Positive/Negative Emotion Detection from RGB-D upper Body Images

Lahoucine Ballihi, Adel Lablack, Boulbaba Ben Amor, Ioan Marius Bilasco, Mohamed Daoudi

Positive/Negative Emotion Detection from RGB-D upper Body Images

Lahoucine Ballihi[1], Adel Lablack[2], Boulbaba Ben Amor[1,2], Ioan Marius Bilasco[2], and Mohamed Daoudi[1,2]

[1] TELECOM Lille, Institut Mines-Télécom, Villeneuve d'Ascq, France
lahoucine.ballihi@telecom-lille.fr
[2] Laboratoire d'Informatique Fondamentale de Lille (UMR 8022) Université Lille 1
adel.lablack, *boulbaba.benamor*, *marius.bilasco*, *mohamed.daoudi*@lifl.fr

Abstract. The ability to identify users'mental states represents a valuable asset for improving human-computer interaction. Considering that spontaneous emotions are conveyed mostly through facial expressions and the upper Body movements, we propose to use these modalities together for the purpose of negative/positive emotion classification. A method that allows the recognition of mental states from videos is proposed. Based on a dataset composed with RGB-D movies a set of indictors of positive and negative is extracted from 2D (RGB) information. In addition, a geometric framework to model the depth flows and capture human body dynamics from depth data is proposed. Due to temporal changes in pixel and depth intensity which characterize spontaneous emotions dataset, the depth features are used to define the relation between changes in upper body movements and the affect. We describe a space of depth and texture information to detect the mood of people using upper body postures and their evolution across time. The experimentation has been performed on Cam3D dataset and has showed promising results.

Keywords: Emotional state, feature extraction, Grassmann manifold, depth features

1 Introduction

Several vision-based systems for automatic recognition of acted emotion have been proposed in the literature. Moreover, there is a move away from the automatic inference of the basic emotions proposed by Ekman [9] towards the inference of complex mental states such as attitudes, cognitive states, and intentions. The real world is dominated by neutral expressions [2] and complex mental states, with expressions of confusion, happiness, thinking, surprise, concentration, anger, etc. [19]. There is also a move towards analyzing spontaneous expressions rather than posed ones, as there is evidence of differences between them [6]. However most of these systems are evaluated only on datasets that are captured in controlled conditions.

The goal of this paper is to propose an approach to detect positive and negative emotions on spontaneous video streams. Positive emotions are expressed in

response to situations where the user enjoys its experience, where negative emotions are commonly expressed in response to situations that the person finds to be irritating, frustrating, or unpleasant. Assuming that expressive body movements and pose is responsible for independent, denotative information which can be mapped to the emotion space, we propose a 2D model which extracts metrics to detect positive/negative emotions that are highly correlated with the facial expressions anger and happy. A 3D model of features is also extracted from depth map video sequences and projected onto a Grassmann Manifold [8]. The 3-D feature set is converted into a feature vector to construct the Grassmann Manifold.

2 Related Work

The majority of approaches so far have made use of 2D image sequences, though a few works have started to use 3D facial geometry data. In addition research has been conducted into analysis of upper body expressions from 3D static data. In this section we discuss previous 2D dynamic facial expression analysis, and then go on to look at the 3D static and dynamic work that has been completed in this area, focusing mainly on the feature extraction stage as this provides the main differences in the analysis of expressions in 3D versus 2D images and upper body sequences.

Many studies have used classifier margins and posterior probabilities to estimate expression intensity without evaluating their performance to the ground truth [26, 22]. Several studies [25, 20] found that classifier margin and expression intensity were positively correlated during posed expressions. However, such correlations have typically been lower during spontaneous expressions. In a highly relevant study, Whitehill et al. [25] focused on the estimation of spontaneous smile intensity and found a high correlation between classifier margin and smile intensity on only five short video clips. Recent studies used other methods such as regression [7, 12, 11] and multiclass classifiers [15]. These studies have found that the predictions were typically highly correlated with expression intensity during both posed and spontaneous expressions.

Most of these previous works are limited since the majority of them focused on posed expressions, which limits the external validity and generalizability of their models. Some studies only coded expressions peak intensities, while others obtained frame-level ground truth, only for few subjects. The results reported on systems trained and tested on acted expressions might not be generalize to spontaneous ones.

Two studies are described in [10]. In the first study the authors evaluate whether two robust vision-based measures can be used to discriminate between different emotions in a dataset containing acted facial expressions under uncontrolled conditions. In the second one, they evaluate in the same dataset the accuracy of a commercially available software used for automatic emotion recognition under controlled conditions. Piana et al. [17] focus on the body movement analysis. A set of 2D and 3D features is introduced and a dictionary learning

method is described to classify emotions. Preliminary results are shown to assess the importance of the studied features in solving the problem. Shan et al. [21] use the fusion of facial expressions and body gestures at the feature level and derive an "affect" space by performing Canonical Correlation Analysis. In [18], the authors apply feature extraction techniques to multi-modal audio-RGB-depth data. They compute a set of behavioral indicators that defines communicative cues coming from the fields of psychology and observational methodology. The proposed approach relies on a set of features extracted from RGB-D upper body images captured in unconstraint environment where the subjects were expressing their emotions spontaneously.

3 Methodology and Contributions

We propose in the following two methods that indicate the positive/negative emotion performed spontaneously by a single person sitting in front of a camera.

3.1 RGB-based Approach

The main steps of this approach are divided into two stages. The first one consists in the different processing steps that allow to extract a normalized face from the input data. The second one consists in locating selected ROI from the face region and applying a specific filtering to each region to indicate the presence of a negative emotion, while the positive emotion is extracted using a neural network.

3.1.1 Image pre-processing

The image pre-processing procedure is a very important step in the facial expression recognition task. The aim of the pre-processing phase is to obtain images that have normalized intensity, are uniform in size and shape, and depict only the face region.

A) Face detection : We use a Boosted Haar like features method. The detection of the face is performed using the Viola-Jones face detector algorithm [24] available in OpenCV library. The selected parameters achieve the best speed and performance.
B) Eye detection : We use a neural network based approach to locate the positions of pupils. We derive only the eye detection code from the STASM library [16] which is variation of Active Shape Model of Coote's implementation. However, it works better on frontal views of upright faces with neutral expressions.
C) Up-right face : We estimate the orientation of the face using the vertical positions of the two eyes. If they are not in the same position, we compute the angle between these two pupil points and correct the orientation by setting the face center as origin point and we rotate the whole frame in opposite direction. It guaranteed a frontal upright position of the face up to 30 degrees in both sideways.

D) Face normalization : We use histogram equalization to normalize image intensity by improving its contrast. It aims to eliminate light and illumination related defects from the facial area.

3.1.2 Positive/Negative emotion indicators extraction

In order to detect negative emotion, we focus on the ROI located in the upper part of the face and includes the variations of AU4 of FACS where eyebrows are lowered and drawn together. We apply Gabor filter to this region of face. In the literature, 2D Gabor filters have been used for texture analysis and classification. Gabor filters have both frequency and orientation selective properties. Therefore a 2D Gabor function is composed of a sinusoidal wave of specified radial frequency which is the spacing factor between the kernels in the frequency domain and orientation which is modulated by a 2D Gaussian. Gabor representation of a face image is computed by convolving the face image I (x, y) with the Gabor filter. Majority of AUs samples associated to negative emotion face images has vertical lines above the nasal root. So, we choose vertical orientation for the Gabor filter with a frequency of $\sqrt{1.3}$, Mu equal to 0, Sigma equal to π and Nu equal to 3 as Gabor parameters. Then real and imaginary responses are added together to find the magnitude response. After a binary thresholding, the sum of the total pixels in the magnitude response of the filter, just above the nasal root is examined by a threshold value to detect a negative emotion. Brighter pixels in the magnitude responses are used as an indicator of negative emotion as depicted in the Figure 1.

In order to detect positive emotion, we use an analytic approach that performs wrapper based feature selection by exhaustive searching of all possible set of feature windows to find informative pixels. For a given emotion class, a mask is created to improve the Multilayer Perceptron (MLP) model's performance as illustrated in the Figure 1.

3.2 Depth-based Approach

In this section we shall present a second methodology for dynamic flows analysis of depth images of the upper body in order to capture its dynamics. As stated by researchers in social psychology, the expression of the emotion by human beings is performed in different ways. In addition to verbal and facial expressions, the body motions are considered as important source of information to interpret the emotions of humans, by their peers. We propose here a geometric framework to extract the body dynamics (movements) from the depth camera observing the subject.

Formally, we propose to map the flow of T depth images $\{I^t\}_{1\leq t\leq T}$ to a Grassmann manifold. Then, it is possible to use tools from differential geometry to quantify similarity/difference between elements and to find the most efficient way to move from one point to another (using geodesics), thus capture the dynamics. As illustrated in Figure 2, a sequence of depth images of the upper body observed by the camera is collected then organized as fragments of a fixed

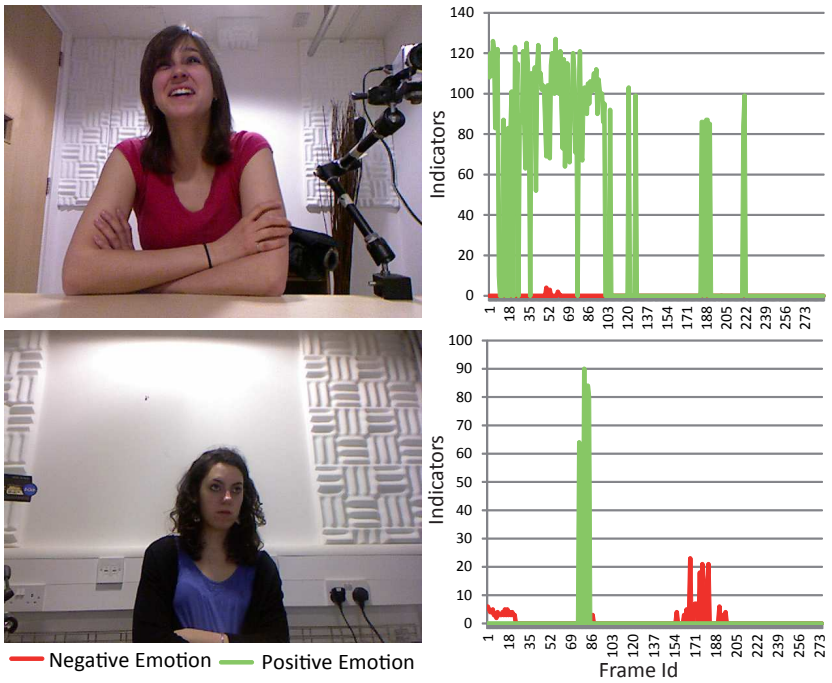

Fig. 1. Negative/Positive emotion indicators extracted from two subjects of the Cam3D dataset

window size, denoted by $W(t)$. Following [23], each video fragment is mapped to a Grassmann manifold by computing the the k-dimensional subspace. When two successive elements on the Grassmann manifold are available, says x and y, we compute the the initial velocity (tangent) vector v from x towards y to characterize the body movements across the observation time. The proposed method has several benefits compared to previous descriptions as optical flow or Motion History Images [3] to capture body movements across a depth image sequence:

- Depth images acquired with MS Kinect or any other cost-effective sensor are usually noisy, of low resolution, usually returns inaccurate depth measurements, and present missing parts. This could affect negatively conventional tools to capture motions across time for example optical flow or History of Motion History. The proposed geometric framework is more robust to these factor for several reasons. The most important one is the use of a sub-space computed from a set of images instead of single image which allows to filter out the noise and handle the low resolution problem.
- The use of the magnitude of the velocity vector between points on the Grassman manifold which represent depth video fragments to capture efficiently

the dynamics of the human body. This feature summarizes robustly the evolution across time of the body movements.

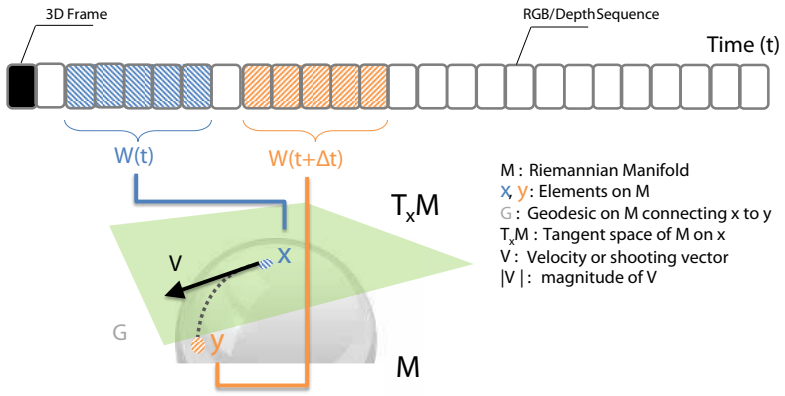

Fig. 2. Overview of our depth image approach: The flow of depth images is mapped through depth video fragments to a Riemannian manifold M on which tools from differential geometry are used to define tangent spaces, compute geodesic distances between elements and define initial velocity vector (element of the tangent space to M on x, $T_x(M)$) which allows to moves from x towards y along the geodesic path connecting them. The velocity vector will be used to capture the human body dynamics.

3.2.1 Modeling the Depth Flows on $G_{m,k}$

Before coming to the depth representation on the Grassman manifold a set of simple pre-processing steps are first applied, that we describe in the following. Given a depth video, we first separate at each frame $z = f(x, y)$ the body region from the background. Then, we reshape z to get a m-dimensional vector. A set of n successive frames, starting at a time t are considered together to form $W(t)$, a fragment of the depth sequence of size $m \times n$. From this matrix, we compute a k-dimensional subspace using the *Singular Value Decomposition* (SVD), where $k \leq n$. As a result, the depth video fragments $W(t)$ are viewed, after mapping, as elements on the Grassmann manifold $G_{m,k}$. Let X_i, $i = 1, \ldots, n$ denote n matrices of size $(m \times k)$ such that $X'X = I_k$. The rows of each X_i correspond to a low-dimensional representation of different "views" of the same object and the columns refer to the same depth pixel tracked across the observed video. Now, the depth videos are preprocessed and mapped to a Grassmann manifold of well-known geometry [5]. It is possible, using tools from differential geometry, to:

– Quantify the similarity of video fragments by interpolating between elements on $G_{m,k}$ and measuring the length of the obtained path (called geodesic path),

- Compute a sample mean of a set of elements,
- Compute the 'Energy' needed to move from an element towards another along the geodesic path connecting them,
- Model variability in a class of elements and perform fragments classification, etc.

In the following we shall recall few useful notations and relevant tools related to the geometry of the Grassmann manifold $G_{m,k}$. It is important to recall that the Grassmann manifold could be identified by a quotient manifold of the spacial orthogonal manifold by $SO(m)/SO(k) \times SO(m-k) = V_{m,k}/SO(k)$, where $V_{m,k}$ is the Stiefel manifold, the set of k-frames in $\mathbb{R}^m$, where a set of k orthonormal vectors in $\mathbb{R}^m$ is called a k-frame [5]. The Grassmann manifold $G_{m,k}$ obtained by identifying those matrices in $V_{m,k}$ whose columns span the same subspace (a quotient manifold). This quotient representation allows to extend the results of the base manifold (here $SO(m)$) such as tangent spaces and geodesics to the quotient spaces ($G_{m,k}$ in our case) [23]. Each point on $G_{m,k}$ can be represented by m-by-k matrix X such that $X^tX = I_k$.

1. **Tangent space**: To perform differential calculus on $G_{m,k}$, one needs to specify tangent space attached to it on a given point, then define an inner product. For the m-by-m identity matrix I, an element of $SO(m)$, the tangent space $T_I(SO(m))$ is the set of all m-by-m skew-symmetirc matrices [23]. For an arbitrary point $O \in SO(m)$, the tangent space at that point is obtained by a simple rotation of $T_I(SO(m))$: $T_O(SO(n)) = OX | X \in T_I(SO(m))$. For any $Y, Z \in T_O(SO(m))$ the inner product on the tangent plane $T_O(SO(m))$ is defined by $< Y, Z >= trace(YZ^T)$, where trace denotes the sum of diagonal elements. With this metric $SO(n)$ becomes a Riemannian manifold.
2. **Subspace angles**: The principal angles or canonical angles $0 \leq \theta_1 \leq \ldots \theta_m \leq \pi/2$ between the subspaces span(X) and span(Y) are defined recursively by $\cos(\theta_k) = \max_{u_k \in span(X)} \max_{v_k \in span(Y)} u_k' v_k$, subject to $u_k' u_k = 1$, $v_k' v_k = 1$, $u_k' u_i = 0$, $v_k' v_i = 0$, $(i = 1, \ldots, k-1)$. The principal angles can be computed from the SVD of $X'Y$ [4], $X'Y = U(\cos\Theta)V'$, where $U = [u_1 \ldots u_m]$, $V = [v_1 \ldots v_m]$, and $\cos\Theta$ is the diagonal matrix defined by :
$$\cos\Theta = diag(\cos\theta_1 \ldots \cos\theta_m)$$
3. **Geodesic distance**: The geodesic distance or the length of the minimal curve connecting span(X) and span(Y) is derived from the intrinsic geometry of Grassmann manifold. It could be expressed with the subspace angles as given by Eq. (1).
$$d(X,Y) = \left(\sum_{i=1}^{q} \theta_i^2 \right)^{1/2} = \|\Theta\|_2 \tag{1}$$

4. **Velocity (tangent) vector**: Let $\alpha(t) : [0,1] \rightarrow SO(m)$ a parametrized curve (or path) on $SO(m)$, then $\frac{d\alpha}{dt}$, the velocity vector at t, is an element of the tangent space $T_{\alpha(t)}(SO(m))$. The physical interpretation of this velocity vector computed along the special case of geodesics connecting $\alpha(0)$ to $\alpha(1)$ is the energy needed to move from $\alpha(0)$ towards $\alpha(1)$ along the minimal path α.

3.2.2 Modeling the Emotional Dynamics on $G_{m,k}$

To capture the body movements, we use the notion of initial velocity vector defined for geodesic between points on the Riemannian manifold. Let $\alpha(t)$ is a constant speed geodesic starting at x and pointing towards y. Then, it is possible to compute the initial velocity v, element of $T_x(SO(m))$. Formally, $v = exp_x^{-1}(y)$ where the exp^{-1} is the inverse exponential map for which the expression is not available analytically. We use the numerical solution described in [23].

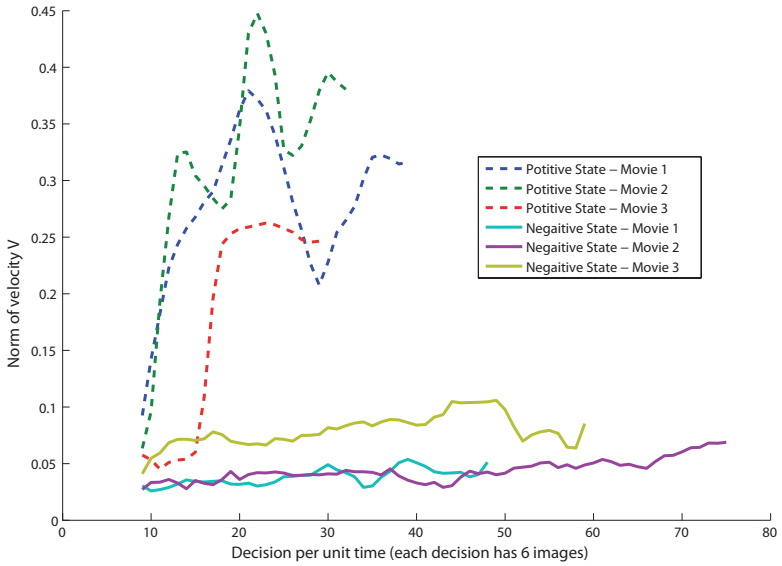

Fig. 3. Norm of velocities against time computed for three positive (dashed lines) and three negative (continuous lines) videos taken from the cam3D dataset.

Accordingly, given a flow of depth images of an observed subject, we split the video to fragments of a defined window size. Then, we measure the energy needed to move from successive elements on the Grassmann manifold to describe the dynamics of the human body. More specifically, we compute the norm of the velocity vectors as features for the observed motions across time. Figure 3 shows the norm of the velocity vector against time for three positive and three negative

emotional videos. It can be seen from these plots that positive emotional states are accompanied by more body movements compared to the negative emotional states. So, the values of the norm of the velocity vector across time is higher when a positive state is expressed by the subject observed by the camera. The feature vectors resulted from the previous step are feed up to a conventional classifier in order to train a classification model. In our experiments, presented in 4.2, we use the Random Forest classifier.

4 Experiments and Results

In order to evaluate the described methods, we tested them on a publicly available dataset called Cam3D [14] that records the spontaneous and natural complex mental states of different subjects using kinect. We use random forest classifier to predict the positivity and negativity of emotions.

4.1 3D database of spontaneous complex mental states

The corpus includes spontaneous facial expressions and hand gestures labelled using crowd-sourcing and is publicly available. The dataset contains a 3D multimodal corpus of 108 audio/video segments of natural complex mental states presented in [14]. The Figure 4 illustrates some examples of still images from the labelled videos.

Fig. 4. Examples of still images from the Cam3D dataset (from [14]).

The dataset has been divided in four groups. The first group induce cognitive mental states : thinking, concentrating, unsure, confused and triumphant. The second group of affective states was frustrated and angry. It was ethically difficult to elicit strong negative feelings in a dyadic interaction. The third group included bored and neutral. It was also hard to elicit boredom intentionally in a dyadic interaction, so this was only attempted in the computer based session by adding a "voice calibration" task. The last group included only surprised. In the computer-based session, the computer screen flickered suddenly in the middle of the "voice

calibration" task. In the dyadic interaction session, surprise was induced by flickering the lights of the room suddenly at the end of the session.

The purpose of our system is to recognize the state of an affective dimension. We are interested to detect positive/negative emotions. In order to identify the negative and positive emotions, we have used the emotion annotation and representation language (***EARL***) proposed by the Human-Machine Interaction Network on Emotion (HUMAINE) [1] that classifies 48 emotions. We use ***EARL*** to split the different mental states categories of Cam3D dataset into positive and negative classes.

4.2 Random Forest-based Classifier

Upper body-based mental state classification is a binary problem which classifies the state of query movie to positive or negative. We carry out experiments with the well-known machine learning algorithm, named Random Forest. Random Forest is a set of learning method that grows a forest of classification trees based on random selected features and thresholds. To give a decision for each window of the video, the input vector is fed to each tree and each tree gives a classification result. The forest selects the result by simple majority voting. We use random forest classification since it is reported that face classification by Random Forest achieves lower error rate than some popular classifiers [13].

4.3 Experimental results

We evaluate our approaches on the 108 videos of the Cam3D dataset. We use a 10-fold subject-independent cross-validation with Random Forest. For each round, images of some videos are randomly selected for testing, with images of remaining videos dedicated for training. For all the 10 rounds of experiments, no common subjects are used in training and testing. The input of random forest classifier is composed by the positive and negative indicators for the 2D approach where it is composed by the initial velocity descriptors for the depth image approach. The Figure 3 illustrates the comparison of the norm of the velocity vector between some negative and positive videos selected from the Cam3D dataset. Furthermore, we propose to combine both modalities by fusing their feature vectors that we feed up to RandomForest. The idea is to combine two sets of feature vectors of 2D and depth image methods and we group two sets of feature vectors into one union-vector (or a supervector).

Table 1. Comparison of the performance of 2D, depth image and RGB-D approaches using the Cam3D dataset.

Approach Based	2D image	Depth image	RGB-D
Random Forest Classifier	63.00%	68.12%	**71.32%**

The Table 1 shows the results of the modalities taken individually and their combinaison. The depth-based approach (which analyze the upper body dynamics) outperforms the color-based approach (which analyzes the face). This states that the movements that exhibits the upper body are more informative than the expressions displayed by the face for spontanious emotion. Their fusion outperforms the modalities taken individually, with 71.32% classification rate, which presents a good achivement on the the Cam3D database. Recall that this database presents difficult challenges as the occlusions, various body postures, different illumination conditions, and low quality of depth images. It is clear from this experiment that the facial expressions and the body dynamics provide synchronized and complementary interpretations of the human intentions.

5 Conclusion

In this paper, we have presented approaches to analyze (1) 2D facial movies, (3) depth movies of the upper body and (3) their combinaison to estimate the complex affect state of subjects in terms of positive and negative mood. These approches have been tested on the challenging Cam3D dataset collected using Kinect sensors. When the 2D-based approach is based on Positive/Negative emotion indicators extracted from the face part of the RGB channel, the Depth-based approch uses a geometric framework able to quantify the body dynamics from the low quality depth videos. The results demonstrates that the combinaison of 2D and depth (RGB-D) outperforms the modalities taken individually. In our future work, we will focus on reducing the impact of the occlusions on the quality of the extracted descriptors and using other strategy of feature fusion of information that comes from RGB-D data.

References

1. *HUMAINE Emotion Annotation and Representation Language (EARL)*, June 2006. `http://emotion-research.net/earl`.
2. S. Afzal and P. Robinson. Natural affect data - collection & annotation in a learning context. In *Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, Sept 2009.
3. Md. Atiqur Rahman Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: Its variants and applications. *Mach. Vision Appl.*, 23(2):255–281, March 2012.
4. Åke Björck and Gene H. Golub. Numerical Methods for Computing Angles Between Linear Subspaces. *Mathematics of Computation*, 27(123), 1973.
5. Yasuko Chikuse. *Statistics on Special Manifolds*. Springer, February 2003.
6. R. Cowie. Building the databases needed to understand rich, spontaneous human behaviour. In *Automatic Face Gesture Recognition, 8th IEEE International Conference on*, pages 1–6, Sept 2008.
7. Abhinav Dhall and Roland Goecke. Group expression intensity estimation in videos via gaussian processes. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3525–3528. IEEE, 2012.

8. Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
9. Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth. *Emotion in the Human Face*. Oxford University Press, 1972.
10. David Antonio Gómez Jáuregui and Jean-Claude Martin. Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions. In *Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop*, EmotiW '13, pages 17–22, New York, NY, USA, 2013. ACM.
11. László A Jeni, Jeffrey M Girard, Jeffrey F Cohn, and Fernando De La Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.
12. Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012.
13. Abbas Kouzani, Saeid Nahavandi, and K Khoshmanesh. Face classification by a random forest. In *IEEE Region 10 Conference: TENCON*, 2007.
14. Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel Riek. 3d corpus of spontaneous complex mental states. In *Conference on Affective Computing and Intelligent Interaction*, pages 205–214, 2011.
15. Mohammad H Mahoor, Steven Cadavid, Daniel S Messinger, and Jeffrey F Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 74–80. IEEE, 2009.
16. Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *10th European Conference on Computer Vision*, 2008.
17. Stefano Piana, Alessandra Staglianò, Antonio Camurri, and Francesca Odone. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In *IDGEI International Workshop*, 2013.
18. Víctor Ponce-López, Sergio Escalera, and Xavier Baró. Multi-modal social signal analysis for predicting agreement in conversation settings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 495–502. ACM, 2013.
19. Paul Rozin and Adam B. Cohen. High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. *Emotion*, 3(1):68–75, 2003.
20. Arman Savran, Bulent Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774 – 784, 2012. 3D Facial Behaviour Analysis and Understanding.
21. Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, pages 1–10, 2007.
22. Keiji Shimada, Yoshihiro Noguchi, and Takio Kuria. Fast and robust smile intensity estimation by cascaded support vector machines. *International Journal of Computer Theory & Engineering*, 5(1), 2013.
23. Pavan K. Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2273–2286, 2011.

24. P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.
25. Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.
26. Peng Yang, Qingshan Liu, and Dimitris N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *ICCV*, pages 1018–1025, 2009.