Lip Motion Recognition for Indonesian Vowel Phonemes Using 3D Convolutional Neural Networks

Maxalmina Satria Kahfi
School of Computing
Telkom University
Bandung, Indonesia
maxalmina.kahfi@gmail.com

Kurniawan Nur Ramadhani
School of Computing
Telkom University
Bandung, Indonesia
kurniawannr@telkomuniversity.ac.id

Anditya Arifianto
School of Computing
Telkom University
Bandung, Indonesia
anditya@telkomuniversity.ac.id

***Abstract*—Lip motion recognition is a technique for interpreting visual data that focuses on the mouth area and aims to recognize lip movement. The development of lip motion recognition is expected to be used to develop communication tools with deaf people and to automate the speech-to-text process visually. In the Indonesian language, the existence of vowel phonemes is needed to produce sounds so that words and sentences in the Indonesian language can be formed. This paper proposes a model that can recognize Indonesian vowel phonemes (/a/, /i/, /u/, /e/, and /o/) in lip movements. We proposed a model that uses 3D Convolutional Neural Networks. The data in this paper were processed by resizing into 112x56 pixel resolution then, proceed to the data augmentation by reversing the data horizontally and add blur to the data. The results of the testing of the vowel phoneme recognition model on lip motion show the highest accuracy rate of 84%.**

***Keywords*—lip motion recognition, convolutional neural networks, Indonesian vowel phonemes, 3D convolution**

I. Introduction

The recognition of lip motion, which is also one of the visual speech recognition, is a technique to interpret visual data. The task focuses on the mouth area and aims to recognize lip motion so that it can be classified in class [1]. The recognition of Indonesian vowel phonemes on lip motion is needed to build the foundation of research on word recognition and even sentences in Indonesian Language on lip motion. Another benefit provided by the introduction of lip motion is making transcripts from a video without audio [2] and helping the deaf person in order to communicate with others. Besides communication, another benefit is the creation of therapeutic tool technology for the learning process to speak for a deaf person [3].

Research related to lip reading is mostly done for English, but very few for the Indonesian language. Lip reading research for the Indonesian language started in 2012 using Artificial Neural Networks (ANN). The study on the recognition of Indonesian vowel phoneme on lip motion was conducted by implementing lip localization followed by lip segmentation [3]. Moreover, the classification process carried out by Artificial Neural Networks (ANN) and achieved an accuracy of 75.9%. However, the performance of the system needs to be developed further so that the accuracy value may increase.

Several studies of computer vision use Convolutional Neural Networks (CNN) as the standard classifier. Since Krizhevsky et al. [4] with the AlexNet won the ILSVRC ImageNet Large Scale Visual Recognition Competition in 2012, CNN has become very popular for research in the field of computer vision such as object detection and classification of visual data into several classes. One of the advantages of CNN is that the entire system is trained in end-to-end. Because it is trained end-to-end, CNN also performs feature extraction so that the process of feature extraction on visual data is no longer needed before the model trains the data. For lip motion recognition problems, CNN has a significant contribution to this problem. A new CNN architecture, LipNet, was created and introduced for the recognition of lip motion for English [5].

The topic to be discussed in this paper focuses on the recognition of Indonesian vowel phonemes (/a/, /i/, /u/, /e/, and /o/) on lip motion. We introduce a model for the topic using 3D CNN. The primary goal of this paper is to perform data classification by the proposed model that can recognize five vowel phonemes. In the Indonesian language, it is difficult to form a word or sentence without involving the role of vowel phonemes. The existence of vowel phonemes is needed to produce sounds so that words and sentences in the Indonesian language can be formed and has a meaning. The use of 3D CNN is expected to capture the spatial features from the data accurately since the previous study that use ANN as the classifier cannot capture the spatial features of an image data. Spatial features refer to the arrangement of pixels and the relationship between them in image data. Despite the primary goal of this research, we hope this research can be used for the basis for visual speech recognition research in the Indonesian language in the form of words or sentences.

Furthermore, studies related to the methods used are in section II. The system design is explained in section III. In section IV, the result of the proposed method is presented and discussed. The conclusion and future works are finally highlighted in section V.