Depth video-based facial emotion analysis

Daniel Barmaimon Mendelberg
SnT - University of Luxembourg

Abstract

This paper presents a system to analyze spontaneous emotions from facial expressions captured with a 3D/depth sensing system. The possibility of controlling the resolution and noise level of the data allows to emulate off-the-shelf depth cameras. A whole pipeline is designed for the study of 3D depth video sequences of human faces. Three main steps define the pipeline: preprocessing, feature extraction and classification. In the preprocessing step the facial point cloud is aligned and reduced to a sorted group of 3D points, hence maintaining an invariant position with respect to the subject's face. This, in turn, allows a local temporal analysis. In the feature extraction step, the mean curvature for each of the points in the new cloud is calculated creating a 2D curvature map. The curvature map is divided into patches, over which a histogram analysis is performed. The bins of the histograms are the features of the system. The system is trained with classification models such as Support Vector Machines or binary trees, and validated with cross-validation. This system is designed as a tool for quantitative analysis of resolution enhancement and denoising algorithms, allowing to study the recognition rate before and after applying them in the context of facial expression recognition from 3D videos.

1. Introduction

Identification of human emotions using facial expressions is a difficult target that is being addressed during the last decades. Some applications fields are the of human-computer interaction and behavioral research. Some examples could be found in medical attendance and patients feelings [23], emphatic teaching [10,22,33] or customer satisfaction analysis [31].

The Facial Action Coding System (FACS) proposed by Ekman in [13] has been the standard during more than three decades for facial emotion recognition. It consists of 44 Action Units (AUs), with each of them describing the activity in a specific muscle or muscle group, producing a variation on the facial expression. More than 7000 combinations of AUs have been observed [12]. These combinations allow to code and classify 6 prototypical expressions of emotions namely, happiness, sadness, anger, fear, disgust and surprise. It is important to remark that the expression of emotions is not part of FACS. Only exaggerated emotions are the target for recognition but not others, nor the combination of the prototypical emotions nor their intensity. This more complex analysis with non-prototypical emotions is studied in other coding systems namely EMFACS [14], AFFEX [19] or MAX [18].

The evolution of research in the domain of emotion recognition from facial expression led to two classes of methods. First class is based on models that contain prior information about human faces. The information could be represented as facial landmarks [11,26,36,41]. Another way to represent the model is in a holistic way; considering the whole face as a model itself. Then differences in the shape variation are measured [15,28,30]. The second class of methods extracts features using image processing filters such as Gabor wavelets [5].

The human face could be considered as a complex system that combines muscle movements to express emotions. Great variety of muscles and interaction between them do not always lead to a clear definition of an emotion. Furthermore, emotions are difficult to define in a unique and unequivocal way, due to the variance in appearance for people of different age, gender, ethnicity or in the intensity in which the emotions are expressed [8, 20, 21].

Acquisition issues make facial expression recognition even more challenging. Illumination variations, scale, pose estimation, internal (mouth inside) and external (facial hair or glasses) occlusions are some of the most common problems that appear in combination with the noise inherent to acquisition systems.

The appearance of new acquisition systems and the evolution of hardware have allowed to carry a 3D analysis of facial expressions. This helps to solve and simplify some of the aforementioned problems namely pose estimation, scale and illumination [35]. As a tradeoff, a problem of higher dimensionality appears, increasing the amount of information to process and the computational time required for this task. Model-based methods take advantage of mapping between 2D textures and 3D points as in [15, 28]. Some other methods use the difference between models to estimate the expressions [15, 28, 30]. Both cases use fiducial points or landmarks to identify specific facial regions or fit the model correctly. In the case of feature-based models, the new features are not only the 3D points and distances [38, 39, 43], but also the curvature locally measured around them [6, 24, 25, 27, 34, 45].

The use of facial expression recognition systems in real-time applications is the main reason that motivated researches to move from static to dynamic analysis. Time is a new source of information but also brings a new dimension. The amount of frames needed to correctly detect an emotion, or the speed to recognize it are variables to study. It happens in [3, 4], where the representation of the sequences as points of a manifold simplifies the problem by reducing its dimensionality.

The cost of acquisition systems and the size of the data to compute are important parameters to consider. In [34], Savran et al. propose a method for recognition of emotion's valence using consumer cameras. Dimensionality is not the only problem there. 3D acquisition systems do not produce a constant number of vertices for each frame, a condition that is necessary in some methods, like [3, 4, 34]. It is corrected during preprocessing. The most recent advances in facial expression recognition are collected in the surveys [9, 32].

In the search of a system in facial expression recognition that could be applicable to real world problems, there are two main issues. The emotions in most of the systems presented are posed. Spontaneous facial emotions is a topic that has not been studied in depth. One of it differences is that variations between expressions are more subtle and difficult to detect when emotions are not posed [16, 29, 46]. The second problem is related with the cost and installation limitation of acquisition systems. 3D acquisition systems are expensive and bulky. The use of affordable acquisition systems that measure the depth such as consumer Time-of-Flight cameras or Kinect cameras eliminates this issues. On the other hand, they have a low resolution and a high level of noise compared with the high performance of other more expensive systems.

The current paper focuses on the recognition of facial expressions from 3D videos. The expressions are spontaneous and without any exaggeration from the subjects. For this purpose a system is developed to extract features from facial regions. The features are based on the curvature over sampled points, and allow the classification of emotions. The system is able to recreate different levels of resolution and noise, allowing to emulate the acquisition of a low resolution commodity camera starting from a 3D high definition database. This allows to evaluate and test the performance of denoising and resolution enhancement algorithms under real acquisition conditions. A whole pipeline is created and presented, where the main contributions include:

- Complete system to detect facial emotions for 3D video sequences, based on local analysis of curvature.
- Ability to control the resolution and the level of noise on clean data, allowing to simulate the acquisition of different vision systems.
- Analysis over spontaneous emotions, that are natural and realistic, which could be used in real world applications.
- Features representation as histogram of curvatures over patches, adding robustness, and making the system independent of the number of points for the face in each frame.

The rest of the paper is structured in the following way: In the Section 2, the proposed pipeline for feature extraction and classification will be described in detail. Section 3 presents and discusses the experimental results and Section 4 collects conclusions and future work.

2. Methodology

The method proposed estimates emotions from 3D facial video sequences, independently of the ethnicity, gender, pose or illumination. The main target is to get the features and perform training and classification models. The features are the curvature values of a group of sampled points from a cloud that represents the face. A whole pipeline has been developed in order to accomplish this task. The section is divided in a part explaining the databased used and three main steps that describe the system: preprocessing, feature extraction, classification.

2.1. The database

The database used to test and evaluate the pipeline was BP4D-Spontaneous [47]. This database has been obtained with a high-resolution 3D acquisition system, [17], at a speed of 25 fps, and contains 30000 - 50000 vertices per frame and 1040 x 1392 2D texture images. There are eight different tasks for each of the 41 subjects. For each of these tasks the emotions were induced spontaneously instead of posing it. To achieve this target the subjects were exposed to different situations such as an interviewer telling a joke, watching a documentary, the presence of an unpleasant smell in the room, etc. The database keeps the track of a group of 83 3D landmarks and 49 2D points. Approximately 500 frames per sequence have labels for AUs. This allows the codification between each frame and the emotion that is represented. For the case of some AUs, the intensity is also given.

Extraction of features over the high resolution and clean of noise database allows to train and recognize the emotions using given labels. In this case the recognition rates are the upper limits for the classification of same sequences with different levels of noise and resolution after applying enhancing and denoising algorithms.

2.2. Preprocessing

A point cloud P_t is defined as $P_t = \left\{p_1^t, ..., p_{k_t}^t\right\}$, where $t = 1, 2, ..., T$, k_t represents the total number of points in a time step t, T is the total number of frames in a sequence, and $p_i^t \in \mathbb{R}^3$ with $i = 1, 2, ..., k_t$. Each point cloud P_t should be prepared for the computation of the curvature. A representation of a cloud of points for a single frame can be observed in Figure 1(a), where the color coding represents depth values with respect to the acquisition system. For each point cloud P_t in a specific sequence, the nose-tip $n_t \in P_t$ is located. The most robust way to get its location is to use the Viola-Jones method [42] over the matching texture image I_t. The mapping between the 2D texture map and the 3D cloud of points is used to locate n_t.

In order to remove the shoulders, neck and back part of the head, P_t is cropped using a sphere of radius r_{sphere} with the center at n_t, as it is shown in Figure 1(b). The result of this cropping step is a reduced point cloud $P_t' \subset P_t$.

P_t' is aligned with respect to P_{t-1}' using Iterative Closest Point (ICP) algorithm [7]. Then Principal Component Analysis (PCA) [44] is performed on P_t', returning the three main directions $\vec{v_1}, \vec{v_2}, \vec{v_3}$ of an orthonormal reference system.

A plane π_t is created to get a sub-sampled sorted cloud of points. Plane π_t is created using n_i and two of the principal directions, $\vec{v_1}$ and $\vec{v_2}$, so it can be written as $\pi_t = \{n_t, \vec{v_1} \times \vec{v_2}\}$ where × is the vector cross product. The vetors $\vec{v_1}$ and $\vec{v_2}$ are respectively represented as red and green axis in Figure 1(c). The resulting plane is plotted in Figure 1(d).

Limits over the plane are found by projecting the 3D points over it, and finding maximum and minimum values in the principal directions used to create the plane. Projection of point cloud P_t' over plane π_t is represented as P_t'', where $P_t'' = \left\{p_{k_t'}''\right\}$ and k' is the total number of points in the cropped point cloud for the time step t. The mathematical definition of this projection is given below.

$$P_t'' = proj_\pi \ P_t' \tag{1}$$

where, for a given vector $\vec{u}$

$$proj_\pi \ \vec{u} = \vec{u} - \frac{\vec{u} \cdot \vec{n}}{||\vec{n}||^2}\vec{n} \quad , \quad \vec{n} = \frac{\vec{v_1} \times \vec{v_2}}{||\vec{v_1} \times \vec{v_2}||} \tag{2}$$

$$\tag{3}$$

Once that P_t'' is calculated, the limits for the grid of sampling points are defined by the projection of each point in P_t'', onto the each axis given by the two principal components $\vec{v_1}$ and $\vec{v_1}$. These projections over the principal directions are defined as W_t^1 and W_t^2, and $W_t^c = \left\{w_1^c, ..., w_{k_t'}^c\right\}$, where $c = 1, 2$. This could be appreciated in the following equations.

$$w_t^c = proj_{\vec{v_c}} \ (p_{t_i}' - n_t), \quad \forall p_{t_i}' \in P_t' \tag{4}$$

$$w_{t_{max}}^c = \max \ \{w_t^c\} \tag{5}$$

$$w_{t_{min}}^c = \min \ \{w_t^c\} \tag{6}$$

where, $i = 1, ..., k_t'$ and the projection of a vector $\vec{u}$ onto a vector $\vec{v}$ is defined as follows:

$$proj_{\vec{v}} \ \vec{u} = \frac{\vec{u} \cdot \vec{v}}{||\vec{v}||^2}\vec{v} \tag{7}$$

We define a parameter $\alpha \geq 1$ and $\alpha \in \mathbb{R}^+$ to control the resolution of the sampled point cloud. It is defined as the quotient of the number of points in the original point cloud res_{init} divided by the number of points of the sub-sampled one res_{sub}. It has as limits 1 for the maximum resolution and k_t' for the minimum, where k_t' is the total amount of points in P_t' int the time step t.

$$\alpha = \frac{res_{init}}{res_{sub}} \quad , \quad \alpha \in [1, k_t'] \tag{8}$$

In order to make classification operations, the number of sample divisions is the same in each of the principal components of the plane. Using the number of points k_t' in a cropped face and the limits over the plane π_i, calculated using Eq. 5 and Eq. 6, it is possible to obtain a grid over the limited part of the plane for sampling. A grid of sampled points over the plane is defined as A_t where $A_t \in \pi_t$. As the grid A is squared, total amount of its points could be expressed as $\beta = \gamma \times \gamma$ where $\beta, \gamma \in \mathbb{N}$. The calculation for γ is expressed below

$$\gamma = \left\lceil \sqrt{\frac{k_t'}{\alpha}} \right\rceil \tag{9}$$

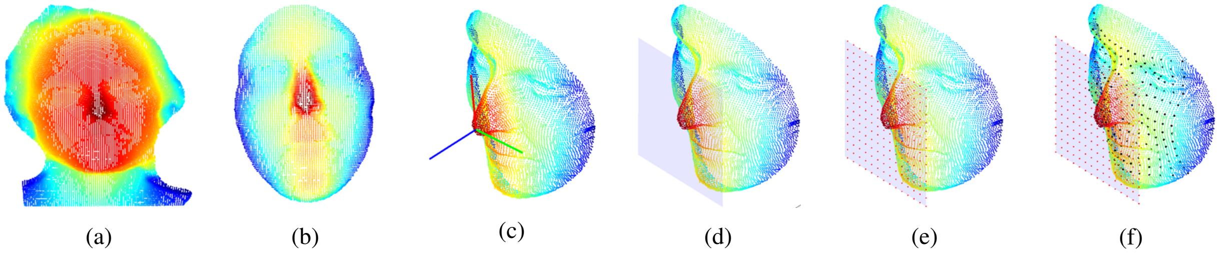

(a) (b) (c) (d) (e) (f)

Figure 1: Preprocessing pipeline: (a) Original point cloud, (b) Cropped point cloud using a sphere, (c) PCA represented over the nose, (d) Plane creation using principal directions, (e) Grid definition, (f) Interpolated point cloud

where $[\cdot]$ is defined as the closest natural number to the exact value of the division. The sampled points in A are equally distanced in $\vec{v}_1$ and $\vec{v}_2$ as shown in Figure 1(e). A is bilinearly interpolated over the cropped point cloud P_t', as it is reflected in Figure 1(f). The new cloud $Q^t = \left\{q_1^t, ..., q_\beta^t\right\}$, where $q_i^t \in \mathbb{R}^3$ and $i = 1, ..., \beta$ and it represents the face for a subject in each frame. Each point in this cloud remains in an invariant position with respect to the face along time.

2.3. Feature extraction

To obtain the features for each frame it is necessary to compute the mean curvature of the interpolated points Q^t. The mean curvature value κ_i^t of each point q_i^t in the interpolated cloud is calculated where the main idea is to fit a plane tangent to the points given in the neighborhood of q_i^t, [37]. It is based on the estimation of the eigenvectors $\vec{v}_g$ of the covariance matrix C_i^t created for the m nearest neighbors of the selected point q_i^t.

$$C_i^t = \frac{1}{m}\sum_{j=1}^{m}(q_j^t - \bar{q}^t)\cdot(q_j^t - \bar{q}^t)^T \tag{10}$$

where $\bar{q}^t$ is the mean of $\left\{q_l^t\right\}_{l=1}^m$. The calculation of eigenvectors is perform as follows.

$$C_i^t \cdot \vec{v}_g = \lambda_g \cdot C_i^t \quad , \quad g \in \{1,2,3\} \tag{11}$$

where λ_j are the eigenvalues of C_i^t, and $vecv_j$ its eigenvectors. The mean curvature κ_i^t is finally given by

$$\kappa_i^t = \frac{min\left\{\lambda_g\right\}}{\sum_{g=1}^3 \lambda_g} \tag{12}$$

The number of neighbors m to measure the curvature of a certain point q_i^t is regulated by a radial parameter r_{dist}. The values for the curvatures of the point cloud Q^t is stored as a matrix M_t that could be interpreted as a 2D curvature map of the face. An example of a curvature map could be observed in Figure 2(a).

We divide curvature maps M_t into an $N \times N$ matrix of squared patches with the same size $S \times S$. For the extraction of the features a histogram analysis is performed for each of the patches. Our considered features are the bins of each of the patches' histograms extracted from a curvature map. The total amount of features is given by two parameters; N that is the number of divisions for the curvature map in one direction, and the number of bins n_b for the histogram representation. The resulting total amount of features is $N \times N \times n_b$. This statistical approach avoids problems related with different number of points in different frames, returning always the same amount of features. The square shape and the number of patches are important characteristics that allow a deeper analysis that is reflected on the classification results. To ensure that important information remains located in the same facial area, independently of the number of patches, a special cropping step is required. It starts by maintaining the same position of the nose-tip as the reference $(0,0)$, then keeping the maximal number of points that allow the division of M_t into an exact natural number of square patches. This cropping results in a new curvature map M_t' that is smaller in size than M_t. This means that the number of points should be distributed $N \times N$ patches of $S \times S$ pixels each one, as shown in Figure 2(b) for the case of $N = 3$. The reference to set the image over the curvature patch is the center of the curvature map, in this case the nose location.

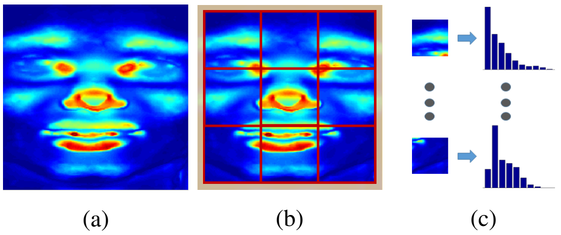

(a) (b) (c)

Figure 2: Feature extraction: (a) Initial curvature map, (b) Cropping and division into patches, (c) Histogram analysis of each patch

2.4. Classification

For each group of task sequences, happiness or sadness, a training and classification process is performed. AUs information given by the database is converted into emotion labels. Time step for the analysis is a frame, instead of the use of a sliding window of frames as was performed in [3, 4] where the main emotion was considered as the one with more appearance in a window subsequence. The reason is to understand the characterization of spontaneous emotions and the difficulty of the differentiation between them. Specific singularities of emotions are lost when a voting process is used.

A Support Vector Machine (SVM) model is created for each of the emotions detected in the group of sequences. This allows to analyze how different is each single emotion independently from others. A binary tree classifier is created to evaluate the classification of the whole system for each of the task group of sequences.

For the evaluation of both SVM and binary tree models, 10-fold cross-validation is performed over each of the task of sequences.

3. Results and discussion

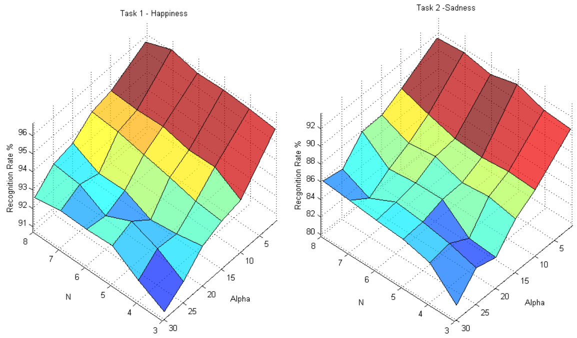
Evaluation of the pipeline is the first analysis to perform. Implementation of the system leads to a tool able of controlling the resolution and level of noise of the point cloud. The new cloud must be centered at the nose and its points are sorted and structured into a matrix, emulating the behavior of a depth commodity camera. Resolution is controlled with the parameter α, while the noise level is controlled with the characteristics that represent the noise. In Figure 3(a), 3(b), 3(c) resolution is changed in absence of noise. Noise added for the evaluation is a Gaussian noise with a mean $\mu = 0$ and a standard deviation σ. In Figure 3(d), 3(e), 3(f) different levels of noise have been added, maintaining $\alpha = 1$. There is a need of knowing the best tunning for the parameters that regulate the classification. Number of patches in which the curvature maps are divided would be use as a variable over different resolutions. The parameter N defines the total number of patches of the experiment as $N \times N$, and α controls the resolution. A binary tree model is created and trained for the evaluation of the whole system. Six different spontaneous emotions appear in each of the tasks; happiness, sadness, surprise, fear, anger and neutral. The results for the two different tasks after 10-fold cross validation of the models are collected in Table 1, Table 2 and displayed in Figure 4. The results shows that the best performance is obtained for 5×5 division of curvature maps with high resolution and 8×8 with low resolution. There is a considerable variation in the performance of the system for the two different tasks. Taking into account that the emotions are spontaneous, happiness facial expression differs naturally much more from neutral expression than sadness which has subtle variations in facial muscles position.

Task 1 - Happiness						
$\alpha \backslash N \times N$	3×3	4×4	5×5	6×6	7×7	8×8
1	95.80	95.93	96.00	95.93	**96.25**	95.68
10	93.55	94.39	94.34	94.66	94.59	**94.66**
15	93.31	93.43	93.76	94.03	94.31	**94.43**
20	92.86	92.55	92.35	92.98	92.94	**93.63**
25	91.83	92.67	93.19	92.51	93.25	**93.40**
30	91.09	91.84	**92.77**	92.69	92.73	92.47

Table 1: Recognition rate of the system for task 1

Task 2 - Sadness						
$\alpha \backslash N \times N$	3×3	4×4	5×5	6×6	7×7	8×8
1	90.84	91.07	**92.07**	91.26	91.84	91.55
10	87.22	87.37	87.06	87.27	88.10	**89.10**
15	85.75	86.35	86.01	87.04	86.63	**87.90**
20	83.19	83.44	85.92	85.87	85.52	**87.84**
25	83.81	84.66	85.18	**85.50**	84.03	85.40
30	81.40	84.29	85.03	85.25	85.59	**85.86**

Table 2: Recognition rate of the system for task 2

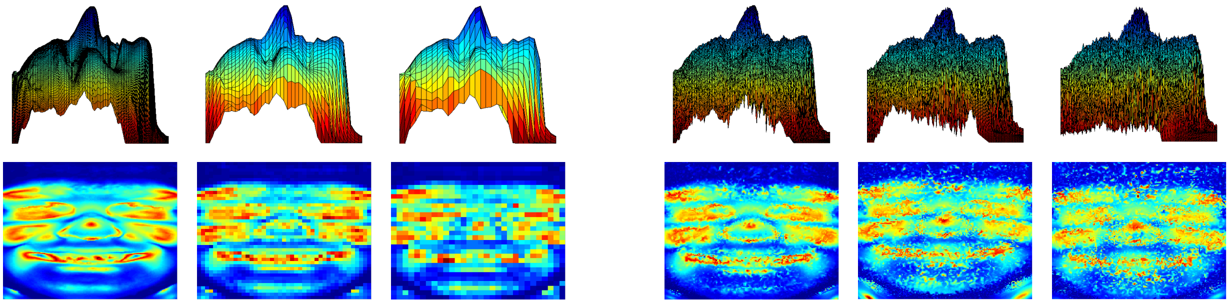


(a) Happiness

(b) Sadness

Figure 4: Evaluation of recognition rate of the system with noiseless data when varying N and α

Curvature is very sensitive to noise. To analyze the effect of noise on the performance of the system, a Gaussian noise with a standard deviation varying from 0 to 5 mm. has been added. The resolution is then reduced by a value of $\alpha = 30$, analog to the resolution obtained by a PMD Camboard Nano time-of-flight camera [40]. Number of patches to divide is 8×8, as it had the best performance for the same resolution in the previous analysis of noise-free sequences. As expected, Figure 5 shows that the recognition rate drops considerably when there is an increase of the noise level. The recognition rate for 'happiness' drops in a much smoother way as



(a) $\alpha = 1$ (b) $\alpha = 15$ (c) $\alpha = 30$ (d) $\sigma = 1mm.$ (e) $\sigma = 3mm.$ (f) $\sigma = 5mm.$

Figure 3: Point clouds and curvature maps for different resolutions and noise levels

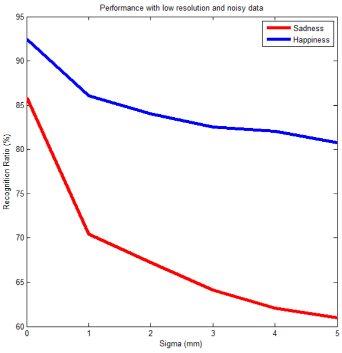


Figure 5: Analysis of the system's performance with low resolution and noisy data, $\alpha = 30, N \times N = 8$

compared to 'sadness' and corroborates that it is relatively easier to detect happiness among others even with a high level of noise. For sadness, however, a drop down to 60% of recognition is obtained for a noise variance of 25 mm. Such a degradation in performance requires an additional preprocessing to attenuate the effect of noise and low resolution.

4. Conclusions and future work

A complete system is developed for the detection and evaluation of facial spontaneous emotions using depth video sequences. The possibility to vary the resolution and the level of noise allows the emulation of current depth commodity cameras. Statistical approach using patches for analysis of the curvature maps gives flexibility and robustness to the method proposed. It is not constraint to a fixed number of points per step of time and counteracts the effects of inaccurate registration. Best results for recognition rate with low resolution demonstrate that it is better to make the patches as small as possible. This leads to use mean curvatures values for the given points as features for the classification system in the case of data with low resolution. Analysis on spontaneous emotions on facial expressions is complicated due the subtle changes that differentiate them. Luckily some expressions, such as happiness, are essentially different by nature from any other. The current system can be used for the analysis of resolution enhancement and denoising algorithms, with the possibility of measuring differences quantitatively.

Future work has several topics to cover. First one will be related with the search of the optimal radius r_{dist} to compute the curvature. This parameter could also be used with different values, allowing a multi-scale classification system for intensity evaluation of spontaneous emotions. Secondly, feature representation could be changed into a radial distribution of the patches. In this way a greater weight will be given to the areas that are closer to the nose-tip. The use of combination of 2D features with curvature features could improve the performance of the system and should be object of study. Adaptations should be made to optimize system for its use in real-time applications. The use of denoising and super-resolution algorithms such as [1, 2] in combination with the current system could improve the recognition rate, allowing the use of the system with off-the-shelf depth commodity cameras.

References

[1] Kassem Al Ismaeil, Djamila Aouada, Bruno Mirbach, and Björn Ottersten. Dynamic super resolution of depth sequences with non-rigid motions. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 660–664, Sept 2013.

[2] Kassem Al Ismaeil, Djamila Aouada, Thomas Solignac, Bruno Mirbach, and Björn Ottersten. Real-time non-rigid multi-frame depth video super-resolution. In *Computer Vision and Pattern Recognition Workshop (CVPRW), IEEE International Conference on*, June 2015.

[3] Taleb Alashkar, Boulbaba Ben Amor, Stefano Berretti, and Mohamed Daoudi. Analyzing trajectories on grassmann manifold for early emotion detection from depth videos.

[4] Boulbaba Ben Amor, Hassen Drira, Stefano Berretti, Mohamed Daoudi, and Anuj Srivastava. 4-d facial expression recognition by learning geometric deformations. 2014.

[5] Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, and Javier R Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[6] Stefano Berretti, Alberto Del Bimbo, Pietro Pala, Boulbaba Ben Amor, and Daoudi Mohamed. A set of selected sift features for 3d facial expression recognition. In *20th International Conference on Pattern Recognition*, pages 4125–4128, 2010.

[7] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[8] Michael Biehl, David Matsumoto, Paul Ekman, Valerie Hearn, Karl Heider, Tsutomu Kudoh, and Veronica Ton. Matsumoto and ekman's japanese and caucasian facial expressions of emotion (jacfee): Reliability data and cross-national differences. *Journal of Nonverbal Behavior*, 21(1):3–21, 1997.

[9] Antonios Danelakis, Theoharis Theoharis, and Ioannis Pratikakis. A survey on facial expression recognition in 3d video sequences. *Multimedia Tools and Applications*, pages 1–39, 2014.

[10] Sidney D'mello and Art Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):23, 2012.

[11] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10):974–989, 1999.

[12] Paul Ekman. 2. methods for measuring facial action. 1982.

[13] Paul Ekman and Wallace V Friesen. Facial action coding system. 1977.

[14] Wallace V Friesen and Paul Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2:36, 1983.

[15] Boqing Gong, Yueming Wang, Jianzhuang Liu, and Xiaoou Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 569–572. ACM, 2009.

[16] Hatice Gunes. Automatic, dimensional and continuous emotion recognition. 2010.

[17] Di3D Inc. `http://www.di3d.com`.

[18] Carroll E Izard. *The maximally discriminative facial movement coding system*. Acad. Computing Services and Univ. Media Services, University of Delaware, 1983.

[19] Carroll Ellis Izard, Linda M Dougherty, and Elizabeth Ann Hembree. *A system for identifying affect expressions by holistic judgments (AFFEX)*. Instructional Resources Center, University of Delaware, 1983.

[20] Rachael E Jack, Roberto Caldara, and Philippe G Schyns. Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, 141(1):19, 2012.

[21] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, page 201200155, 2012.

[22] Ashish Kapoor, Selene Mota, and Rosalind W Picard. Towards a learning companion that recognizes affect. In *AAAI Fall symposium*, pages 2–4, 2001.

[23] Bee Theng Lau. Portable real time emotion detection system for the disabled. *Expert Systems with Applications*, 37(9):6561–6566, 2010.

[24] Pierre Lemaire, Mohsen Ardabilian, Liming Chen, and Mohamed Daoudi. Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[25] Pierre Lemaire, Boulbaba Ben Amor, Mohsen Ardabilian, Liming Chen, and Mohamed Daoudi. Fully automatic 3d facial expression recognition using a region-based approach. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 53–58. ACM, 2011.

[26] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[27] Ahmed Maalej, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, and Stefano Berretti. Shape analysis of local facial patches for 3d facial expression recognition. *Pattern Recognition*, 44(8):1581–1589, 2011.

[28] Iordanis Mpiperis, Sotiris Malassiotis, and Michael G Strintzis. Bilinear models for 3-d face and facial expression recognition. *Information Forensics and Security, IEEE Transactions on*, 3(3):498–511, 2008.

[29] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[30] Gang Pan, Song Han, Zhaohui Wu, and Yuting Zhang. Removal of 3d facial expressions: A learning-based approach. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2614–2621. IEEE, 2010.

[31] Nancy M Puccinelli, Scott Motyka, and Dhruv Grewal. Can you trust a customer's expression? insights into nonverbal communication in the retail context. *Psychology & Marketing*, 27(10):964–988, 2010.

[32] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression

recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[33] Abdolhossein Sarrafzadeh, Hamid Gholam Hosseini, Chao Fan, and Scott P Overmyer. Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems. In *Advanced Learning Technologies, 2003. Proceedings. The 3rd IEEE International Conference on*, pages 336–337. IEEE, 2003.

[34] Arman Savran, Ruben Gur, and Ragini Verma. Automatic detection of emotion valence on faces using consumer depth cameras. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 75–82. IEEE, 2013.

[35] Arman Savran, BüLent Sankur, and M Taha Bilge. Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. *Pattern recognition*, 45(2):767–782, 2012.

[36] Thibaud Senechal, Kevin Bailly, and Lionel Prevost. Impact of action unit detection in automatic emotion recognition. *Pattern Analysis and Applications*, 17(1):51–67, 2014.

[37] Craig M Shakarji et al. Least-squares fitting algorithms of the nist algorithm testing system. *Journal of Research-National Institute of Standards and Technology*, 103:633–641, 1998.

[38] Hamit Soyel and Hasan Demirel. Facial expression recognition using 3d facial feature distances. In *Image Analysis and Recognition*, pages 831–838. Springer, 2007.

[39] Hao Tang and Thomas S Huang. 3d facial expression recognition based on automatically selected features. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[40] PMD Technologies. `http://www.pmdtec.com/products_services/reference_design.php`.

[41] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001.

[42] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[43] Jun Wang, Lijun Yin, Xiaozhou Wei, and Yi Sun. 3d facial expression recognition based on primitive surface feature distribution. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1399–1406. IEEE, 2006.

[44] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

[45] Mingliang Xue, Ajmal Mian, Wanquan Liu, and Ling Li. Fully automatic 3d facial expression recognition using local depth features. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1096–1103. IEEE, 2014.

[46] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[47] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.