2014 13th International Conference on Machine Learning and Applications

Facial expression recognition using kinect depth sensor and convolutional neural networks

Earnest Paul Ijjina
Ph.D Research Scholar
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: cs12p1002@iith.ac.in

C Krishna Mohan
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: ckm@iith.ac.in

***Abstract*—Facial expression recognition is an active area of research with applications in the design of Human Computer Interaction (HCI) systems. In this paper, we propose an approach for facial expression recognition using deep convolutional neural networks (CNN) based on features generated from depth information only. The Gradient direction information of depth data is used to represent facial information, due its invariance to distance from the sensor. The ability of a convolutional neural networks (CNN) to learn local discriminative patterns from data is used to recognize facial expressions from the representation of unregistered facial images. Experiments conducted on EURECOM kinect face dataset demonstrate the effectiveness of the proposed approach.**

***Keywords*-Facial expression recognition; convolutional neural networks (CNN);**

I. Introduction

Face is the index of mind and actions speak louder than words. One of the active areas of research in computer vision is human behavioral analysis. The key elements in human behavioral analysis include body language, hand gestures and facial expressions which play a crucial part in semantic analysis of multimedia content. The spatio-temporal variation of facial muscles results in the creation of various facial expressions that convey emotions [1]. Paul Ekman's research confirms that the 43 facial muscles can create 10,000 unique facial expressions. The broad vocabulary of facial expressions makes it an ideal candidate for human computer interaction. Some of the application of automatic facial expression recognition include clinical psychology, pain assessment, lie detection and multimodal human computer interface. This paper tries to address the problem of facial expression recognition by considering 1) kinect depth sensor for the acquisition of facial information 2) depth information for representation of facial expression and 3) convolutional neural network for facial expression recognition.

In the literature, most of the existing approaches use facial information acquired using a RGB camera. Some of the other approaches use 3D scanning [2] and thermal imaging [3] to acquire the facial information. Though facial information acquired using 3D scanning is accurate and invariant to illumination conditions when compared to other approaches, it requires specialized expensive equipment and controlled capturing environment. As illumination conditions affect facial information captured using an RGB camera, a low cost RGB-D camera like kinect sensor is the most suitable choice for capturing facial information.

Over the years several facial recognition approaches were proposed using various features and recognition techniques. Chun Fui Liew *et al.* [4] evaluated five most commonly used feature spaces with seven classification methods to identify the most effective features for facial expression recognition. Xiaohua *et al.* [5] proposed the use of Weber Local Descriptor (WLD) and Histograms of Oriented Gradients (HOG) to capture local features in the sub-regions of an image and used chi-square distance measure to identify the facial expression. Yongqiang *et al.* [6] proposed the use of temporal evolution of relationships among different facial activities to recognize facial expressions using Dynamic Bayesian Network. Taheri *et al.* [7] considers an expressive face as a superposition of natural face and a sparse facial expression component and uses a dictionary-based component separation algorithm to recognize the facial expressions. Xiaoli *et al.* [8] proposed the transformation of 3D facial information into 2D range image for representation and analyzed various approaches for generation of range image and features for 3D facial expression recognition. Inchul Song *et al.* [9] used deep convolutional neural network with 5 layer to develop a face expression recognition system for smart-phone. Samira Ebrahimi Kahou *et al.* [10] used deep convolutional neural network to analyze facial expression in video frames and other neural network architectures to recognize emotions in videos.

Most of the existing approaches rely on RGB information to recognize the most common facial gestures like normal, smile and disgust. We use depth information to recognize facial expression like *open mouth*, *occlusion of mouth by hand* and *occlusion by paper* from unregistered facial im-

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.70

ages. Over the years, CNNs were used for various computer vision tasks [11]. In this paper, we try to address the problem of facial expression recognition using CNN and features generated from depth information only. The rest of the paper is organized as follows: section II outlines some of the challenges associated with the use of depth information for facial expression recognition. Section III describes the use of CNN architecture for facial expression recognition and section IV presents the experimental results. The last section provides conclusions and future work.

II. Depth information for facial expression recognition

In contrast to the RGB information which is most widely used for facial expression recognition, depth information is insensitive to illumination conditions. Thus, an approach relying only of depth information for facial expression recognition is inherently tolerant to illumination variations. One of the most commonly used depth sensor in research community, gaming and entertainment industry is the kinect sensor. So far, many computer vision tasks were able to leverage the depth information captured using kinect sensor to improve the performance of various computer vision tasks [12]. Some of the challenges involved in the use of depth information generated from a kinect depth sensor for facial expression recognition are : 1) low accuracy of depth sensor affects the quality of facial information captured 2) the accuracy on left side of the captured frame is low compared to the right side due to the use of a single sensor 3) the random noise in the captured depth information due to environmental and other factors and 4) ambiguous edges of objects.

Figure 1 shows the RGB, depth and gradient direction information for the nine facial expressions: a) *neutral* b) *smile* c) *open mouth* d) *strong illumination* e) *occlusion of eyes by sun-glasses* f) *occlusion of mouth by hand* g) *occlusion by paper* h) *left profile*, and i) *right profile* of a subject in EURECOM kinect face dataset [13]. The gradient direction images shown in the last row of the figure shows visual evidence confirming the afore mentioned challenges involved in the use of depth information for facial expression recognition. Thus, only a limited set of expressions can be recognized using depth information. The expressions in EURECOM kinect face dataset that can be recognized using depth information only are: *neutral*, *open mouth*, *occlusion of mouth by hand*, *occlusion by paper*, *left profile*, and *right profile*. As depth information is independent of variation in illumination and color of the (reflecting/refracting) surface one cannot distinguish between the expressions *neutral*, *strong illumination*. Also, the expression *neutral* and *occlusion of eyes by sun-glasses* cannot be distinguished when the subject wears glasses like the one shown in Figure 1. Smile expression cannot be recognized accurately due to the low accuracy of depth information captured using Kinect sensor as shown in Figure 1(b).

As gray value in a depth image is indicative of the distance of a particular point from the depth camera i.e., white value indicates a point closest to the camera and a black point indicates the farthest point from the camera. Thus, the subject's distance from the camera affects the corresponding facial depth information captured. Most of the facial datasets contain scaled, aligned faces in the same pose. The facial expression images in EURECOM kinect face dataset are not aligned, scaled and have minor facial pose and tilt, making facial expression recognition more challenging. The physical variations of the subjects (in terms of hair style, shape of face etc.,) and the inconsistency in the execution of expressions like use of left/right hand for *occlusion of mouth* adds more complexity to the recognition approach. The following section describes how these challenges are addressed in the proposed approach.

III. Proposed approach

The architecture of the proposed approach consists of a preprocessing step, that generates the representation of facial image used in the proposed approach from depth information generated by a kinect sensor as shown in Figure 2. This facial representation is used by deep convolutional neural network classifier to recognize the facial expressions. The following sub-section describes the steps involved in preprocessing of kinect depth information.

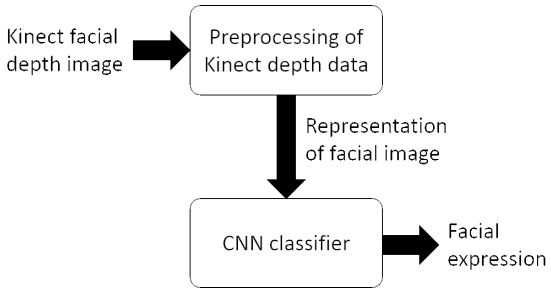


Figure 2. Overall architecture of the proposed approach

A. Pre-processing of kinect depth images

During preprocessing, the facial depth data generated by a kinect sensor is processed to generate the facial representation of input data, by normalizing the variation in subjects distance from the kinect sensor using gradient direction and by background subtraction as shown in Figure 3. Figure 4 shows the typical (a) to (d) images used and generated during the pre-processing of kinect facial depth data given in Figure 3.

The masking of gradient direction information eliminates the noise because of background, thereby improving the

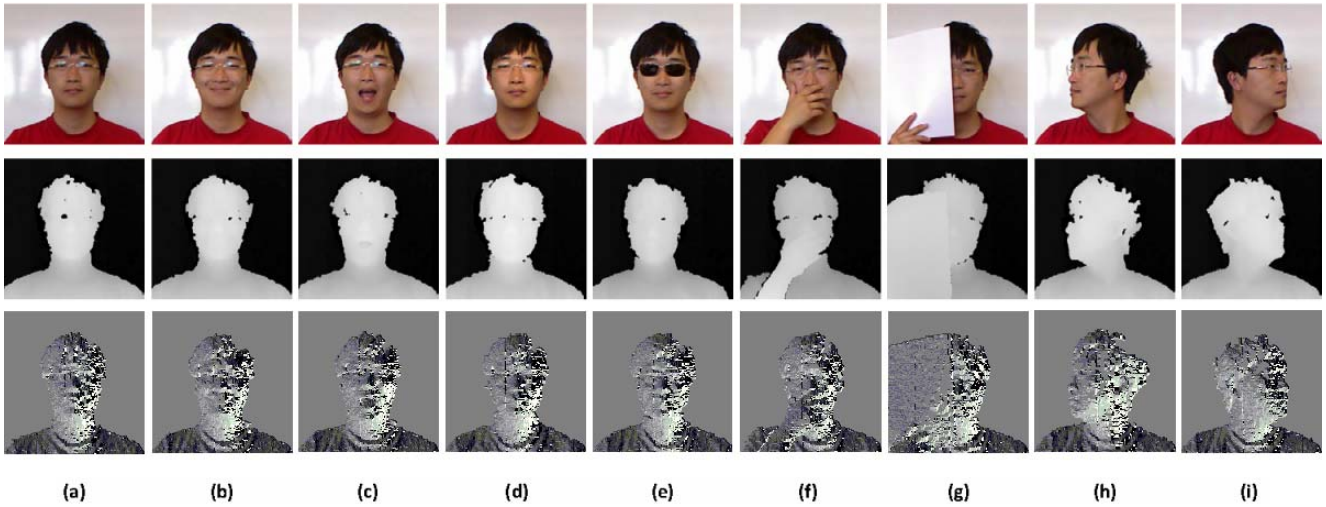


Figure 1. RGB, depth and gradient direction information for facial expression in EURECOM kinect face dataset

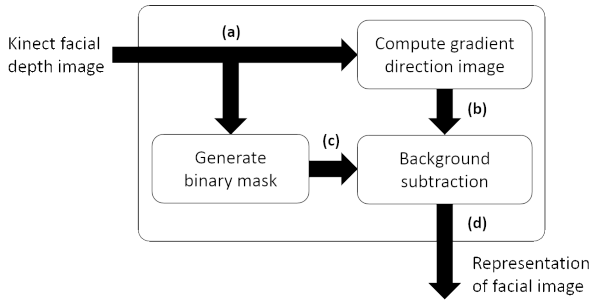
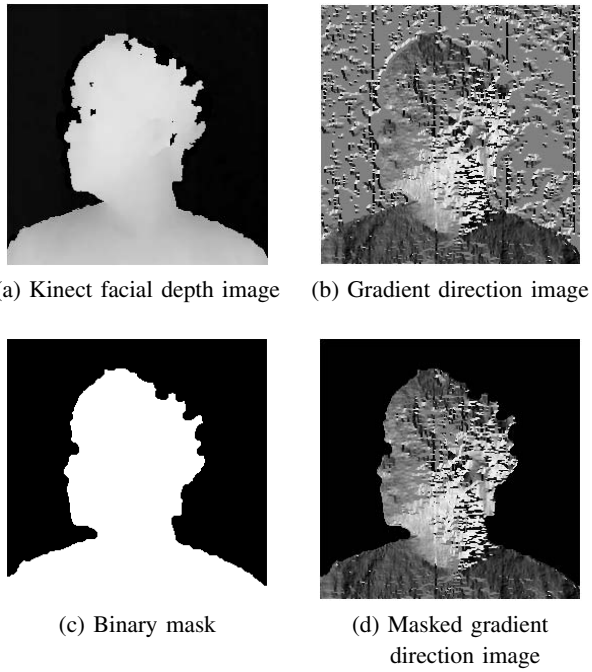


Figure 3. Preprocessing of facial depth data to generate facial representation



(a) Kinect facial depth image (b) Gradient direction image

(c) Binary mask (d) Masked gradient direction image

Figure 4. Images used and generated during pre-processing of kinect depth facial images

performance of the proposed approach. The use of deep convolutional neural network architecture for facial expression recognition using this masked gradient direction image is explained in the following subsection.

B. CNN for facial expression recognition

A deep convolutional neural network [14] is a deep neural network with alternating convolution and subsampling layers. A deep convolutional neural network can discriminate various inputs by learning the local patterns in input data. In this paper, this ability of a CNN is exploited to recognize the facial expression from the masked gradient direction image, explained in the previous subsection. The CNN classifier [15] used in the proposed approach is shown in Figure 5.

The dimensions of input data, size of templates used in convolution and sub-sampling layers, the number of feature maps used in each layer are shown in Figure 5. The 12 feature maps of size 5×5 generated by the last layer of CNN are traversed in row major order resulting in a column vector of dimension 300×1, which is used by the neural network to assign a label (facial expression) to the input facial image. Back-propagation in batch mode is used to train the CNN classifier. The details of the experimental setup and results are explained in the following subsection.

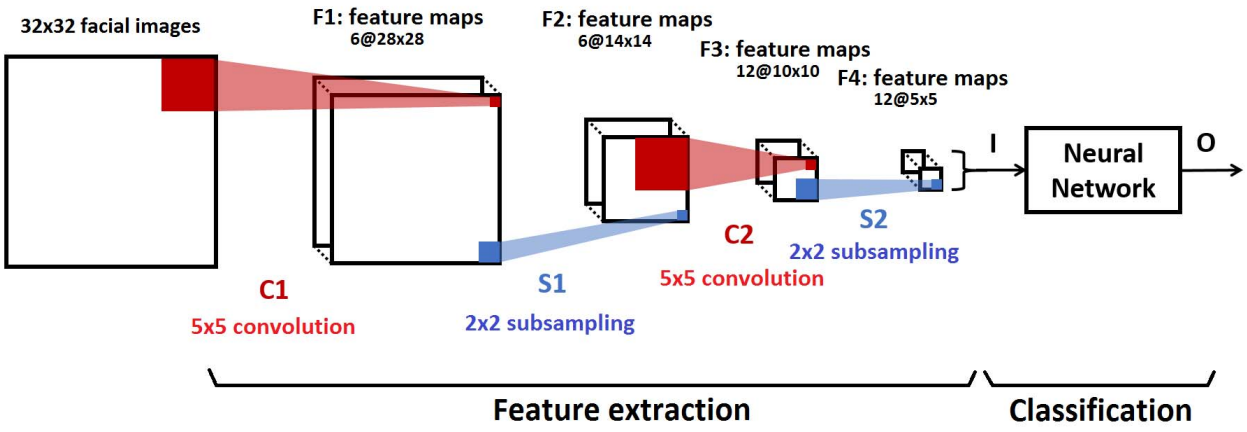


Figure 5. Architecture of deep convolutional neural network classifier

C. Experimental setup and results

The experiments were conducted on EURECOM Kinect face dataset [13] due to the availability of wide range of facial expressions that can be recognized using depth information. The dataset consists of RGB-D images of 52 (14 female, 38 male) subjects with 9 states of facial expressions per person captured in two sessions, resulting in a total of 936 images of size 256×256. The actions considered in this approach are: *occlusion paper, open mouth, occlusion mouth, right profile, left profile* and *neutral*. These facial images are preprocessed as explained in section III-A. The resulting 256×256 are down-sampled to 32×32 and given as input to the deep convolutional neural network classifier, described in the previous section for classification. Two-fold cross-validation across sessions is used to evaluate the performance of the proposed approach. The CNN classifier is trained in batch mode with a batch size of 24 for 1000 epochs to obtain an average classification performance of 87.98%. The confusion matrix of the proposed approach is shown in Figure 6 and the plot of classification error vs training iteration is shown in Figure 7.

The performance of the proposed approach for recognition of *occlusion paper, occlusion mouth, left profile* and *right profile* facial expressions is high when compared with the remaining expressions due to unambiguity of evidential information to recognize these expression. The low recognition performance for *open mouth* and *neutral* facial expressions is due to unregistered facial images in EURECOM kinect face dataset.

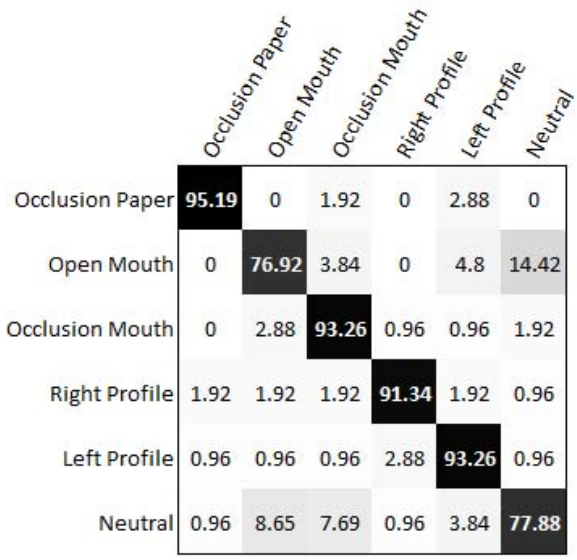


Figure 6. Confusion matrix of the proposed approach

IV. Conclusion

In this paper, we propose an approach for facial expression recognition from unregistered facial depth images captured by a kinect sensor, using deep convolutional neural network. The invariance of depth information to illumination variations and the ability of a convoutional neural network to recognize local patterns in input are exploited to recognize facial expression from depth information. In spite of the high noise in depth information and the use of unregistered facial images, the proposed approach is able to accurately recognize the facial expressions. The future work includes exploration of other approaches for pre-process the depth information and other machine learning approaches for classification.

References

[1] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression

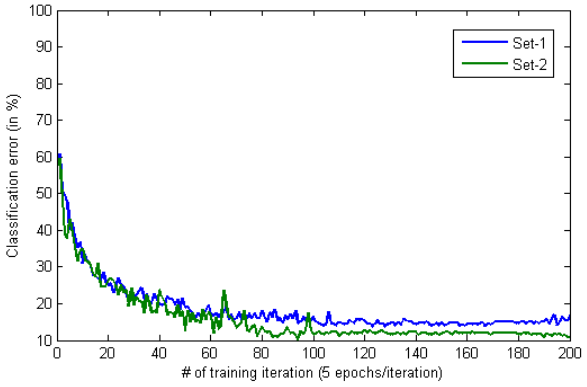


Figure 7. Plot of classification error vs training iteration

recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3422–3429.

[2] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image Vision Comput.*, vol. 30, no. 10, pp. 683–697, 2012.

[3] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, Nov 2010.

[4] C. F. Liew and T. Yairi, "A comparison study of feature spaces and classification methods for facial expression recognition," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2013, pp. 1294–1299.

[5] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. Ren, "Feature fusion of hog and wld for facial expression recognition," in *2013 IEEE/SICE International Symposium on System Integration (SII)*, Dec 2013, pp. 227–232.

[6] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2559–2573, July 2013.

[7] S. Taheri, V. Patel, and R. Chellappa, "Component-based recognition of facesand facial expressions," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 360–371, Oct 2013.

[8] X. Li, Q. Ruan, and G. An, "Analysis of range images used in 3d facial expression recognition," in *TENCON 2013 - 2013 IEEE Region 10 Conference (31194)*, Oct 2013, pp. 1–4.

[9] I. Song, H.-J. Kim, and P. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2014, pp. 564–567.

[10] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 543–550.

[11] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2010, pp. 253–256.

[12] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[13] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data," in *Proceedings of the 11th International Conference on Computer Vision - Volume Part I*, ser. Asian Conference on Computer Vision (ACCV) 2012. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 133–145.

[14] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *IEEE International Symposium on Circuits and Systems (ISCAS 2010)*, 2010, pp. 253–256.

[15] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, Technical University of Denmark, Asmussens Alle, Denmark, 2012.