A Facial Motion Capture System Based on Neural Network Classifier Using RGB-D Data

Fateme Zare Mehrjardi
Computer Engineering Department, Yazd University, Iran
Zarefateme41@yahoo.com

Mehdi Rezaeian[3]
Computer Engineering Department, Yazd University, Iran
mrezaeian@yazd.ac.ir

Abstract

Analysis of live and dynamic movements by computer is one of the areas that draws a great deal of interest to itself. One of the important parts of this area is the motion capture process that can be based on the appearance and facial mode estimation. The aim of this study is to represent 3D facial movements from estimated facial expressions using video image sequences and applying them to computer-generated 3D faces. We propose an algorithm which can classify the given image sequences into one of the motion frames. The contributions of this work lie mainly in two aspects. Firstly, an optical flow algorithm is used for feature extraction, that instead of using two subsequent images (or two subsequent frames in a video), the distinction between images and the normal state is used. Secondly, we realize a multilayer perceptron network that their inputs are matrices obtained from optical flow algorithm to model a mapping between person movements and database movement categories. A three-dimensional avatar, which is made by means of Kinect data, is used to represent the face movements in a graphical environment. In order to evaluate the proposed method, several videos are recorded in order to compare the available modes and discovered modes. The results indicate that the proposed method is effective.

Keywords: motion capture, depth data, 3D model, facial expression, Kinect, optical flow.

1. Introduction

Motion capture is the process of recording a live motion event and translating it into actionable data which allows for a 3D recreation of the performance. In other words, transforming a live performance into a digital performance. The Motion capture includes the motion capture of the entire body, the hands, the face etc. The motion capture process can be divided into offline and online. In the former, the data are processed after being recorded and saved in order to separate noisy data, and then applied to the computerized character. While in the latter, the data are recorded and applied to the character simultaneously. They are also associated with noisy data. Motion capture process can also be divided into two marked and unmarked types. In the former, some marks are located on different parts of the body. In this method there are problems such as marks overlapping and the person not being comfortable with the special covering. So as to overcome the problems existent in the marked method, unmarked motion capture used. This method is less precise than the marked method (Skogstad et al., 2010; Lindequist and Lönnblom, 2004). This study aims to capture some facial expressions such as mouth, eyebrows, eyes and cheeks by using colorful images and depth data that are obtained from the RGB-D camera. The rest of this paper is structured as follows. A brief review of the related works is represented in the next section. The main concepts are introduced in Section 3. Section 4 discusses the proposed method. Subsequently, the results of the experiments are mentioned in Section 5 and finally, Section 6 concludes the paper.

[3] Corresponding author

2. Related Works

In (Michel and El Kaliouby, 2003) it is used a facial feature tracker to deal with the problems of face localization and feature extraction in spontaneous expressions. The employed tracker uses a face template for primary localization of 22 face key points from video sequences. Moreover, a filter has been used for tracking the situation of these points in the video sequences. For each expression, a vector of feature displacements is calculated by taking the Euclidean distance. These replacements are given to the support vector machine classifier. Then the SVM model determines the mode of an unseen frame by means of feature replacements amounts.

In (Ge et al., 2008), a system is proposed for reconstructing human facial expressions. The proposed system is composed of four key modules: face detection, feature extraction, classification and artificial emotion generation. In this system, it is assumed that there is only one face in the image and the face take up a significant area in the image. Furthermore, OpenCV library is used to separate the face region from an image and extract facial features. After the separation of the face area, facial feature extraction is conducted to locate the positions of the eyebrows, eyes, eyelids, mouth and wrinkles. These positions can be obtained by some key points along with special mathematical properties. Considering the Facial Action Coding System (FACS) the twenty-two facial muscles are closely related to human facial expressions. In fact, the human facial expressions are due to the movements of facial muscles beneath the skin. Each facial muscle is represented by means of a pair of key points called driven points and fixed points. The driven points are points which can move along the facial expression, while the fixed points are points which are fixed during a facial expression. The elastic forces generated from the facial muscles changes were calculated as the feature vector that is elaborated in the article (Ge et al., 2008). Afterwards, this vector is given to the SVM classifier as an input. A total of 600 videos are prepared for six facial expressions (100 videos for each facial expression), namely happiness, sadness, fear, disgust, anger, and surprise. All the videos are captured from one person. Then, all the data are randomly divided into two groups 480 videos for training and 120 videos for testing. At the end, a robot is employed in order to simulate the expression modes on it.

Li (Li et al., 2013) proposed a system for representing the face expression modes. The user performs this task in a natural environment without any mark and records a 3D facial video of a person with Kinect. In the face-tracking phase, a set of predefined key points on the face are tracked using improved active appearance model. The head 3D models and the 3D movements of these key points are computed. Then, the controlled parameters are derived for facial animation according to the underlying facial model. The face tracking role in this system is to extract the defined movement units by means of Candide-3 standards. Candide-3 standard is a very simple and publicly available face model which has been very popular in many research labs around the world. Afterward, a communicative channel is created for transforming the obtained controlled parameters from the face tracking to the 3ds Max software for face animation.

Cao et al. (Cao et al., 2014) introduce FaceWarehouse, a database of 3D facial expression models for visual computing. With an off-the shelf RGBD depth camera, raw datasets from 150 individuals aged 7–80 from several ethnicities were captured. For each person, they captured her neutral expression and 19 other expressions, such as mouth-open, smile, angry, kiss, and eyes-shut. For each RGBD raw data record, a set of facial feature points on the color image, such as eye corners, mouth boundary and nose tip are automatically localized, and manually adjusted if the automatic detection is inaccurate. They then deform a template facial mesh to the depth data as closely as possible while matching the feature points on the color image to their corresponding locations on the mesh. Starting from the 20 fitted meshes (one neutral expression and 19 different expressions) of each person, the individual-specific expression blendshapes of the person are constructed.

In (Seddik et al., 2013), a system is proposed that is able to recognize the facial expressions and then fitted to a 3D face virtual model using the depth and RGB data captured from Microsoft's Kinect camera. This system starts working by detecting the face and segmenting its regions. It identifies the facial expressions using eigenface criterion on the RGB images and reconstructs the face from the filtered depth data.

3. Main concepts

3.1. Optical flow algorithm

An optical flow (OF) algorithm calculates the changes of brightness patterns from one frame to another. This task is done by using the intensity values of the neighboring pixels. Algorithms that calculate the changes for all the image pixels are referred to as dense optical flow (DOF) algorithms. Sparse optical flow (SOF) algorithms estimate the changes for a selected number of pixels in the image. DOF algorithms, such as Horn-Schunck method (Nourani-Vatani et al., 2012), calculate the changes at each pixel by using public constraints. These methods generally prevent from feature extraction but are not robust to noise. SOF algorithms such as the Lucas-Kanade approach (Nourani-Vatani et al., 2012), assume local smoothness. These methods provide more robustness to noise but offer a much sparser flow field. Due to this increased robustness, the SOF algorithms are preferable over DOF algorithms for many applications. In this research, Lucas-Kanade approach is used.

3.2. Steps of facial expression detection

Each facial expression recognition system should go through some steps before classifying the emotions. The first step consists of separating the face from other available objects in the image or video. When the face is recognized, the face changes and movement should be followed in order to be able to detect the facial emotions. There might be obstacles such as illumination or background changes which can make a difficult detection procedure. When the face is separated, the system must find features such as the lips, eyebrows, cheeks movement, etc. In the last stage after extracting the features to determine what emotions are represented, facial expressions are classified. Indeed, the system must have been trained with a database that contains a variety of expressions to be able to recognize the face emotion regardless of the issues such as the age, gender, race, skin color. Briefly, a facial expression recognition system consists of the following stages (Rangari and Gonnade, 2014):

- Face detection and tracking it
- Feature extraction
- Expression mode classification

3.2.1. Face detection and face tracking

The first step in facial expression analysis is to detect the face from the image or the given video frames and following it in various frames. Recognizing the face within an image is termed as face detection or face localization whereas tracking it across the different frames of a video sequence is termed as face tracking. Face detection algorithms and tracking it from feature detection algorithms are due to the finding of a special representative in images. One of the popular and improved methods in this domain is Kanade-Locase-Tomasi tracking method. Kanade and Locase developed a feature extraction algorithm in which two images are corresponded and it is assumed that the second frame in one consecutive frame of images; due to the small inner frame movement, there is a conversion of the first image. In the year 2004, Viola and Jones (Viola and Jones, 2004) developed a learning method based on the forehead detection algorithm (Vinay, 2012). The algorithm works based on the Ada boost learning algorithm and the techniques such as integral image and attentional cascade make the Viola-Jones algorithm highly efficient and precise.

3.2.2. Feature Extraction

In recent years, many various algorithms have been proposed to extract the facial points from images or video sequences of faces. The four basic approaches are as follow:

3.2.3. Geometry-based approaches

In geometry-based method, the features are extracted according to the geometrical relations such as the positions and the feature width. Geometry-based features describe the shape of the face and its components, such as the mouth or the eyebrow. Although in the geometry-based method less memory is needed, its application is difficult (Dhawan and Dogra, 2012; Yen and Nithianandan, 2002).

3.2.4. Appearance-based approach

In appearance-based methods, the images are converted to a vector. In fact, each image is represented with one point in a space with 'n' dimensions. Due to the multidimensionality of the images, the computing process is very difficult and time consuming. In this manner, statistical methods are used to reduce the images' dimensions in order to be able to obtain the data frequency. In this approach, without losing the original information, the images could be represented in a sub-space with less dimensions so called "feature sub space". For recognition, the image is taken to the sub space and there the similarity between the new image and the given image is analyzed. These subspaces can be divided into two: linear and nonlinear types. From among linear sub spaces it can be referred to the principal component analysis (PCA (Karamizadeh et al., 2013; Kaur and Sarabjit Singh, 2015)), local binary pattern (LBP (Huang et al., 2011; Liu et al., 2011)) and the linear discriminant analysis (LDA (Kaur and Sarabjit Singh, 2015)). LBP algorithm which is used in this study will be explained later in feature extraction section. Each of these three methods, considering their own statistical view point, define one sub space and transfer the data to that sub space. For data comparison it is enough to analyze the similarity between the transferred vectors in the regarding sub space with the entered data to the sub space. Linear models are not able to separate the information that its frequency is nonlinear. For overcoming this problem, the data are transferred to a space with high dimensions and then linear functions are used for their separation (Dhawan and Dogra, 2012; Yen and Nithianandan, 2002).

3.2.5. Template-based approach

In template-based methods, a single template or a multiple template which are designed by means of energy functions are used for feature extraction. This method is easily applicable, but it suffers from an insufficient fitness supply and a need for a lot of memory. This method cannot also work with images with complicated backgrounds.

Some methods for feature extraction use changeable templates for face feature extraction. These are flexible templates that are made by means of a priori knowledge of the form and size of various features. The templates' size and form can be changed in order to be well fitted. Each of these templates is evaluated by means of an energy function. The least amount of the energy function corresponds to the best fitness of the image. These methods along with some changes in the scale and the head circular movement, work well in recognizing the eyes and the mouth. But nose and eyebrow modeling has always been a complicated task (Bhatt and Shah, 2011; Bagherian et al., 2009).

3.2.6. Color-based approach

This approach uses skin color to isolate the face area. Any non-skin color region within the face is viewed as a candidate for eyes or mouth. The effectiveness of such techniques on images with various backgrounds is rather limited, due to the diversity of ethnical backgrounds (Bhatt and Shah, 2011; Bagherian et al., 2009).

3.2.7. Expression Mode classification

After the face detection stages and feature extraction, the final stage in face detection system is an appropriate modeling to classify the extracted features of special modes. Classification is actually a two-staged process. In the first step, a model is prepared based on the collection of data available in the database. The training data collection consists of records, samples, examples or objects that each contains a set of attributes. For each instance a class label is denoted. Each of the training data collection samples is called a training sample that have been selected randomly from the data collection. In the second step, the trained model is used to classify appropriately the new data sample. In this study, the neural network classifier is used.

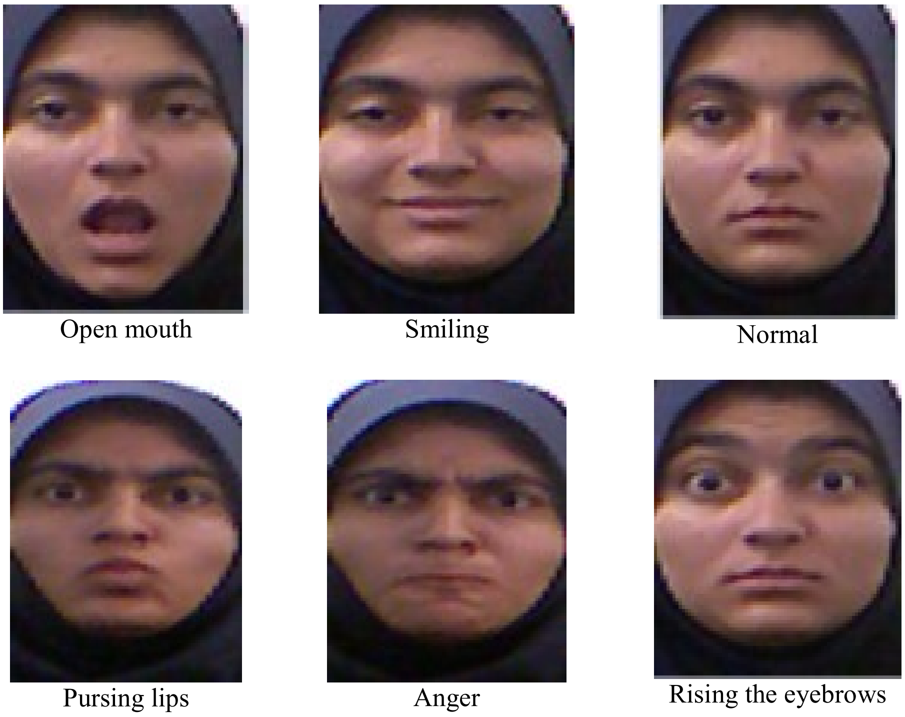
4. Proposed Method

4.1. Data collection procedure

In this study, data are obtained from the Kinect camera that benefits from colorful images and depth data. Kinect can record colorful and depth data simultaneously at 30 frames per second. The data are collected from the person who initially pose in front of the camera with normal face mode and then the various modes are represented.

It should be noted that data are obtained at different distances from the Kinect camera and in different lighting conditions.

Figure 1 shows various facial modes in our database.



Open mouth

Smiling

Normal

Pursing lips

Anger

Rising the eyebrows

Figure 1. Peak of facial expression in database

4.2. Preprocessing

In this stage, a video is prepared using the color data captured from Kinect camera. The face region in each frame is obtained from the video using Viola-Jones algorithm (Figure 2).

Because of different distance from the Kinect camera, the obtained images from the face must be re-sized, in order to have the same size. At the end, the colored images are converted to gray-scaled images.

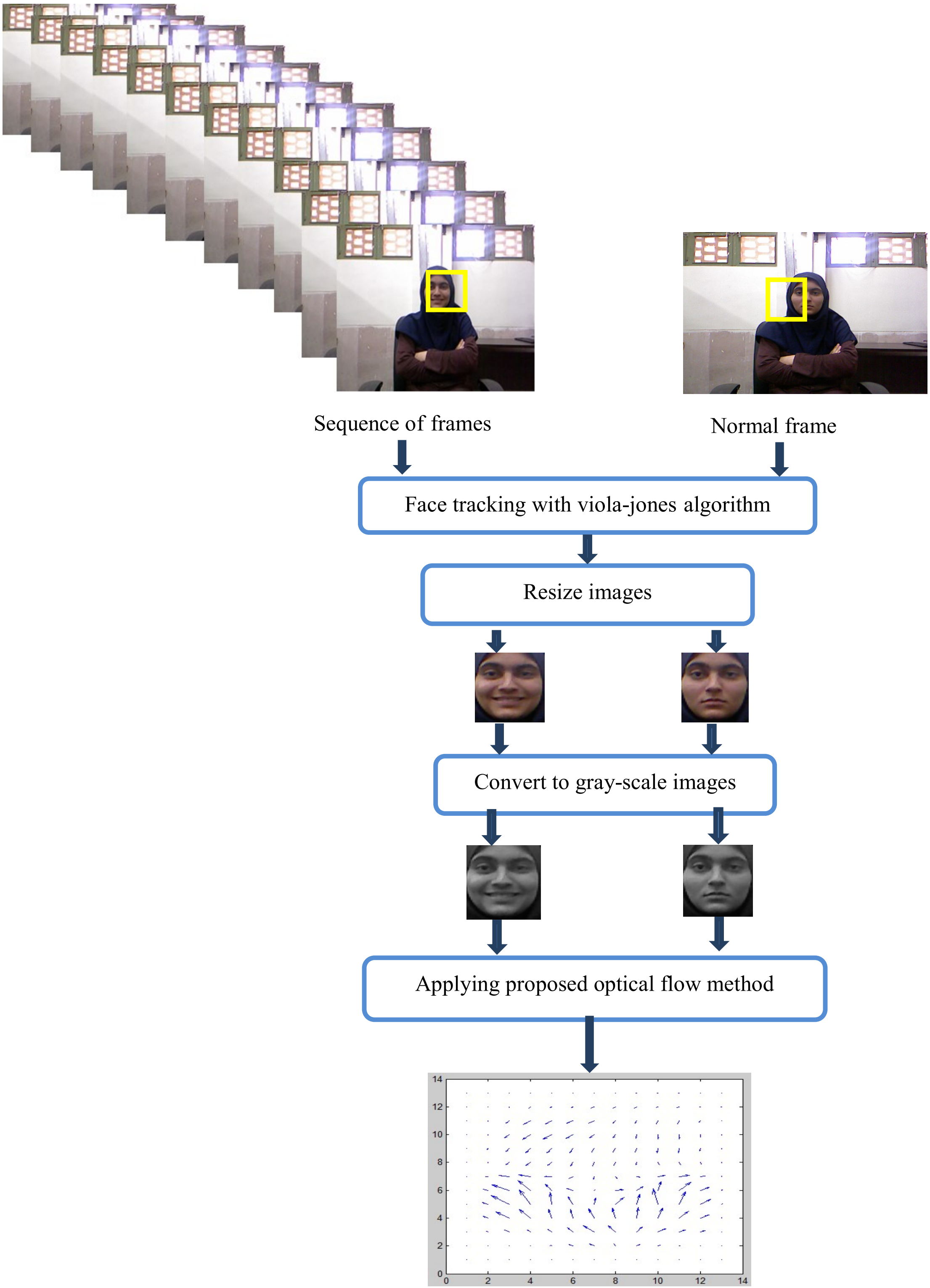
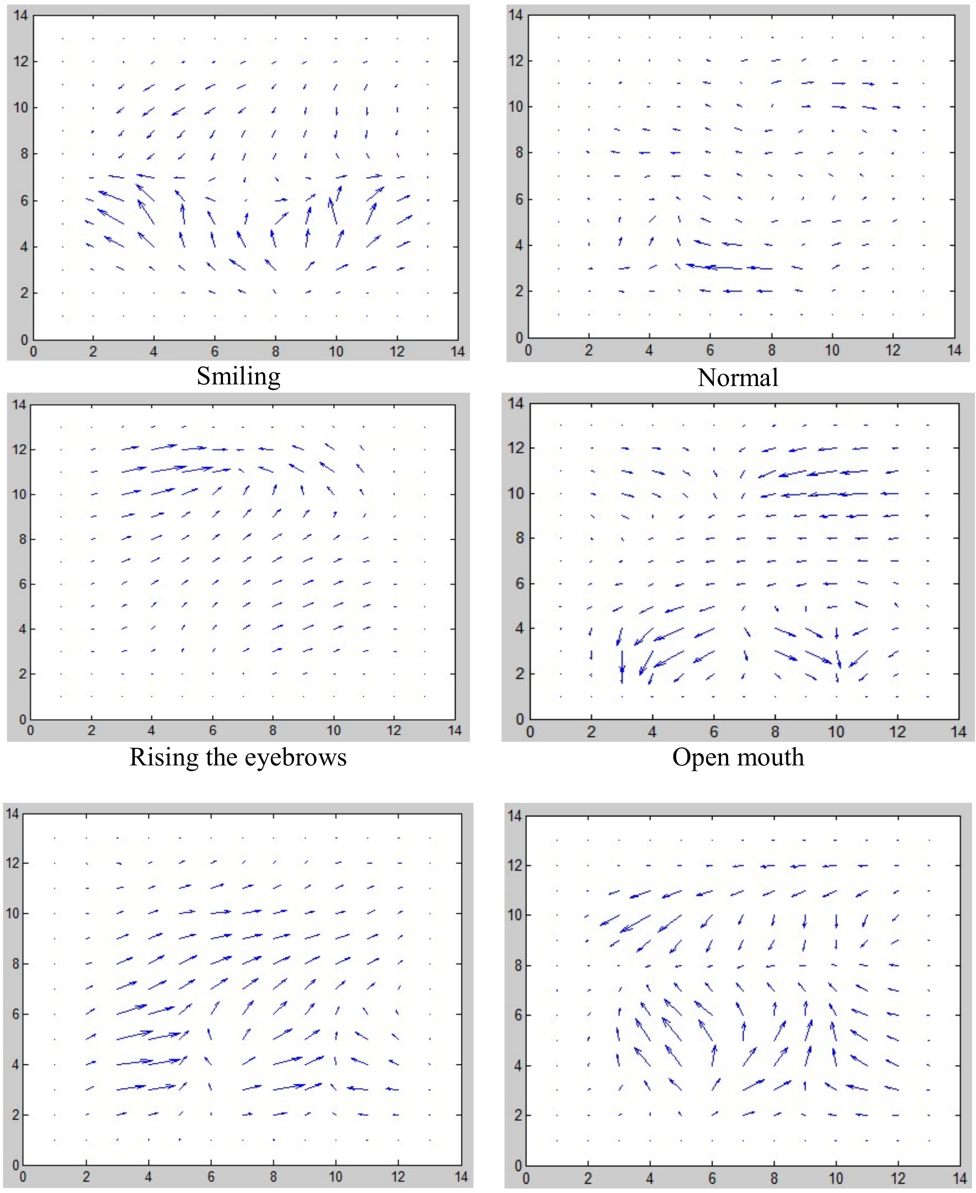

Sequence of frames
Normal frame
Face tracking with viola-jones algorithm
Resize images
Convert to gray-scale images
Applying proposed optical flow method
0
2
4
6
8
10
12
14

Figure 2. Facial expression recognition in proposed method

4.3. Feature extraction

Optical flow method is used for feature extraction. It should be noted that since the data are started from the normal state, instead of using the optical flow distinctions of two subsequent frames, the optical flow distinction of all images with the normal image is taken into consideration. In two consecutive frames, differences of all the pictures with the normal picture are taken into account. Figure 3 shows feature vectors of facial expression of our database. Matrices ‘U’ and ‘V’ values that are obtained from this algorithm are used as feature vectors. The ‘U’ matrix represents the position and the ‘V’ matrix represents the change of direction. In the following, the proposed method is combined with some other feature extraction methods (LBP uniform approach and LBP circular approach) and the obtained results will be mentioned.



Smiling

Normal

Rising the eyebrows

Open mouth

Pursing lips

Anger

Figure 3. Feature vectors of facial expression in database

The original LBP operator labels the pixels of an image by means of decimal numbers called Local Binary Patterns or LBP codes, which encode the local structure around each pixel. It proceeds thus as illustrated in figure 4: Each pixel is compared with its eight neighbors in a 3x3 neighborhood by subtracting the center pixel value. The resulting strictly negative values are encoded with 0 and the others with 1. A binary number is obtained by concatenating all these binary codes in a clockwise direction starting from the top-left one and its corresponding decimal value is used for labeling. The derived binary numbers are referred to as Local Binary Patterns or LBP codes.

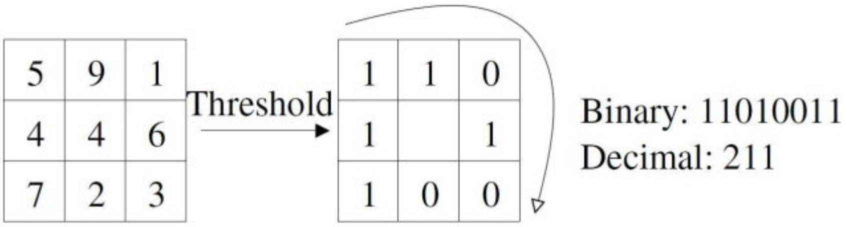


Figure 4. An example of the uniform LBP operator (Huang *et al*., 2011**)**

Formally, given a pixel at (x_c ,y_c), the resulting LBP can be expressed in decimal form as:

$$LBP(x_c ,y_c) = \sum_{i=0}^{7} s(g_i - g_c)2^i \quad (1)$$

Where g_c and g_i are respectively gray-level values of the central pixel and surrounding pixels in the 3x3 neighborhood, and function s(x) is defined as:

$$s(x,y) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2)$$

One limitation of the basic LBP operator is that its small 3x3 neighborhood cannot capture dominant features with large scale structures. To deal with the texture at different scales the operator was later generalized to use neighborhoods of different sizes. A local neighborhood is defined as a set of sampling points evenly spaced on a circle which is centered at the pixel to be labeled. The sampling points that do not fall within the pixels are interpolated using bilinear interpolation, thus allowing for any radius and any number of sampling points in the neighborhood. Figure 5 shows some examples of the extended LBP operator where the notation (P, R) denotes a neighborhood of P sampling points on a circle of radius of R.

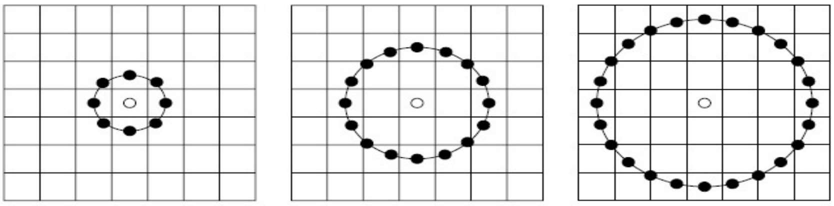


Figure 5. Examples of the circular LBP (Huang *et al*., 2011**)**

Formally, when given a pixel at (x_c ,y_c), the resulting LBP can be expressed in decimal form as:

$$LBP_{P,R}(x_c ,y_c) = \sum_{p=0}^{p-1} s(i_p - i_c)2^p \quad (3)$$

where i_c and i_p are respectively gray-level values of the central pixel and P surrounding pixels in the circle neighborhood with a radius R, and function s(x) is defined as (Huang *et al*., 2011; Liu *et al*., 2011):

$$s(x,y) = \begin{cases} 1 & x \geq 0 \\ 0 & x \prec 0 \end{cases} \quad (4)$$

4.4. Mode Classification

After the feature extraction stage, neural network is used for classifying the modes. In this study, the utilized expressions are normal, smiling, open mouth, rising the eyebrows, anger and pursing modes. In fact, they are some selective modes for face movements. It should be noted that the modes can be increased but in this case we work with these six modes. This paper used three layers feed-forward neural network (Figure 6). The proposed neural network includes 800 nodes for the input layer (400 nodes for U matrix and 400 nodes for V matrix), 100 nodes for the hidden layer and 6-nodes for output layer. From the collected data 70% are used for training, 15% for validation and the last 15% are used to evaluate the neural network.

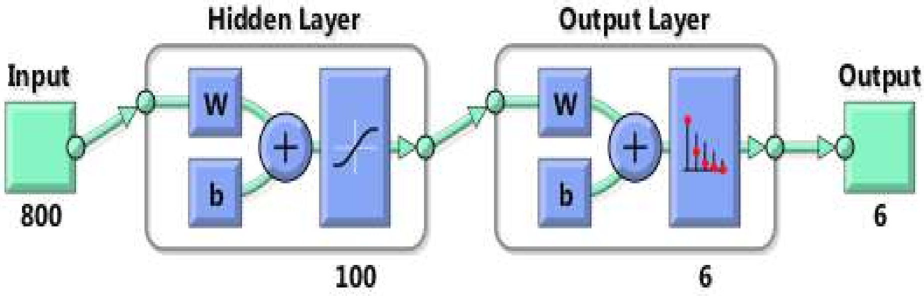


Figure 6. Proposed feed-forward neural network classifier

4.5. 3D model Generation

Face region is separated precisely from video frames by using a segmentation method based on skin color. The depth data corresponding to this separated area is taken for a 3D representation from depth data corresponding to each frame. At the end, a file is prepared for each frame consisting of face points with 6 features: X, Y, depth, red, green and blue color. These data are used for producing a 3D model and a graphical avatar for each frame (Figure 7). Figure 8 shows 3D model of some facial expressions.

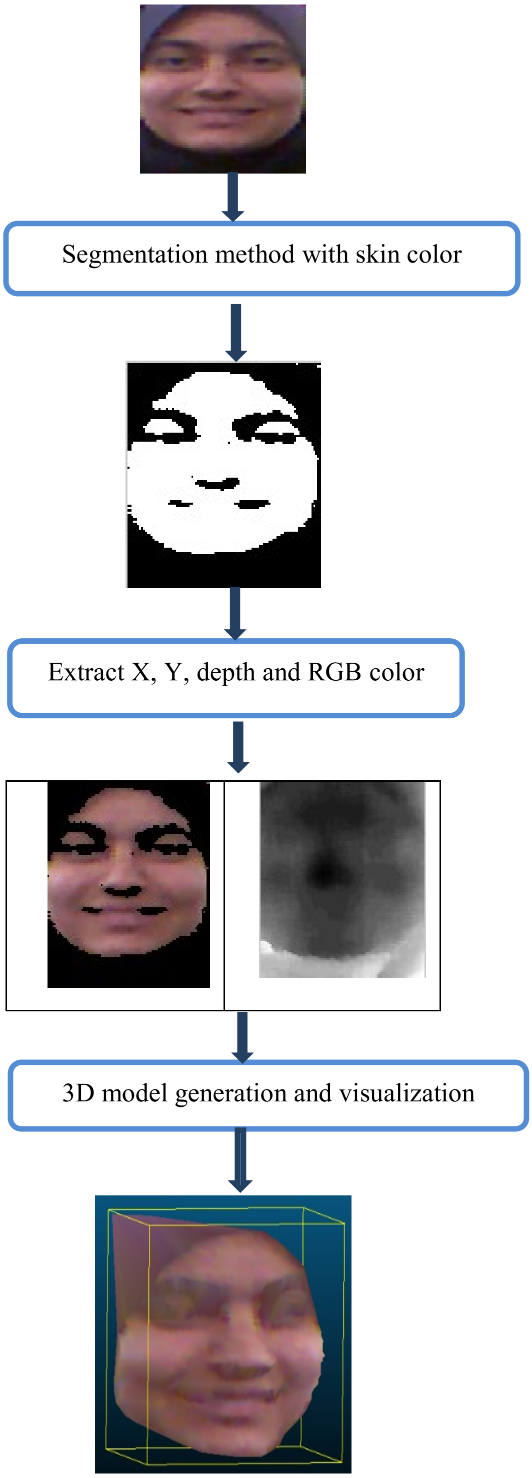
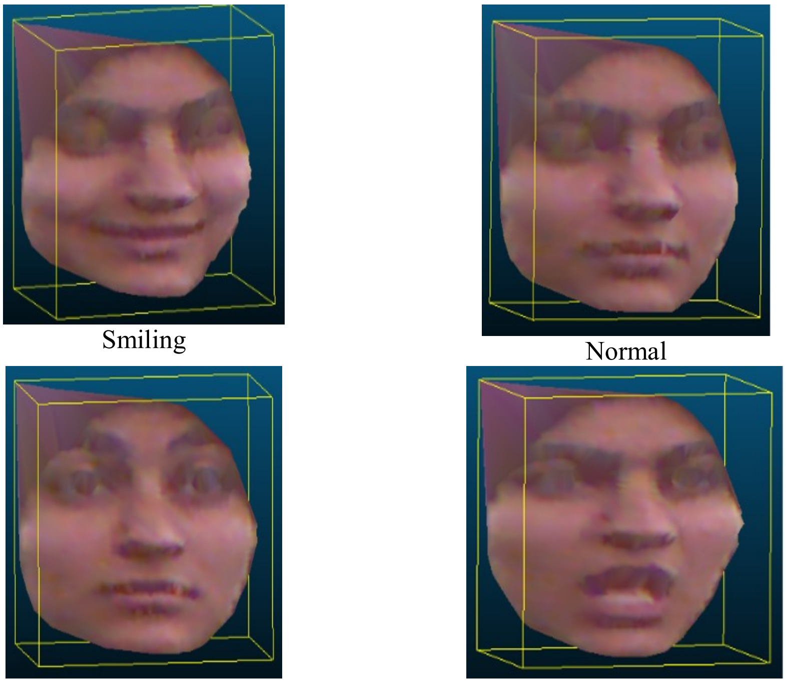

Segmentation method with skin color
Extract X, Y, depth and RGB color
3D model generation and visualization

Figure 7. Avatar 3D model generation



Smiling

Normal

Rising the eyebrows

Open mouth

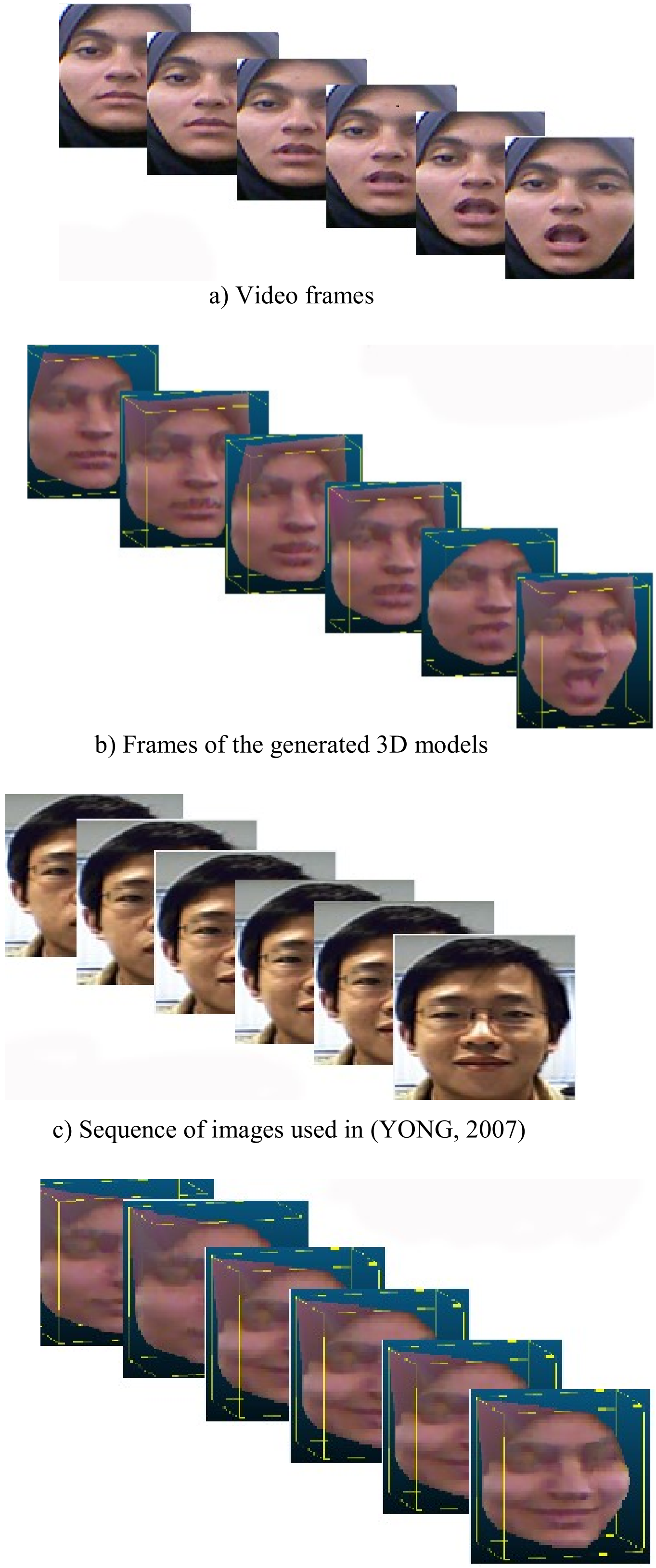
Figure 8. 3D model of some facial expressions

5. The experimental results

As previously described, the matrices U and V obtained from optical flow algorithm are used as feature vectors and feed-forward neural network is used for classification. For this purpose, 11008 data are used for training the neural network, and out of these data, 4449 are normal data, 1247 smiling, 2473 data for open mouth, 1658 data for rising the eyebrows, 643 data for anger and 538 data for pursing lips mode.

These data are obtained from three different persons with different distances (.5-1.5 meters) from Kinect camera and in different lightning (different hours of a day and changing the light situation). Five people (three people who took part in the training procedure and two persons who did not take part in the training procedure are called ‘seen’ and ‘unseen’ face, respectively) are used for evaluating this network. For simplicity and brevity, specific numbers are used to illustrate different expressions. Namely number 1 is used for normal mode, number 2 for smiling, number 3 for open mouth, number 4 for rising the eyebrows, number 5 for anger and number 6 for pursing lips mode. Tables 1 and 2 depict numerical results on various videos. The asterisk* sign is used for denoting unseen faces. It should be noted that the video lengths are different.

In test procedures, single video feature vectors consisting of different expressions are given to the neural network and the network produces the corresponding labels for each frame as output. If there is a mode in a video which is not available in the data base, the nearest available mode's label to this mode is produced. For example, in test3 and test6 videos, the surprise expression (that have been showed with number 7) is recognized as open mouth expression. At the end, considering the certain numbers of subsequent similar labels (at least 10 frames, because the minimum number of one modes' frames is related to “rising the eyebrow” mode that takes 10 frames), the expressions are detected, and a 3D show of these expressions are represented. For instance, in test8 videos that have been obtained from unseen face, the “smiling” and “open mouth” expressions are well recognized, but expressions related to rising the eyebrows are not detected properly and all the corresponding frames to this expression are regarded as normal expression. Figure 9 shows example of generated 3D models.



a) Video frames

b) Frames of the generated 3D models

c) Sequence of images used in (YONG, 2007)

d) Sequence of generated 3D models based on images in c

Figure 9. Results of our facial motion capture system(a,b,c,d)

Table 1. Experimental results for seen faces

Test	Facial expressions in video	Detected facial expressions
Test1	1 2 1 2 1 2 1 2 1 3 1 3 1 3 1 3 1 4 1 4 1 4 1 4 1	1 2 1 2 1 2 1 2 1 3 1 3 1 3 1 3 1 4 1 4 1 4 1 4 1
Test2	1 2 1 2 1 3 1 3 1 4 1 4 1	1 2 1 2 1 3 1 3 1 4 1 4 1
Test3	1 2 1 2 1 3 1 3 1 4 1 4 1 7	1 2 1 2 1 3 1 3 1 4 1 4 1 3
Test4	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 5 1 5 1 5 1 6 1 6	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 5 1 5 1 5 1 6 1 6
Test5	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 6 1 5 1 5 1	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 6 1 5 1 5 1
Test6	1 2 1 3 1 4 1 5 1 6 1 7 1	1 2 1 3 1 4 1 5 1 6 1 3 1

Table 2. Experimental results for unseen faces

Test	Facial expressions in video	Detected facial expressions
Test7 *	1 2 1 2 1 3 1 3 1 4 1 4 1 5 1 6	1 2 1 2 1 3 1 3 1 4 1 4 1 5 1 6
Test8*	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 4 1	1 2 1 2 1 2 1 3 1 3 1 3 1 1 1 1 1 1 1 1

Table 3 shows the mode detection accuracy of the proposed method and its combination with two other methods (uniform LBP and circular LBP) for different people. The overall accuracy of the proposed procedure is calculated as this way one video is chosen as input, and after mode detection the three aforementioned steps are applied on this video. The obtained feature vectors are given to the neural network and the corresponding labels to each frame are regarded as output. Afterwards, the overall accuracy is calculated from the confusion matrix. However, it should be noted that the expression detection criteria are the observation of a certain number of subsequent similar labels and in the case of observing a limited or sparse number of different labels the final label would not change. Figure 10 and 11 show result of different methods for seen and unseen data respectively. Table 4 shows results for seen data with proposed method and uniform LBP.

Table 3. Overall accuracy of neural network using different methods

Test	proposed method + uniform LBP	proposed method + circular LBP	proposed method
Test1	96.7	95.8	96.3
Test2	95.9	96.9	94.4
Test3	88.3	81.3	93.8
Test4	91	92	95.2
Test5	94.7	92.6	96.3
Test6	92.6	92	92.3
Test7*	47.2	82.1	92.7
Test8*	55.9	84.7	79.1

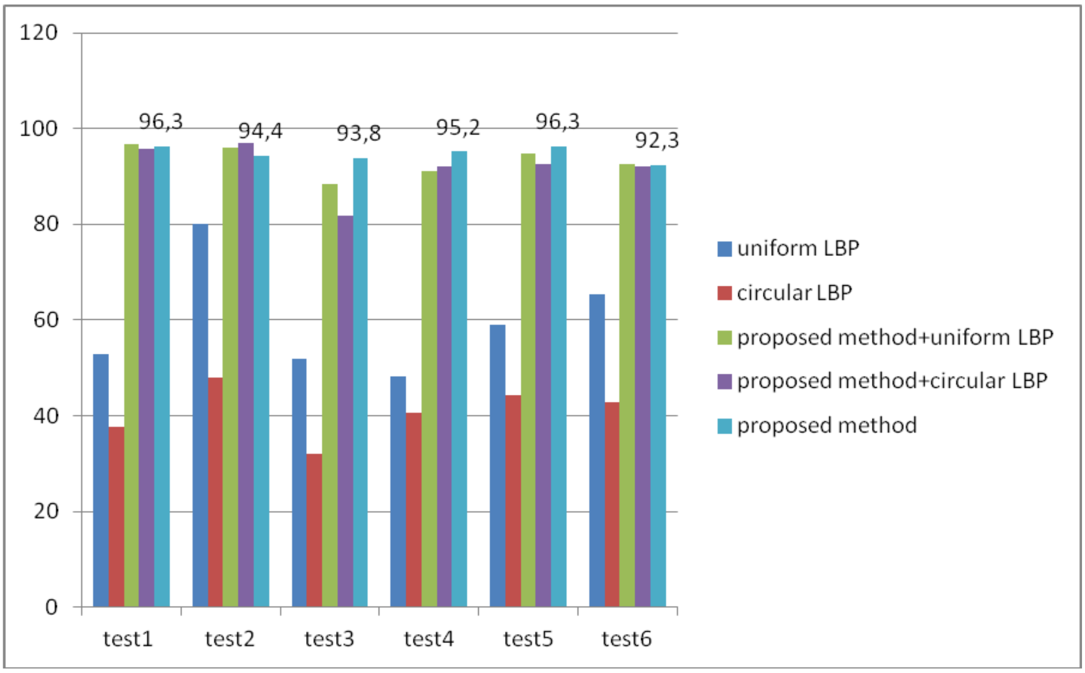


Figure 10. Accuracy of different method for seen faces

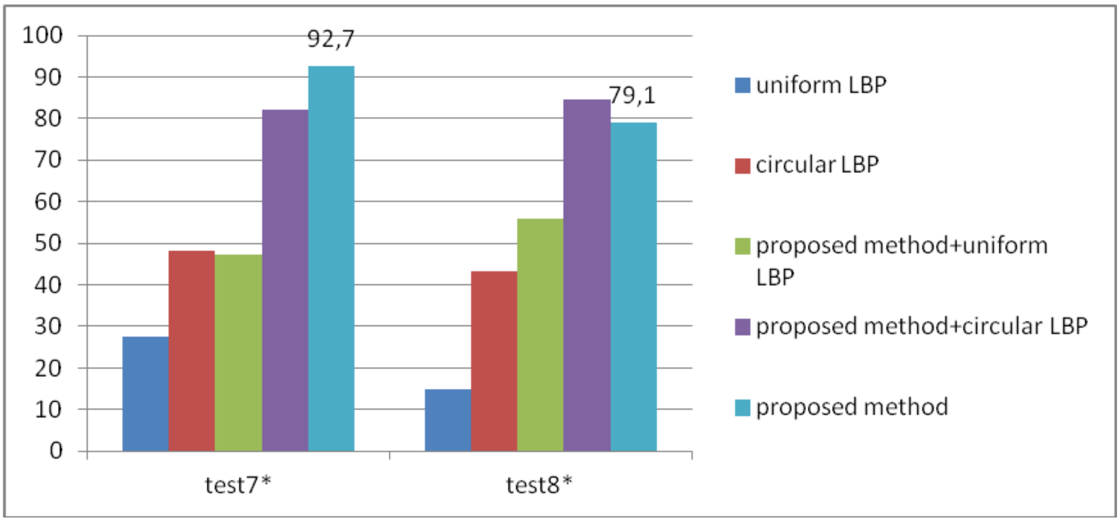


Figure 11. Accuracy of different method for unseen faces

Table 4. Results of frame detections for proposed method +uniform LBP

Test	Facial expressions in video	Detected facial expressions
Test1	1 2 1 2 1 2 1 2 1 3 1 3 1 3 1 3 1 4 1 4 1 4 1 4 1	1 2 1 2 1 2 1 2 1 3 1 3 1 3 1 3 1 4 1 4 1 4 1 4 1
Test2	1 2 1 2 1 3 1 3 1 4 1 4 1	1 2 1 2 1 3 1 3 1 4 1 4 1
Test3	1 2 1 2 1 3 1 3 1 4 1 4 1 7	1 2 1 2 1 3 1 3 1 4 1 4 1 3
Test4	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 5 1 5 1 5 1 6 1 6	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 5 1 5 1 5 1 6 1 6
Test5	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 6 1 5 1 5 1	1 2 1 2 1 2 1 3 1 3 1 3 1 4 1 4 1 4 1 6 1 5 1 5 1
Test6	1 2 1 3 1 4 1 5 1 6 1 7 1	1 2 1 3 1 4 1 5 1 6 1 3 1

6. Conclusion

The aim of this research is to capture face motion by means of color and depth data obtained from Kinect camera. In fact, in a video, a facial expression starts from normal state goes to peak state and then returns to normal mode. The first task for obtaining the facial expressions is removing the face region from other components in each frame. In this study, Viola and Jones algorithm is used to separate the face region from other components and optical flow method is used for feature extraction. In the proposed optical flow method, the distinctions between all frames with the normal frame is taken into consideration, rather than using the distinctions between two consecutive frames. The obtained matrices are used as feature vectors and feed-forward neural network is used for classifying the modes. In the final stage, after the discovery of available modes in video, a 3D show of facial expressions are represented that it can be used to create 3D animations and interactive digital games. Based on the results of the present study, our system is able to recognize the facial expressions of different unseen people, but considering the existing hardware condition, it would not be able to work at real time.

References

Skogstad, S. A. van D., Jensenius, A. R., & Nymoen, K. (2010). "*Using IR optical marker based motion capture for exploring musical interaction*". Proceedings of the conference on new interfaces for musical expression, 407–410.

Lindequist, J., & Lönnblom, D. (2004). "*Construction of a Motion Capture System*". School of mathematics and systems engineering.

Li, D., Sun, C., Hu, F., Zang, D., Wang, L., & Zhang, M. (2013). "*Real-time performance-driven facial animation with 3ds Max and Kinect*". 3rd International Conference on Consumer Electronics, Communications and Networks, 473–476.

Michel, P., & El Kaliouby, R. (2003). "*Real time facial expression recognition in video using support vector machines*". Proceedings of the 5th international conference on Multimodal interfaces, 258–264.

Ge, S. S., Wang, C., & Hang, C. C. (2008). "*Facial expression imitation in human robot interaction*". 17th IEEE International Symposium on Robot and Human Interactive Communication, 213–218.

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2014). "*Facewarehouse: A 3d facial expression database for visual computing*". IEEE Transactions on Visualization and Computer Graphics 20, 3, 413-425.

Seddik, B., Maamatou, H., Gazzah, S., Chateau, T., & Ben Amara, N. E. (2013). "*Unsupervised facial expressions recognition and avatar reconstruction from Kinect*". 10th International Multi-Conference on Systems, Signals & Device, 1–6.

Nourani-Vatani, N., Borges, P. V. K., & Roberts, J. M. (2012). "*A study of feature extraction algorithms for optical flow tracking*". Proceedings of Australasian Conference on Robotics and Automation.

Rangari, S., & Gonnade, S. (2014). "*A survey on Facial Expression Recognition*". International Journal of Research in Advent Technology, 2(9).

Viola, p., & Jones, M. J. (2004). "*Robust real time face detection*". International Journal of Computer Vision, 57(2), 137-154.

Vinay, B. (2012). "*Face expression recognition and analysis: the state of the art*". arXiv preprint arXiv:1203.6722.

Dhawan, S., & Dogra, H. (2012). "*Feature Extraction Techniques for Face Recognition*". International Journal of Engineering, Business and Enterprise Applications, 1–4.

Yen, G. G., & Nithianandan, N. (2002). "*Facial feature extraction using genetic algorithm*". Proceedings of the Congress on Evolutionary Computation, 2, 1895–1900.

Bhatt, B. G., & Shah, Z. H. (2011). "*Face feature extraction techniques : a survey*". National conference on recent trends in engineering & technology.

Bagherian, E., Rahmat, R. W., & Udzir, N. I. (2009). "*Extract of Facial Feature Point"*. International Journal of Computer Science and Network Security, 9(1), 49-53.

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). "*An overview of principal component analysis*". Journal of Signal and Information Processing, 4(3), 173–175.

Kaur, A., & Sarabjit Singh, T. (2015). "*Face Recognition Using PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) Techniques*". International Journal of Advanced Research in Computer and Communication Engineering, 4(3), 308-310.

Huang, D., Member, S., Shan, C., Ardabilian, M., Wang, Y., & Chen, L. (2011). "*Local binary patterns and its application to facial image analysis : a survey*". IEEE transactions on systems man and cybernetics part c (applications and reviews), 41(6), 765–781, IEEE, 2011.

Liu, W., Wang, Y., & Li, S. (2011). "*LBP feature extraction for facial expression recognition local binary patterns (LBP)*". Journal of Information & Computational Science, 8(3), 412–421.

YONG, Y. (2007). *Facial expression recognition and tracking based on distributed locally linear embedding and expression motion energy*. National University of Singapore, Ph.D. Diss. 145p.

Mehdi REZAEIAN (b. Sep. 22, 1970) has received his B.Sc. degree in electronics engineering in 1994 from KNT University, Tehran, Iran. He received his M.Sc. degree in Biomedical engineering from University of Tehran in 1997. In 1999, he joined the faculty of engineering at University of Tehran for five years. In 2010, he received his Ph.D. degree from Swiss Federal Institute of Technology (ETH) in Zurich. Since 1998, his main research is in image processing and computer vision topics. Now, he is an assistant professor of Department of Computer Engineering at Yazd University, Iran. He published more than 50 papers in journals and conferences.

Fateme Zare MEHRJARDI (b. May 14, 1991) received his BSc in computer engineering (2012), MSc in Artificial Intelligence (2014) from Yazd University of Iran. Her current research interests include different aspects of Artificial Intelligence applied in Image processing, Pattern recognition and Machine learning.