TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 13, No. 1, January 2015, pp. 180 ~ 186
DOI: 10.11591/telkomnika.v13i1.6872

An Efficient Lip-reading Method Using K-nearest Neighbor Algorithm

Rashed Mustafa[1, 2, 3], Dingju Zhu[1, 2, 4, 5]
[1]Laboratory for Smart Computing and Information Sciences, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Department of Computer Science and Engineering, University of Chittagong, Bangladesh
[4]Shenzhen Public Platform for Triple-play Video Transcoding Center
[5]School of Computer Science, South China normal University Guangzhou, China
*****Corresponding author, e-mail: rashed.m@siat.ac.cn

Abstract

Many studies have been carried out on lip reading, most of those works are based on color images, while some essential features might not be obtained, like inner lip information. In this paper, RGB-D camera will be introduced for improving the recognition rate of lip reading. We try to complete lip reading through using only gray-scale images. Thirteen groups of words are given, and we present eight features for classification. Volunteers are asked to sit in the front of RGB-D camera. For each word we select 15 frames. K-nearest neighbor algorithm (KNN) is used to select the same words between different volunteers.

Keywords: *lip reading, visual feature extraction, face detection, KNN*

Copyright © 2015 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction

Lip reading, also known as speech reading, is a technique of understanding speech by visually interpreting the movements of lips, face and tongue using the information provided by the context, language, and any residual hearing [2]. It is different from speech recognition because in speech recognition the speaker is audible, but in lip-reading only the motion of lips and other facial features like gestures etc is available. Although a lip-reading system does not necessitates the availability of knowledge about context, language or residual hearing, but any information about the above three features is definitely exploited by such as systems in [2-5]. Lip-reading is a complex art of observation, inference and inspired guesswork. Even the most skilled lip reader cannot accurately identify every word, because different sounds can be made with the lips in the same position: inferring likely words from the context is often necessary. Fast speech, poor pronunciation, bad lighting, faces turning away, hands over mouths, moustaches and beards, all these make lip reading more difficult or even impossible. Moreover, we form many sounds in the middle of our mouth. Others come from the back of our mouth and even in our throat. These latter are absolutely impossible to speech read. There are numerous homophones in languages. Same words uttered differently by the person’s lips. In order to effectiveness of speech reading, we have to know the subject being discussed. A downside of speech reading is that it is very tiring, especially with someone who is hard to speech read in the first place.

The sensitivity of lip movement and accuracy of lip location are of great challenge for lip reading. Most of the researches used frontal facial image that contains predominant visible speaking information. Yao et al [1] showed several methods of lip reading, especial in details for feature extraction: direct on pixels method, model method and mixture feature extraction method. Most lip reading system can be capable of 80%~90% with clean environment, while it is only 50%~80% in natural environment. Ong E.J. et al [12] attempts to tackle it by building visual sequence classifiers that are based on salient temporal signatures. This method integrate AdaBoost [6] algorithm, which can be applied to multi-class recognition in the lip reading domain. Also, profile view (PV) [1, 12] contains protrude lengths like lip heights and protrusion lengths, which are important features for lip reading. The experiments on speaker-dependent

Received October 19, 2014; Revised November 26, 2014; Accepted December 16, 2014

isolated word speech recognition have shown the improvements of the use of integration of PV and FV (frontal view).

The rest of this paper can be organized according to the following sections: data collection through RGB-D camera will be describes in section 2, section 3 gives an overview of lip reading system, section 4 describes feature extraction system in gray scale images and K-nearest neighbor algorithm (KNN), methodology and experimental results shown in section 5 finally section 6 concludes this paper.

2. Data Sets

To achieve the goal of this research, the organization of the data is categorized into two sections: i) Audio and ii) Video. Each audio and video data are related to 5 male and 5 female. Each person has 13 videos that are concentrating to utter traditional Chinese characters. Each video split into some variable number of segments, for example xyz.avi file is the segmented part of a person which is responsible to utter three times /yi1/ and three times /er3/. The audio files are marked as xyz.wav contains audio data for the corresponding video files. The preprocessing phase is to extract significant frames from the video files by using corresponding audio files. It can be shown that 1776 frames extracted for one person. After that features would be identified which are feed to a machine-learning tool to identify specific characters are uttering by test person.

Table 1. 13 groups of Chinese words based on image information

/Yi1/, /er4/	/Jiu3/, /shi2/	/Fo1/,/fu1/,/da1/	/Mi1/,/mu1/,/fa1/
/San1/, /si4/	/Pa1/,/po1/,/pi1/	/Pu1/,/ma1/,/mo1/	
/Wu3/, /liu4/	/De1/,/di1/,/du1/	/Ba1/,/bo1/,/bi1/	
/Qi1/, /ba1/	/Ta1/,/te1/,/ti1/	/Tu1/,/na1/,/ne1/	

Table.1 gives the thirteen groups. And our objection is to classify 13 groups of Chinese words based on image information. Female and male volunteers' images which are screenshot of the original video are given in Figure 1.

Figure 1. Five female (up) and five male (down) in original vide

3. Lip Reading System

In section 2 we described some common video scene change detection methods (SAD, HD, ECR and SCDSW). All methods devoted to detect sudden or gradual transitions of scenes. In those works we didn't find any indication that how often the scenes are changing. Our proposed method is based on a ground truth of frequent and infrequent nature of changing scenes. The experiment carried out on 121 unstructured videos with arbitrary length containing objectionable and benign scenes. At first all Key frames are extracted using an open source tool ffmpeg [17]. Then summarize the result for instance Table 1 demonstrated the extracted key frames, its types and hit or misses status of different video genres.

3.1. Key Frames Selection

A key frame is a frame encoded without reference to any images in another frame [7]. In digital video editing, a key frame is a frame used to indicate the beginning or end of a change made to the signal. For example, a key frame could be set to indicate the point at which the audio will have faded up or down to a certain level [7]. There are several methods to extract key frames such as: k-means, hard cut between frames, perceived motion energy level, adaptive key frames using unsupervised clustering etc [8-10]. In this research we applied a different technique for extracting the key frames. The frames between beginning and ending time to utter any character have been considered as key frames. In order to achieve this we split the audio files according to utterance of characters then ffpmeg [11] tool utilized for key frames extraction.

3.2. Face Detection

To obtain the exact lip region, generally, some works [14, 15] focus on lip region directly, rather than acquiring face region firstly. Considering it works in natural environment, at first we get face region from camera. Here, we adopt Haar-like features [13] to detect face. For lip region, it is directly determined through the relative position of the lip in the face generally. While, it might get error rectangle since lip might be deformed during speaking. To solve the problem, nostril detection is adopted because the nostril’s relative position with mouth is stable while speaking. Figure 2 shows detected-face using the above methods.

4. Grey Scale Feature Extraction and KNN

4.1. Grey Scale Feature Extractions

Feature extraction phase is considered to be the most important phase in pattern recognition system. In this phase, a set of features have to be extracted from the image. These features are the base of the classification. Feature extraction is considered as a dimension reduction technique because the image is shorthanded to less dimensions and values using mathematical models rather than dealing with the overall image. There are many methods for feature extraction and these methods are application dependent. The mostly common methods can be classified mainly into four approaches. These approaches and their corresponding methods can be classified. For our research histogram feature are chosen as statistical features to extract some descriptors from mouth images in our data set.

Statistical methods analyze the spatial distribution of gray values by computing local features at each point in the image, and deriving a set of statistics from the distributions of the local features. With this method, the textures are described by statistical measures. Depending on the number of pixels defining the local feature, the statistical methods can be further classified into first-order (one pixel), second-order (two pixels) and higher-order (three or more pixels) statistics [19]. The most commonly used statistical methods are histogram properties, autocorrelation, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run-Length (GLRL) and Local Binary Pattern (LBP). Because of image sizes in our dataset are relatively small and number of images is very large, ordinary histogram features are chosen for their simplicity and fast calculation. We chose a set of 8 features that can be calculated from ordinary histogram.

Image histogram, as mentioned before, is a first order statistics which is one pixel level. There are many statistical measures that can be extracted from the histogram using first order probability distribution. These features are as following:

1) Mean value (average) among the intensity of pixel values as given in Equation (1).

$$\bar{M} = \frac{1}{n}\sum_{i=1}^{n} r_i\,,\ where\ 0 \le r_i \le G \tag{1}$$

2) Variance gray level values which shows how the pixel values differs from the mean value of gray levels and can be calculated as given in Equation (2).

$$\sigma^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{M})^2}{n} \tag{2}$$

3) Standard deviation that identifies the variation or dispersion of gray level values exists from the average gray level. Standard deviation is computed as the square root of variance given in Equation (2).

4) Skewness is used as a measure of symmetry or lack of symmetry of the normal distribution of the gray levels of the image. Skewness can be calculated as given in Equation (3).

$$s = \frac{1}{n}\frac{\sum_{i=1}^{n}(r_i - \bar{M})^3}{\sigma^3} \tag{3}$$

5) Kurtosis is used as measure whether data are peaked or flat relative to the normal distribution of the gray levels of the image. It can be measured as given in Equation (4).

$$k = \frac{1}{n}\frac{\sum_{i=1}^{n}(r_i - \bar{M})^4}{\sigma^4} \tag{4}$$

6) Maximum gray level value appeared in the image.
7) Minimum gray level value appeared in the image.
8) Mode value which represent the frequently repeated gray level value in the image.

4.2. KNN

In pattern recognition, the k-nearest neighbor algorithm (KNN) is a method for classifying objects based on closet training examples in the feature space. KNN is a way of supervised and instance-based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. The KNN algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simple assigned to the class of its nearest neighbor [20]. The KNN algorithm is very simple and easy to implement. There is not an explicit model. And the KNN is particularly effective method for the type of data variables including the characteristics.

In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point [20]. The KNN algorithm measures the distance between a query scenario and a set of scenarios in the data set. Usually Euclidean distance is used as the distance metric. The Euclidean distance measuring is given as following [15].

$$\mathrm{d_E(x,y)} = \sum_{\mathrm{i=1}}^{\mathrm{N}}\sqrt{\mathrm{x^2 - y^2}} \tag{5}$$

Here, we describe a classical example of the KNN classification. The test sample should be classified either to the first class of blue squares or to the second class of red triangles. If k = 3 it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If k = 5 it is assigned to the first class.

To start with, k parameter selection depends upon the data set; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when k = 1) is called the nearest neighbor algorithm. The accuracy of the KNN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification [23].

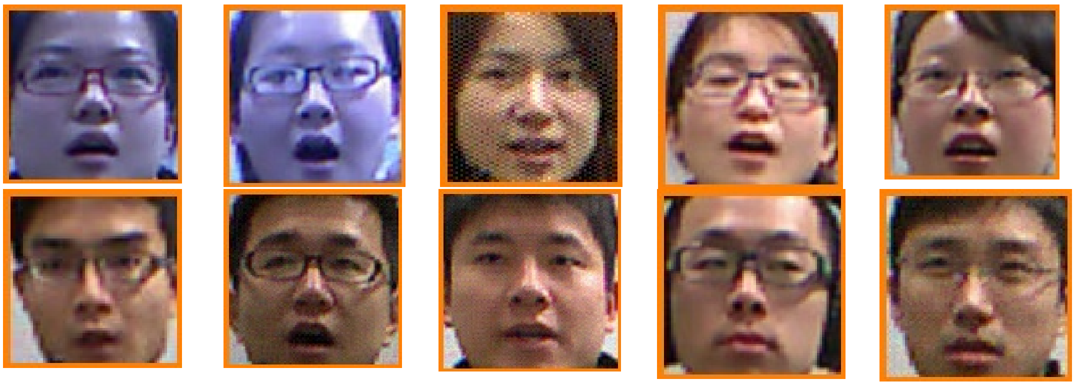

Figure 2. Five female (up) and five male (down) of face detection

4.3. Mouth Region Location

Location of Mouth Region in face images is considered to the key role in lip reading system. In this stage, mouth region is then located on each frame and stored in a database. A separate database of character is also maintained. The position of mouth region is matched with the characters to determine what the speaker has spoken. In recent years, many studies on lip segmentation have been proposed to address a wide range of image segmentation problems in image processing and computer vision. State of the art lip reading systems use several techniques for mouth region location. In the former stage, some face detection are missed more or less, such as video 3 of male, others may be lost. To segment lip region we obtain the perpendicular distance H between the top and the bottom and the horizontal distance W between the left and the right. It has been observed that the horizontal motion of lips is restricted in a region [Wmin, Wmax] and vertical motion is restricted in a region [Hmin, Hmax]. Due to the difference between each person, we can adjust the parameters to get good result of mouth region location.

Those data should be grouped into thirteen groups respectively. We might not get all frames for each word, that's mean missing data also should be figured out. Here it is supposed that the first frame is detected, and then missing frame is coped from the latest frame. In our experiment, fifteen frames are gotten for each word. Besides features are calculated into standard features by Equation (6):

$$V_{\text{standard}} = \frac{V_{\text{orignal}}}{\max(V) - \min(V)} \tag{6}$$

There are 2 classes in groups "000", "001", "002", "003", "004"; 3 classes in rest groups. So we give grouping information like Figure 3, which G1 include, subgroups of 2 classes and G2 include subgroups of 3 classes.

5. Methodologies and Results

5.1. Preprocessing Data

There are 9 volunteers, in which one is cancelled because of a poor result of mouth location (See section 3.3).

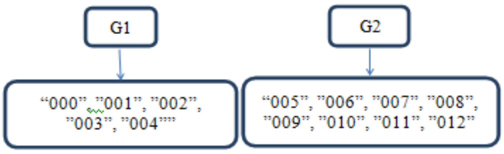

Figure 3. Grouping indexes

5.2. Features Classification

Eight features for each word have been considered where we choose 15 frames for each word.

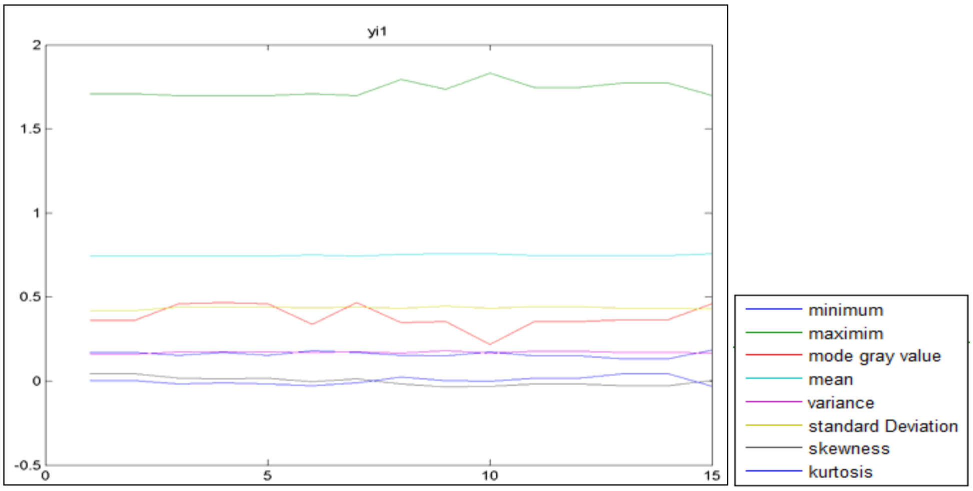

Figure 4. Giving an example of eight features changed through time

Here is the way to classify features. For instance, there are 30 words (2 words * 3 times *5 persons) in female group of “000”. The first word “yi1” is chosen as the tested word, and the distances between tested word and all word in the group are gotten. Because the subgroup of “000” belongs to G1, we choose median of distance as measures. Scores are calculated based on median of distances. If it is belonged to G2, tertile analysis is calculated. It is labeled to ‘1’ which its distance is smaller than median and others labeled to ‘0’. That is how we get the ground truth of each subgroup. The precision is calculated by Equation (7),

$$\text{pres} = \frac{\#\text{ of right matching}}{\text{lenght of group}} \qquad (7)$$

Where right matching is the same label as ground truth.

Table 2 shows the precision of each subgroup when the test word is the first frame in each group. We can see that it almost larger than 0.5 for G1 and larger than 0.3.

Table 2. Precision of each subgroup

	Index of groups	Precision
G1	000	0.5333
	001	0.5333
	002	0.4000
	003	0.4667
	004	0.4667
G2	005	0.4889
	006	0.4000
	007	0.4889
	008	0.6222
	009	0.5333
	010	0.5778
	011	0.5778
	012	0.4889

6. Conclusion

Experiment performs well for those features. The quality of images is better for lip location It leads that face detection works on all key frames between beginning and ending when people sound a word. Through the experiment, we can figure out the considering time sequence between continues frames can do a big favor for recognizing word. We selected eight features for classification, just like minimum, maximum, mode gray value, mean, variance, standard deviation, skewness and kurtosis. Some works using VLBP as a feature for expression recognition [24] which works so well. Maybe VLBP can be used for lip reading in the future.

Acknowledgements

This research was supported in part by Shenzhen Technical Project (grant no. HLE201104220082A) and National Natural Science Foundation of China (grant no. 61105133) and Shenzhen Public Technical Platform (grant no. CXC201005260003A).

References

[1] Yao HX, Gao W, Wang R, Lang XB. A survey of lipreading-one of visual languages*. Chinese Journal of Electronics*, 29(2): 239-246, 2001.
[2] H McGurk, J MacDonald. Hearing lips and seeing voices. *Nature*. 1976; 264(5588): 746–748.
[3] TF Cootes, A Hill, CJ Taylor, J Haslam. The Use of Active Shape Models for Locating Structures in Medical Images*. Image and Vision Computing*. 1994; 12(6): 355-366.
[4] TF Cootes, CJ Taylor, DH Cooper, J Graham. Active Shape Models Their Training and Application. *Computer Vision and Image Understanding*. 1995; 61(1): 38-59.
[5] A Ilampapur, It Jam, TE Weymouth. *Indexing in video databases*. SPIE/IS&T Proceedings on Storage and Refrieval in Image and Video Databases. 1995; 2420 (San Jose): 292-306.
[6] www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf
[7] N Dimitrova, T McGee, H Elenbaas. *Video keyframe extraction and filtering: A keyframe is not a keyframe to everyone*. Proc. CIKM. 1997; 113–120.
[8] T Liu, HJ Zhang, F Qi. A novel video key frame extraction algorithm based on percived motion energy model. *IEEE transactions on circuits and systems for video technology*. 2003; 13(10).
[9] C Huang, B Liao. A robust scene-change detection method for video segmentation. *IEEE Transactions on Circuits and System for Video Technolog*y. 2001; 11(12).
[10] HJ Zhang, A Kankanhalli, S Smoliar. Automatic partitioning of full-motion videol*. ACM Multimedia Syst*. 1993; 1(1): 10–28.
[11] www.fmpeg.org
[12] Ong EJ, Bowden R. *Learning temporal signatures for lip reading.* IEEE International Conference on Computer Vision Workshops. 2011; 958-965.
[13] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*. 2001; 511-518.
[14] Saitoh T, Konishi R. *Profile lip reading for vowel and word recognition.* International Conference on Pattern Recognition. 2010; 1356-1359.
[15] Sagheer A, Tsuruta N, Taniguchi RI, Maeda S. *Visual speech features representation for automatic lip-reading*. IEEE International Conference on Acoustics, Speech and Signal Processing. 2005; 781-784.
[16] TF Cootes, A Hill, CJ Taylor, J Haslam. The Use of Active Shape Models for Locating Structures in Medical Images. Image and Vision Computing. 1994; 12(6): 355-366.
[17] TF Cootes, GJ Edwards, CJ Taylor. *Active Appearance Models.* Proc. European Conf. Computer Vision. 1998; 484-498.
[18] X Xie. A Review of Recent Advances in Surface Defect Detection using Texture analysis Techniques. *Electronic Letters on Computer Vision and Image Analysis*. 2008; 7(3):1-22.
[19] D Unay, A Ekin, M Cetin, R Jasinschi, A Ercil. *Robustness of Local Binary Patterns in Brain MR Image Analysis.* Proceedings of the 29th Annual International Conference of the IEEE, EMBS Cité Internationale, Lyon, France. 2007; 2098-2101.
[20] http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm.
[21] Paul Lammertsma. K-nearest-neighbor algorithm: http:/*/paul.luminos.nl/download/document/knn.pdf*
[22] Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, Uri Shaft. When is “nearest neighbor” meaningful?. Database Theory—ICDT’99, pp. 217-235, Springer Berlin Heidelberg. 1999.
[23] Nolan, Graham. Improving the k-Nearest Neighbour Algorithm with CUDA. Honours Programme, the University of Western Australia. 2009.
[24] Jian, Kong, Zhan Yong-zhao, Chen Ya-bi. Expression Recognition Based on VLBP and Optical Flow Mixed Features. Image and Graphics, ICIG'09, Fifth International Conference. 2009; 933-937.