Visual Speech Recognition for Daily Indonesian Words Based on Combination of Double Difference and Image Projection Method

To cite this article: A Nasuha et al 2019 J. Phys.: Conf. Ser. 1413 012040

View the article online for updates and enhancements.
Visual Speech Recognition for Daily Indonesian Words Based on Combination of Double Difference and Image Projection Method

A Nasuha, F Arifin, D Irmawati and N Hasanah
Electronics and Informatics Engineering Education Department,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Email: arisnasuha@uny.ac.id

Abstract. In visual speech recognition, there are two approaches to lip feature extraction, which are based on appearance and shape. The appearance-based approach is usually better, because it provides visual features that not only include lip structure but also the visibility of the teeth and tongue. However, the disadvantage of this approach is that it produces too many features. Integration of double difference and horizontal-vertical image projection, is part of the appearance approach, in addition to using image projection as dimensional reduction. In previous study, the method has succeeded in recognizing 5 daily indonesian words, with data in form of videos recorded inside the room. In this study, we used the same words with 5 additional new words recorded outside the room. MLP (Multi Layer Perceptron) and SVM (Support Vector Machine) are used as classifiers. The word recognition process is evaluated using 10-fold cross validation. The method tested reached 88.92% on classification accuracy and 0.9948 on AUC (Area Under ROC Curve).

1. Introduction
The automatic visual speech recognition, consists of 3 main stages, namely data acquisition, feature extraction, and introduction. Because of the characteristics of color and lip movement: lip color does not always contrast enough with the surrounding environment, and the shape of the lips keeps changing during the speech process [1], it is necessary to use an extraction method that is suitable for both characteristics.

The method for presenting the visual features of lips on video, based on space and time, can be linked into two categories [2], namely stationary-based methods and motion-based methods. The first method will display a video divided into an image frame, technically the feature extraction that is run on each frame. This method is called the static feature method. The second method feature will be taken from the relative movement between the object and the observer, namely the velocity distribution that appears in the conversation pattern in the picture. In other words, information is related to the composition of the objects collected and the speed of changes in composition.

Another type of feature extraction that can be found in the literature is extract according to the type of information source they process [3], which is based on the shape and appearance. ASM (Active Shape Model) and ACM (Active Contour Model) are examples of extraction based on shapes, while DFT (Discrete Fourier Transform) and DCT (Discrete Cosine Transform) are appearance-based examples. Appearance-based lip extraction provides visual features that not only contain lip structures, but also about the visibility of the tongue and teeth. The disadvantage of this extraction is the huge number of features.
There is a simple method for effective segmentation and lip tracking, namely frame difference method, which is one technique in the background subtraction method. This method is widely used in video surveillance and motion detection. The advantage of this method is the low computational process.

Application of frame difference and double difference algorithm for automatic visual speech recognition has studied by Nasuha et. al. [4]. In the research, frame difference is combined with horizontal-vertical image projection. Moreover, important issue of the frame difference, namely selection of the threshold value, had been discussed. Data on the research is five daily Indonesian words recorded from ten subjects in the controlled environment. This paper will discuss the application of the above method on recognition ten daily Indonesian words and data recorded outdoor.

2. Materials and Methods

2.1. Materials

In this research, we use video recorded from 18 volunteers, 12 men and 6 women, with age distribution from 19 to 21 years. All video are in color and focused around subject’s mouth in frontal face. In these video, all volunteers are recorded outdoor. Each of volunteers says ten simple words in Indonesian, which is widely used in daily life, i.e. “saya” (I), “mau” (shall, will, want), “makan” (eat), “minum” (drink), “mandi” (take a bath), “tolong” (help, please), “sapu” (broom), “meja” (table), “kursi” (chair, seat), and “sudah” (already, past, done). Each word is repeated three times, therefore there are 540 video data. The original frame size is 320x240 pixels. Each volunteer is recorded in 0.75 seconds for each word, in 30 frames per second. Example of frame from each subject can be seen in figure 1. Figure 2 is example of sequence of lip images pronouncing “mau”.

2.2. Methods

2.2.1. Double Difference. A double difference [5] is an algorithm that places three consecutive frames as a group, calculates the difference between two consecutive frames (see equation (1) and equation (2)) and performs logical AND operations on two frame differences (see equation (3)). Wherein $I(x,y,k)$ and $I(x,y,k-1)$ are intensity value in same pixel $(x,y)$ at the current and previous time, $dI(x,y)$ and $dI(x,y)$ are difference intensity value, $T$ is a specific threshold value. Illustration of the algorithm can be seen in figure 3.
The algorithm can better get the moving object region if the speed of the object moves and the frame speed matches.

\[ d_{I_1}(x,y) = \begin{cases} 1, & |I(x,y,k) - I(x,y,k-1)| \geq T \\ 0, & |I(x,y,k) - I(x,y,k-1)| < T \end{cases} \]  
\[ d_{I_2}(x,y) = \begin{cases} 1, & |I(x,y,k-1) - I(x,y,k-2)| \geq T \\ 0, & |I(x,y,k-1) - I(x,y,k-2)| < T \end{cases} \]  
\[ dI(x,y) = \begin{cases} 1, & d_{I_1}(x,y) \cap d_{I_2}(x,y) = 1 \\ 0, & d_{I_1}(x,y) \cap d_{I_2}(x,y) = 0 \end{cases} \]

Double difference has been known to overcome the disadvantages of frame difference, namely foreground aperture and ghosting.

2.2.2. Horizontal-Vertical Image Projection. Image projection is a one-dimensional representation of image content. Horizontal-vertical image projections are histograms from horizontal and vertical directions of binary imagery. The vertical-horizontal image projection is defined in equation (4) and equation (5).

\[ P_{\text{hor}}(y_0) = \sum_{x=0}^{N-1} I(x,y_0) \text{ for } 0 < y_0 < N - 1 \]  
\[ P_{\text{ver}}(x_0) = \sum_{y=0}^{M-1} I(x_0,y) \text{ for } 0 < x_0 < M - 1 \]

The \( P_{\text{hor}} \) and \( P_{\text{ver}} \) are image projection in horizontal and vertical way respectively, \( I(x,y) \) is intensity value in pixel \( (x,y) \), \( M \) and \( N \) are width and height of image in pixel respectively. Each row and each column of image becomes a bin in the histogram. The stored count in a bin is the number of 1-pixels that appear in that row or column. This method can extract image features quickly and easily. This method has proven successful for Amazigh handwritten character recognition [6]. Illustration of this method can be seen in figure 4.

The basic steps for the method are as follow:
1) Convert video from RGB color space to grayscale
2) Manually cropping the image in each frame to get area around the lip, from 320x240 pixels into 160x120 pixels, and down sampling to get 32x24 pixels
3) Apply double difference method, using threshold value 5
4) Folding the lip image to reduce the impact of the illumination mismatch of the left and right lip areas, therefore the size of each frame becomes 16x24 pixels, and total of data becomes doubled
5) Apply vertical and horizontal image projection, to extract the features of the lip image, which will be used as input of classifier. The results of feature extraction for each frame is \( (16+24) \) features, or 840 features for each video.
6) Apply word recognition using MLP (Multi Layer Perceptron) and SVM (Support Vector Machine)

For MLP classifier, we use three layers, and the number of hidden units is varied to get the best result. We use sigmoid for activation function, and also regularization to prevent overfitting. For SVM classifier, the kernel is varied, namely linear, polynomial, RBF (Radial Basis Function) and sigmoid. We use automatic parameter search [7] to get optimum parameter for learning for each kernel.

To evaluate the model of the method, we need a way to estimate the test risk. Because of limitation of the data, we use cross-validation [8] instead of classical approach, separate the data into training and testing part. In this research we use 10-fold cross-validation.

There are many ways to measure the performance of the classifier, such as accuracy and AUC. Some researchers prefer to use the AUC [9], because AUC is a better measure of accuracy based on formal definitions of discriminancy and consistency. On the other hand, some others choose accuracy, because it is more intuitive and easy to understand, especially when the number of data for each class of the same amount. For this study, we used the accuracy and AUC to get the benefits of both.

3. Result and Discussion

The result of overall experiments is shown in figure 5 and figure 6. Figure 5 shows comparison of CA and AUC for the proposed method using MLP as classifier. Figure 6 shows comparison of CA and AUC using SVM as classifier.

In this study, when we used MLP as classifier, the best result is 88.65% on CA measurement, achieved when using 200 hidden units and 0.9937 on AUC, achieved when using 50 hidden units. When we used SVM as classifier, the best kernel is RBF. These results were 88.92% on CA, and 0.9948 on AUC.

When compared with previous studies, the results of word recognition in this study declined. This is caused by at least two things, first, namely the video data taken outside the room, where the condition is not controlled. Interference is low brightness and the presence of shadows. Secondly, the number of target classes is twice that of previous studies, while the amount of learning data available is limited.

Another difference with previous research is that the results of the SVM classifier outperform MLP in word recognition. Several studies comparing two types of classifiers have been carried out by other researchers, with results that outperform each other. Because this research did not focus on the problem, more in-depth research was not conducted.

4. Conclusion

This paper discusses the application of combination of double difference method and horizontal-vertical image projection for automatic visual speech recognition, in recognizing ten daily Indonesian words. The proposed method achieves 88.92% on CA and 0.9948 on AUC, when using SVM with RBF kernel.

Our future work in this area involves evaluating the proposed method on classifying Indonesian visemes, not Indonesian words. Other works include studying the use of adaptive or optimum method in determining the threshold value.
5. References

[1] A. Nasuha, T. A. Sardjono, and M. H. Purnomo, “Lip segmentation and tracking based on Chan-Vese model,” presented at the 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2013, pp. 155–158.

[2] M. Li and Y. Cheung, “A novel motion based lip feature extraction for lip-reading,” in Computational Intelligence and Security, 2008. CIS’08. International Conference on, 2008, vol. 1, pp. 361–365.

[3] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audio-Visual Automatic Speech Recognition: An Overview,” p. 39, 2004.

[4] A. Nasuha, F. Arifin, T. A. Sardjono, H. Takahashi, and M. H. Purnomo, “Automatic Lip Reading for Daily Indonesian Words Based on Frame Difference and Horizontal-Vertical Image Projection,” JATIT, vol. 95, no. 2, pp. 393–402, Jan. 2017.

[5] Y. Kameda and M. Minoh, “A human motion estimation method using 3-successive video frames,” presented at the International Conference on Virtual Systems and Multimedia, Gifu, Japan, 1996, pp. 135–140.

[6] Y. Es-saady, A. Rachidi, M. El yassa, and D. Mammass, “Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character,” Int. J. Adv. Sci. Technol., vol. 33, Sep. 2011.

[7] N.-E. Ayat, M. Cheriet, and C. Y. Suen, “Optimization of the SVM kernels using an empirical error minimization scheme,” in Pattern Recognition with Support Vector Machines, Springer, 2002, pp. 354–369.

[8] G. Seni and J. F. Elder, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions, vol. 2. Morgan & Claypool, 2010.

[9] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 3, pp. 299–310, 2005.

Acknowledgement
We would like to express our gratitude to Zulfa Kurniawan and Arif Puji, due to the meaningful discussion about video processing.