleinsmith and N. Bianchi-Berthouze, "Form as a cue in the automatic recognition of non-acted affective body expressions" *Affective Computing and Intelligent Interaction (Lecture Notes in Computer Science)*, **6974**, 155-164 (2011) DOI: 10.1007/978-3-642-24600-5_19

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 39-58 (2009) DOI: 10.1109/TPAMI.2008.52

[11] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 1683-1699 (2007) DOI: 10.1109/TPAMI.2007.1094

[12] R. Kaliouby and P. Robinson, in Real-Time Vision for Human-Computer Interaction, Branislav Kisacanin, Vladimir Pavlovic, and Thomas Huang, Eds. (Springer US, 2005), pp. 181-200.

[13] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences" *IEEE Transactions on Systems, Man and Cybernetics Part B*, **36**, 433-449 (2006) DOI: 10.1109/TSMCB.2005.859075

[14] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: a survey," In Proceedings of International Conference on Automatic Face and Gesture Recognition (FG 11), IEEE, pp. 827-834, March (2011) DOI: 10.1109/FG.2011.5771357

[15] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions" *IEEE Transactions on Cybernetics*, **44**, 315-328 (2014) DOI: 10.1109/TCYB.2013.2253768

[16] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC 13) , ACM, pp. 21-30, (2013) DOI: 10.1145/2512530.2512532

[17] M. F. Valstar, B. Jiang, M. M´ehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," In Proceedings of the Ninth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, pp. 921-926, March (2011) DOI: 10.1109/FG.2011.5771374

[18] 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 Proceedings of International Multimedia Conference (AVEC 13), ACM, pp. 3-10, October (2013) DOI: 10.1145/2512530.2512533

[19] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," In Proceedings of CVPR workshop on Embedded Computer Vision, IEEE, pp. 1-6, June (2007) DOI: 10.1109/CVPR.2007.383420

[20] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition" *Pattern Recognition Letters*, 30, 1049-1058 (2009) DOI: 10.1016/j.patrec.2009.03.003

[21] H. Meng, N. Pears, M. Freeman and C. Bailey, In Embedded Computer Vision, Advances in Pattern Recognition, B. Kisa˘canin, S.S. Bhattacharyya, and S. Chai, Eds. (Springer, 2009), pp. 139-162

[22] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates" *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**, 257-267 (2001) DOI: 10.1109/34.910878

[23] W.T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," In Proceedings of International Workshop on Automatic Face and Gesture Recognition (FG 94), IEEE, pp. 296-301, December (1994)

[24] C. Yang, R. Duraiswami and L. Davis, "Fast multiple object

tracking via a hierarchical particle filter," In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 05), IEEE, pp. 212-219, October (2005) DOI: 10.1109/ICCV.2005.95

[25] T. Cover and P. Hart, "Nearest neighbor pattern classification" *IEEE Transactions on Information Theory*, **13**, 21-27 (1967) DOI: 10.1109/TIT.1967.1053964

[26] C. Cortes and V. Vapnik, "Support-vector networks" *Machine Learning*, **20**, 273 (1995) DOI: 10.1023/A:1022627411411

[27] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, Support vector regression machines, Advances in Neural Information Processing Systems 9 (MIT Press, Cambridge, MA, 1997), pp. 155-161

[28] S. de Jong, "Simpls: an alternative approach to partial least squares regression" *Chemometrics and Intelligent Laboratory Systems*, **18**, 251-263 (1993) DOI: 10.1016/0169-7439(93)85002-X