proceedings

MDPI

Proceedings

Depth-Based Lip Localization and Identification of Open or Closed Mouth, Using Kinect 2 †

Mina Zohoorian Yazdi * and Mohsen Soryani

School of Computer Engineering, Iran University of Science & Technology, Tehran, Iran; soryani@iust.ac.ir

* Correspondence: mina.zjy93@gmail.com; Tel.: +98-915-704-4251

† Presented at the 15th International Workshop on Advanced Infrared Technology and Applications (AITA 2019), Florence, Italy, 17–19 September 2019.

Published: 23 September 2019

Abstract: A depth-based technique for lip localization and also mouth state analysis that is an important step in many applications such as lips reading, yawning detection and emotion recognition, is presented throughout this work. This is done by employing depth images captured by the Kinect V2 device. At first, using the depth information, we use the face's depth information to reduce the search area for the lips by developing a nose point detection, Second, we further reduce the search area by using a depth segmentation algorithm to separate the mouth area. Finally, with the reduced search range, we present a method for mouth state identification based on depth information. Comparing this work with other researchers' work using the databases prepared by authors and the VAP_RGBD dataset, we found that our method, which involves using of depth information, can solve the problem of varying illumination conditions. Experimental results demonstrated an accuracy of 91% for lip localization, and 86% for open or closed mouth state detection.

Keywords: Lip Segmentation; Mouth State; Depth Information; RGB_D; Kinect

1. Introduction

There are several techniques for lip area detection. A large category of techniques are model-based. In these techniques, the structure of lip area is described by a set of model parameters. These techniques include snakes, active contour models and several other parametric models [1]. Before visual features can be extracted from an input image, a robust face and lip detection is required. Also mouth analysis methods are mainly focused on lips segmentation. The automatic detection of facial features is still a work in progress. Even though numerous methods have been developed to achieve this goal but yet different problems remain. Those problems are caused by lighting condition changes, noise and the different characters of human faces [2]. Galatas et al. [3] proposes the use of Viola Jones Algorithm to detect the face and another Viola Jones Algorithm pass to localize the mouth region within the face. However, its limitation includes a constant distance between the speaker and the recording device, controlled illumination and a simple background. Similarly, Navarathna et al. [3] also uses the Viola Jones method to detect both the face and lips, but instead of utilizing the entire face image, the lower half of the face was used. Unlike the use of a simple background from Galatas et al, Navarathna et al. used images that were recorded within a car environment. Hassanat et al. [4] proposes the use of a color-based technique for lip segmentation. This study makes use of color spaces to categorize pixels as either lip or non-lip using artificial neural networks and shows a method for fusion of the existing different color spaces that comes from different color models. Lüsi et al. [5] also uses the method for the real-time imitation of lips movements of an Estonian speaker on a 3D avatar is proposed. This research is focused on obtaining the RGB and depth information from a Kinect 2 camera, processing it and moving the avatar

accordingly. For the more stable extraction of information about lip shape and movement, cosine function is used. The depth info helps to determine the closest speaker and stabilize the movement of avatar and tracking of lips. Kalbkhani et al. [6] also uses the method for lip area detection and lip segmentation in color face images based on local information. Local information is standard deviation of pixels of different parts in lower half part of face image. For detection of the lip area, an enhanced version of Lip-Map, proposed by Hsu et al. [7] has been used in this paper. This Lip-Map is multiplied by saturation component and after partitioning the lower half image into some equal parts, in order to find the lip area, standard deviation of pixels is calculated in each area. Then, for separating lip pixels from skin pixels, optimum threshold value is obtained by Otsu's method. Yingyu Ji et al. [8] proposes the use of a method for detecting the mouth state by extracting contour features. When the mouth is open, the internal contour of the mouth is extracted according to its external rectangle to define the mouth opening degree N for analyzing the mouth open state. In this paper an algorithm is presented, using Kinect camera's RGB-D data, for lip detection and investigating of the state of the mouth. The rest of this paper is structured as follows: in Section 2 the proposed approach is explained. The results are presented in Section 3 and finally Section 4 concludes the paper.

2. Proposed Approach

The proposed algorithm consists of three steps. These stages are as follows and each one will be explained in details in the following:

1 Nose tip detection
2 Lips segmentation and mouth area detection
3 Open or closed mouth detection

Figure 1 Shows the block diagram of the algorithm.

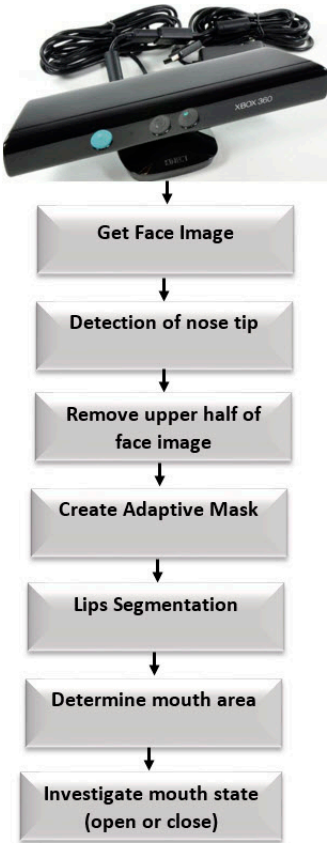

Figure 1. Block diagram of our algorithm.

2.1. Nose Tip Detection

When head rotation is limited, the nose tip pixel has the minimum depth value in the depth map. When the head rotation exceeds a specified interval, for example, the rotation angle of the head is greater than +20 or less than −20 degrees, glasses or cheeks will have the lowest depth. When the pitch angle is greater than +15 or less than −15 degrees the chin or forehead has the smallest depth.

A method for detecting nose tip has 2 steps; using the "out of range" filter to remove invalid depth data and determining the location of pixel with the lowest depth. This method can accurately detect the nose tip but due to small changes that occur in depth information because of noise, at times, the minimum depth value may be mistaken as the glasses or hair. To eliminate this unwanted change, a Border Mask with dimensions chosen according to the number of rows and columns of the depth image is multiplied by the image to remove the hair area around the face (Figure 2) [9]. Then a median filter (7 * 7) is applied to the depth image. Now, in the new image, the minimum depth value is obtained and is used as a threshold and a binary image is created. To ensure that the minimum value of the original depth image is not discarded after convolving the median filter, the surroundings of the minimum depth is also considered. A squared structural element with the same size as the median filter is used to dilate the binary image that was created. The result of the expanded binary image is multiplied by the initial depth image element by element so that a wider search area is obtained. Now, the minimum depth pixel is found on this binary image. If more than one minimum point is found, the average of the column values is obtained and the maximum row value is considered. Now, as shown in Figure 3, the bottom half of the face image will be used to estimate the location and state of the mouth.

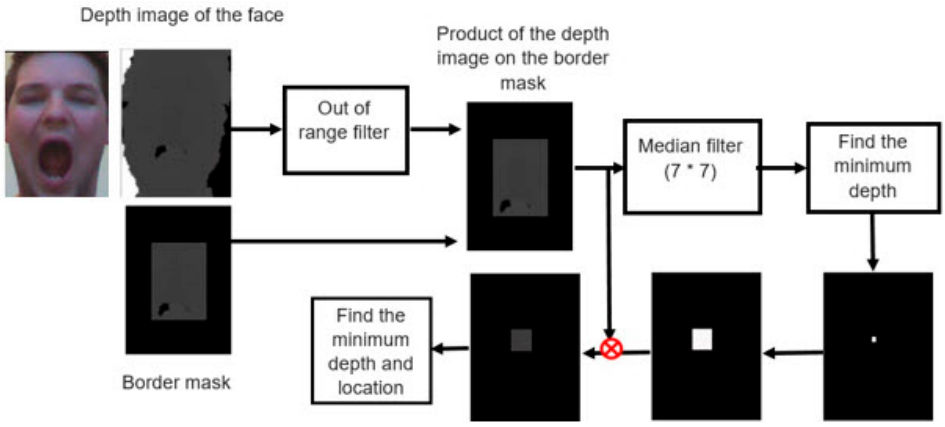

Figure 2. Block diagram of the nose tip detection [9].

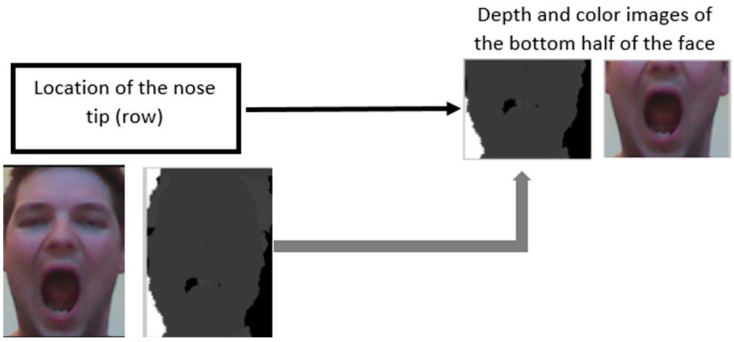

Figure 3 Block diagram for separating the bottom half of the face.

2.2. Lip Segmentation and Mouth Area Detection

In order to identify the location of lips and mouth in the lower face area, an adaptive mask as shown in Figure 4 is applied on the depth image. The depth value of Pixel (*i, j*) is compared with the average depth of its local neighborhood. The neighborhood size is measured by relation (l):

$$S(i,j) = \frac{(f * \bar{s})}{d(i,j)} \quad (1)$$

where *f* is the focal length of the camera, *d*(*i, j*) is the depth value in millimeter for pixel (*i, j*) and $\bar{S}$ is the average height of the mouth and is considered as 50 mm.

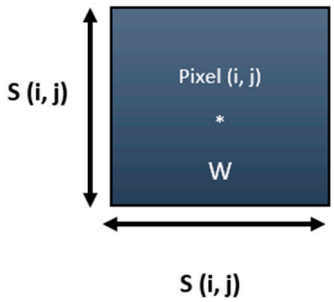

Figure 4. Adaptive mask used to detect the mouth area using depth information.

If the absolute difference between the average depth of the window W and depth *d*(*i, j*) is higher than a threshold, pixel(*i, j*) is set to 1, otherwise it is set to 0. Convolving the mask with the depth image acts as an edge detector and can identify edges of the mouth. The average depth of the window W is calculated as:

$$\mu_w(i,j) = \frac{\sum_{(u,v)\in w} V(u,v) * d(u,v)}{\sum_{(u,v)\in w} V(u,v) + \varepsilon)} \quad (2)$$

Here $V(u,v)$ is a masking parameter in order to consider only valid depth values and ε is a small constant. If the depth value of a pixel(u,v) is in the range of valid depth values, then $V(u,v)$ is set to one, otherwise it is considered zero:

$$V(i,j) = \begin{cases} 1 & if\ \ d(i,j)\ is\ valid \\ 0 & Otherwise \end{cases} \quad (3)$$

Now by comparing *d*(*i, j*) (depth of the window center) with the average window depth, a mask (m) is created:

$$m(i,j) = \begin{cases} 1 & if\ |\mu_w(i,j) - d(i,j)| > \tau \\ 0 & Otherwise \end{cases} \quad (4)$$

Where $\tau = 5mm$ was determined empirically. Figures 5–7 show examples of the mouth area detection using this method for both open and closed mouths.

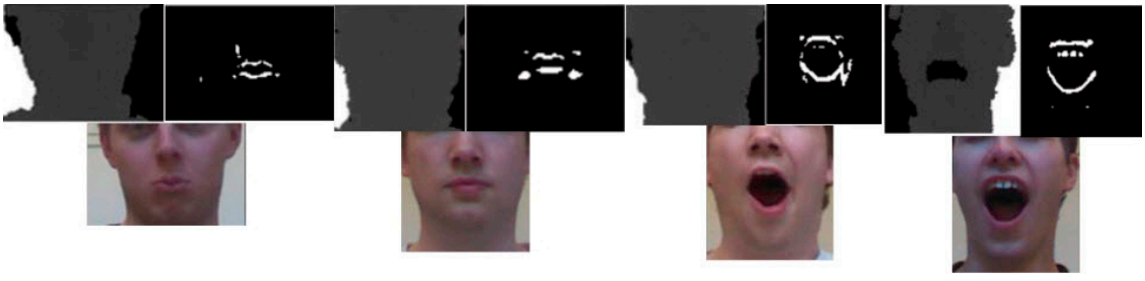

Figure 5. Segmented results that relies on having a clean binary image mouth by applying the adaptive mask on the depth image (first row) for 4 different individuals (second row).

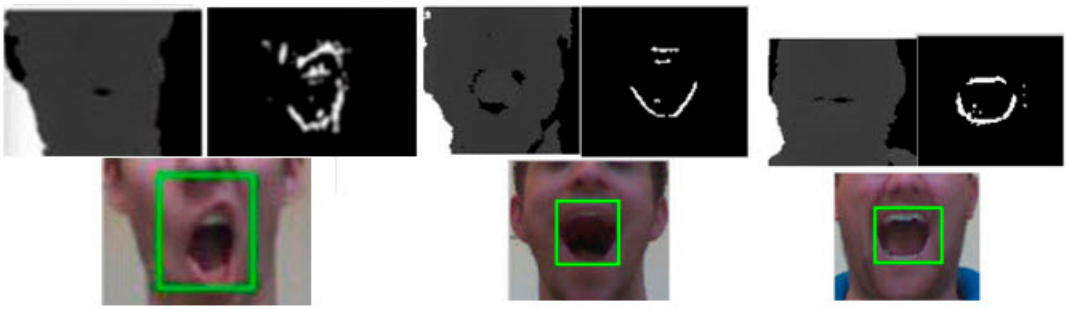

Figure 6. Results of detecting open mouths (second row) by applying the adaptive mask on the depth image (first row) for 4 different individuals.

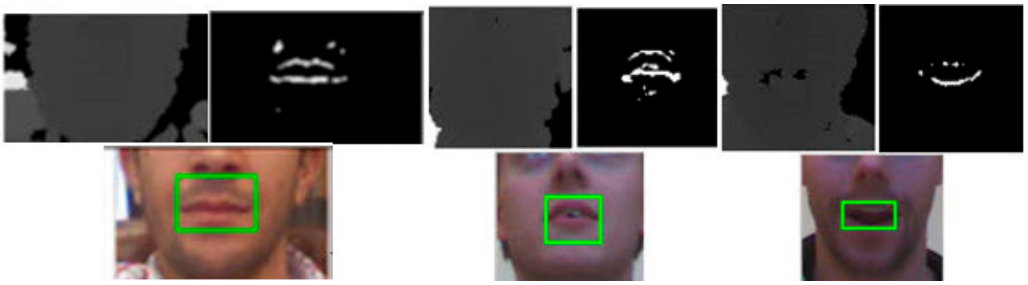

Figure 7. Results of detecting closed mouths (second row) by applying the adaptive mask on the depth image (first row) for 4 different individuals.

2.3. Open or Closed Mouth Detection

We are able to recognize an open mouth if the following two conditions are true:

1: The difference between the maximum depth and the average depth of the pixels in the mask area (ROI) should be greater than a threshold:

$$If\ (Maximum\ Depth - Average\ Depth) > T \tag{5}$$

then Condition 1 is True.

2: If the number of pixels in the mask area with a depth greater than the average depth of the area plus T is greater than $\frac{1}{5}$ of total pixels in the area (N), the second condition is true:

$$S = No.\ of pixels\ \ in\ the\ mask\ for\ which\ Depth(i,j) > AverageDepth + T \tag{6}$$

if $S > \frac{1}{5}$N then Condition 2 is true.

If both Conditions 1 and 2 are true, an open mouth is detected otherwise the mouth is considered as closed.

3. Experimental Results

The suggested algorithms in this paper have been evaluated on the VAP_RGBD dataset [10]. The dataset contains faces of 31 individuals, each individual has been pictured in 17 different positions, and each state has been repeated 3 times. As a result, there are 3 * 17 * 31 = 1581 RGB images and 1581 depth images in the data set. Color images have 32 bit standard bitmap format with resolution of 1280 × 960 pixels. Depth images are text files where each pixel of the depth image is represented by its depth value. The value of depth is measured using a millimeter-scale distance by the Kinect camera. Depth images have a resolution of 480 × 640 pixels where valid values for depths range from 400 to 4000 mm. Some outlaws in depth images are: undefined depth values −1, depth values that are too close 0 and depth values that are too far 4095. Another dataset prepared by authors, using an Xbox One 360 Kinect camera, includes 10 sequences of images from different

people with 10,000 frames. People were located in the distance of 50 to 100 cm from the camera. Videos were recorded with 30 frames per second. RGB images were saved using Bitmap format and depth images were saved as text files with 480 * 640 resolution. Valid values for depth range from 50 to 3000 mm.

3.1. Nose Tip Detection

As previously stated in Section 2, the nose tip is determined by finding the lowest depth value in the depth image. The use of depth information eliminates the algorithm's dependency on illumination variations and reduces errors in the following stages.

3.2. Lip Segmentation and Mouth Area Detection

Using the lips segmentation and mouth localization algorithm explained in Section 2, experiments were conducted on 140 images of open mouths and 40 images containing semi-open or closed mouths. The algorithm successfully identified 136 out of the 140 cases of open mouths and 31 out of the 40 cases in which the mouth is semi-open or closed. Table 1 represent the results obtained from these experiments. Results of the Viola Jones algorithm [11] are also included for comparison. The proposed mouth area detection algorithm detected 167 out of 180 cases correctly with an accuracy of 91% while Viola Jones' algorithm has only detected 106 cases correctly with 58% accuracy. Viola Jones ' algorithm has high accuracy in detecting closed mouths. The performance of the proposed algorithm in detecting both closed and open mouths is better.

Table 1. Comparison of the results of Viola Jones mouth detection algorithm and the proposed algorithm.

Total Samples = 180	**Viola Jones Algorithm**		**Proposed Algorithm**	
	Closed Mouth	**Open Mouth**	**Closed Mouth**	**Open Mouth**
Total No. of Closed mouth = 40	35	5	31	9
Total No. of Open mouth = 140	69	71	4	136

3.3. Open or Closed Mouth Detection

The area of the mouth in the bottom half of the face was determined by applying the adaptive mask (2.2) to the depth information. From the dataset, 50 images of open mouth and 37 images of semi-open or closed mouth were extracted for this experiment. Results show that the algorithm is able to correctly identify an open mouth 43 out of the 50 input cases and a semi-open or closed mouth 32 out of the 37 input. The detection of an open or closed mouth achieved an accuracy of 86%. Table 2 shows the results of this step.

Table 2. Results for the detection of open or closed mouth.

#Total Samples: 87	**Predicted Closed Mouth**	**Predicted Open Mouth**
#Closed mouth: 37	32	5
#Open mouth: 50	7	43

4. Conclusions

In this paper, an algorithm was developed for segmentation of lips and investigating the state of the mouth that was initiated by nose tip identification through finding the minimum data in the face depth image. Having the nose tip location, the lower part of the face was separated from the whole image. In the resulting image, by applying an adaptive mask which acts as an edge detector the mouth area was identified with 91% accuracy and the state of the mouth was investigated with 86% accuracy. Using depth information which is not sensitive to illumination changes is the main advantage of our work.

References

1. Wang, J.; Xiong, R.; Chu, J. Facial feature points detecting based on Gaussian Mixture Models. *Pattern Recognit. Lett*. **2015**, *53*, 62–68.
2. Agrawal, S.; Khatri, P. Facial expression detection techniques: Based on Viola and Jones algorithm and principal component analysis. In Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 21 February 2015; pp. 108–112.
3. Galatas, G.; Potamianos, G.; Makedon, F. Audio-visual speech recognition incorporating facial depth information captured by the Kinect. In Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27 August 2012; pp. 2714–2717.
4. Hassanat, A.B.; Alkasassbeh, M.; Al-awadi, M.; Esra'a, A.A. Color-based lips segmentation method using artificial neural networks. In Proceedings of the 6th International Conference on Information and Communication Systems (ICICS), Amman, Jordan, 7 April 2015; pp. 188–193.
5. Lüsi, I.; Anbarjafari, G.; Meister, E. Real-time mimicking of estonian speaker's mouth movements on a 3d avatar using kinect 2. In Proceedings of the 2015 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 28–30 October 2015; pp. 141–143.
6. Kalbkhani, H.; Amirani, M.C. An efficient algorithm for lip segmentation in color face images based on local information. *J. World's Electr. Eng. Technol.* **2012**, *1*, 12–6.
7. Hsu, R.L.; Abdel-Mottaleb, M.; Jain, A.K. Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 696–706.
8. Ji, Y.; Wang, S.; Lu, Y.; Wei, J.; Zhao, Y. Eye and mouth state detection algorithm based on contour feature extraction. *J. Electron. Imaging* **2018**, *27*, 051205.
9. Fong, K.K. IR-Depth Face Detection and Lip Localization Using Kinect V2. Master's Theses, California Polytechnic State University, San Luis Obispo, CA, USA, 2015.
10. Hg, R.I.; Jasek, P.; Rofidal, C.; Nasrollahi, K.; Moeslund, T.B.; Tranchet, G. An rgb-d database using microsoft's kinect for windows for face detection. In Proceedings of the 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, Naples, Italy, 25–29 November 2012; pp. 42–46.
11. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001.