Malaysian Journal of Computing, 5 (2): 609-618, 2020

eISSN: 2600-8238 online

GESTURE RECOGNITION SYSTEM FOR NIGERIAN TRIBAL GREETING POSTURES USING SUPPORT VECTOR MACHINE

Segun Aina[1], Kofoworola V. Sholesi[2], Aderonke R. Lawal[3], Samuel D. Okegbile[4] and Adeniran I. Oluwaranti[5]

[1-5]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife
[1]s.aina@oauife.edu.ng, [2]sholesikofoworola@gmail.com, [3]alawal@oauife.edu.ng,
[4]sokegbile@oauife.edu.ng, [5]niranoluwaranti@oauife.edu.ng

ABSTRACT

This paper presents the application of Gaussian blur filters and Support Vector Machine (SVM) techniques for greeting recognition among the Yoruba tribe of Nigeria. Existing efforts have considered different recognition gestures. However, tribal greeting postures or gestures recognition for the Nigerian geographical space has not been studied before. Some cultural gestures are not correctly identified by people of the same tribe, not to mention other people from different tribes, thereby posing a challenge of misinterpretation of meaning. Also, some cultural gestures are unknown to most people outside a tribe, which could also hinder human interaction; hence there is a need to automate the recognition of Nigerian tribal greeting gestures. This work hence develops a Gaussian Blur – SVM based system capable of recognizing the Yoruba tribe greeting postures for men and women. Videos of individuals performing various greeting gestures were collected and processed into image frames. The images were resized and a Gaussian blur filter was used to remove noise from them. This research used a moment-based feature extraction algorithm to extract shape features that were passed as input to SVM. SVM is exploited and trained to perform the greeting gesture recognition task to recognize two Nigerian tribe greeting postures. To confirm the robustness of the system, 20%, 25% and 30% of the dataset acquired from the preprocessed images were used to test the system. A recognition rate of 94% could be achieved when SVM is used, as shown by the result which invariably proves that the proposed method is efficient.

Keywords: *Gaussian Blur, Greeting, SVM, Recognition.*

Received for review: 10-08-2020; Accepted: 14-10-2020; Published: 28-10-2020

1. Introduction

Greeting gesture is generally known as a non-vocal gesture in which the body is seen to communicate messages in actions in order to replace speech or used together with speech. According to Thovuttikul (2011), gestures are culture-specific and can convey very different meanings in different social or cultural settings. Greetings, which are a form of gesture, can be used to regularize patterns of reciprocal behavior among group members (Akindele, 1990). Tribal greeting postures are also known to be forms of gesture.

Automated gesture recognition is the ability of the computer to understand gestures with the goal of interpreting them via mathematical algorithms for the purpose of executing commands based on these gestures. Nigeria is a country that has more than 300 tribes with distinct cultural differences, but the major tribes in Nigeria are Yoruba, Igbo and Hausa. Despite

the fact that Nigeria has about 300 tribes with almost different cultures, Yoruba culture is known to have a well-defined and documented greeting postures when compared to others. Hence, this work focused on recognizing different gestures for Yoruba greeting postures.

According to Adediran (1984), the Yoruba people are one of the largest African ethnic groups in the Sahara Desert and are concentrated in the western region of Nigeria. The Yoruba people see the hero Oduduwa as a source of ultimate political authority. Mythology holds that all Yoruba people descended from a hero called Oduduwa and shared a common language and culture for centuries. Various Yoruba towns include *Ikenne*, *Ibadan*, *Ile-Ife*, *Ilorin* and others. The uniqueness of the Yoruba people has received a lot of attention from different research in the past decades. According to Schleicher (1997), a Yoruba girl kneels down, while greeting her parents and a Yoruba boy prostrates. This rule also applies to any younger person who greets an older man or woman. The younger person, when greeting usually looks down, while kneeling or prostrating to show respect for the elder. While prostrating for a king, a man also touches the ground for the king once to the left and once to the right to pay homage to the King, while a woman bends down more, while kneeling down (Schleicher, 1997).

A Support Vector Machine (SVM) model was developed to identify these postures, which have been preprocessed using Gaussian blur functions. Hummel *et al.* (1987) Gaussian blur or convolution against the Gaussian kernel is a common model for image and signal degradation. In image processing, Gaussian blur is the outcome of using a Gaussian function to blur an image, which in effect reduces image noise and details. SVM on the other hand are supervised learning models with associated learning algorithms that analyze data by constructing sets of hyper-planes in a definite or high dimensional space which is later used for regression and classification analysis and other tasks like outlier's detection. According to Cortes & Vapnik (1995), SVM implements the idea of inputting vectors that are non-linearly mapped to a high dimensional feature space where the linear decision surface is constructed.

The remainder of this paper is organized as follows: Section 2 discusses the related work; Section 3 describes the construction of the dataset for greeting gestures, while also presenting the outlines of the adopted methods for body gesture features extractions. The sketching method of SVM is also presented. Section 4 presents the results of the proposed solutions, while Section 5 concludes the paper and offers some suggestions for future work.

2. Related Work

Huang *et al.* (2009) developed a vision-based hand gesture recognition system using principal components analysis (PCA), Gabor filters and SVM. Feature extraction was done using Gabor filters. PCA was used to reduce the dimension of the feature vector output from the Gabor filters. The classification was done by training the SVM classifier. A recognition rate of 95.2% was achieved as indicated by the experimental results, confirming that the SVM can continue to produce accurate results after validation, even when trained using small datasets. This advantage of the SVM makes it a suitable technique for the presented solution. The presented solution is, however, limited to only hand gestures and does not take into cognizance the extraction of the overall body feature, which is very important when modeling any greeting gesture recognition system.

In a similar work, Ma *et al.* (2018) proposed a model that is based on the recognition of general and fine-grained human action in video sequences. This model applied a convolutional neural network (CNN) in processing each image obtained from the video sequences. The outcome of the validation shows a precision accuracy between 54% and 71%. Despite effectively combining the appearance and motion in its unified framework, the obtained precision accuracy is not satisfactory.

Pigou *et al.* (2014) proposed an automated recognition system for sign language using CNN. The model's architecture consists of two CNNs, one for extracting upper body features and the other for extracting hand features. Dropout and data augmentation were the main

approaches used to reduce over-fitting. The experiment result shows an accuracy of 95.68% with a false positive rate of 4.13% which was caused by noise movement. Although this model achieved a remarkable accuracy, while the model also considered extracting body features along with hand features, this model used two CNNs for extraction and may not be suitable in a case where only limited datasets can be gathered or available.

Zhu *et al.* (2017) adopted a 3D convolution and convolutional long short-term memory (LSTM) networks to implement a multi-modal gesture recognition system. The proposed model firstly learns short-term spatiotemporal features of gestures through a 3D convolutional neural network (CNN) and then learns long-term spatiotemporal features by convolutional LSTM networks based on the extracted short-term spatiotemporal features. The proposed method demonstrates the accuracy of 98.89% on the validation set of the SKIG dataset. To use CNN and convolutional LSTM for classification, a large dataset may be required.

Owing to the limited datasets available, we adopted the Gaussian Blur, since it is very effective for the removal of Gaussian noise, reducing-edge blurring, rotationally symmetric and it is computationally more efficient than other filters such as median filters and Gabor filters. Instead of using two layers of the filter in this model, the moment was used for feature extraction because it preserves data characteristics for interpretability and also controls over-fitting when it is unsupervised. Apart from the fact that SVM can be used to train small datasets, it can be used for both feature selection and classification which results in better performance and accuracy. The risk of over-fitting is also less and it is relatively memory efficient and more effective in high dimensional spaces.

3. System Description

In order to design the proposed gesture recognition system, video collection and extraction of image frames, as well as preprocessing of images, feature extraction and classification of greeting gestures were carried out. The recognition process for the gesture recognition system is presented in Figure 1.

3.1 Video Collection, Extraction of Image Frames and Preprocessing of Images

Lighting condition, scale variability and posed angle of greeting gesture were some critical factors that were considered while collecting videos for Yoruba greeting gesture recognition. The videos of two greeting gestures from fifty (50) males and 50 females were collected with the same background color. Image frames were extracted from these videos and were properly labeled after analyzing and removing blurred images from the set of images. The image frames as shown in Figure 2 and Figure 3 were converted from red, green and blue (RGB) scale to grayscale. The converted image frames were further resized to obtain uniformity in frame sizes, while also reducing the frame size. These image frames were grouped appropriately in other to make up a dataset.

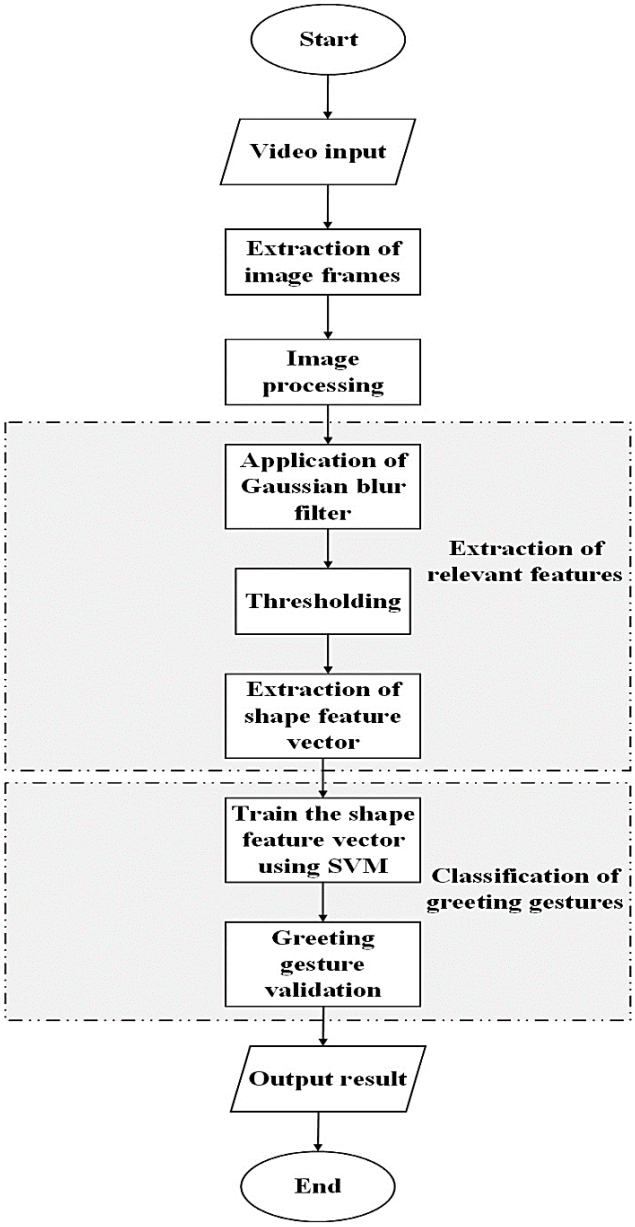

Figure 1: Flowchart for Gesture recognition process.

Figure 2: A Yoruba female greeting posture.

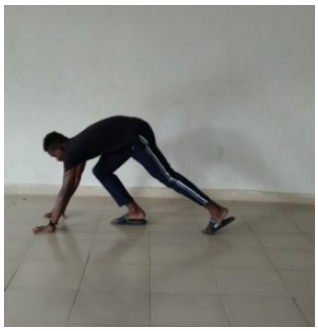

Figure 3: A Yoruba male greeting posture.

3.2 Feature Extraction of Body Features

Features of body gestures were collected through the following steps:

(i) Gaussian Blur Filter

Gaussian blur filter was applied to the preprocessed images in order to remove extra noises from the images. The Gaussian function in one dimension can be represented by:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \quad (1)$$

The Gaussian function in two dimensions can be represented by:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (2)$$

where x is the distance from the origin of the image in the horizontal axis, y is the distance from the origin of the image in the vertical axis and σ is the standard deviation of the Gaussian distribution.

Arora *et al.* (2008) stated that thresholding is an important technique for image segmentation because segmented images obtained from thresholding has the advantage of smaller storage space, fast processing speed and ease in manipulation compared with a grey level image. Hence, thresholding was applied in order to detect and extract the edges of the images, thereby leaving the edges white and non-edges black which could be seen in Figure 4 and Figure 5.

Figure 4: A Yoruba female greeting posture after thresholding.

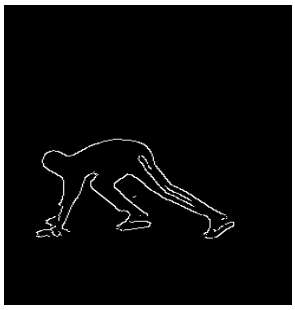

Figure 5: A Yoruba male greeting posture after thresholding.

(ii) Moments and Hu Moments

One of the common transformations and series expansion features in a feature extraction process is Moments (Kumar & Bhatia, 2014), where moment normalization strives to make the process of recognizing an object in an image size translation and rotation independent. A moment is a quantitative measure that was used to describe the image. The orientation and the position of the person in the image could be easily represented by centroid, variance and axis of the orientation image moments. Hu moments were later used for shape matching. Hu moments are the result of using central moments to calculate a set of seven numbers that are

constant to image transformation. The seventh Hu Moments changes in sign only when a shape is a mirror image of the other which invariably shows that the images are identical. The extracted feature vector from shape matching was then normalized so as to scale the variables in the vector to have values between 0 and 1. Hu (1962) computed moment as:

$$mu_{ji} = \sum_{x,y} array(x,y).(x-\bar{y})^{j}.(y-\bar{y})^{i}, \quad (3)$$

where $(\bar{x}, \bar{y})$ is the mass center of the image.

(iii) **Classification of Greeting Gestures**

An SVM classifier looks for the ideal hyper-plane which makes sure that worse case generalization errors are minimized, which is known as Structural risk minimization (SRM). A non-linear SVM can be used for classification between two classes. Gamma (γ) and C are parameters that are used in the Radial basis function (RBF) kernel SVM. These parameters decide the performance of an SVM model. Vert *et al.* (2004) stated that the RBF kernel on two samples x and y represented as feature vectors in some input space is defined as:

$$K(x,y) = \exp\left(\frac{-|x-y|^2}{2\sigma^2}\right), \quad (4)$$

where $|x-y|$ is the squared Euclidean distance between two feature vectors, σ is a free parameter also known as the C parameter in SVM. An equivalent definition of the RBF kernel involves a parameter gamma, γ which can be defined as:

$$\gamma = \frac{1}{2\sigma^2}. \quad (5)$$

Therefore, the kernel equation can be represented as:

$$K(x,y) = \exp(-\gamma|x-y|^2). \quad (6)$$

The gamma parameter determines how far or close a single training example can affect classification, where low and high gamma values mean 'far' and 'close' respectively. The C parameter trades off the maximization of the decision function margin against the correct classification of training examples. A larger C will encourage a smaller margin if the decision function is better at classifying all training points correctly. For a lower value of C, a larger margin will be accepted, therefore a simpler decision function, at the cost of training accuracy. For SVM, a high gamma value leads to more accuracy but biased results and vice-versa. Similarly, a large "C" value indicates poor accuracy but low bias and vice-versa. There should be a fine balance between variance and bias in any machine learning model. Table 1 shows the map of gamma and C with Variance and Bias.

Table 1: Map of Gamma and C with Variance and Bias.

	Large Gamma	**Small Gamma**	**Large C**	**Small C**
Variance	Low	High	High	Low
Bias	High	Low	Low	High

4. Results and Discussion

The datasets of the greeting gestures which consist of 1000 image frames (500 male and 500 female gestures) were classified as training datasets and testing datasets. Various C values and Gamma values were specified in order to determine the best SVM parameters to be used for classification as shown in Table 2.

Table 2: Various Gamma and C Values.

	C	Gamma
Values	0.01	e^{-4}
	0.1	e^{-3}
	1	e^{-2}
	10	0.1
	100	0.2
	1000	0.5

When 75% of the dataset was used for training and 25% of the dataset was used for testing, an accuracy of 94% was obtained. Precision for the kneeling and prostrate gestures were 89% and 99% respectively as shown in Table 3 and the confusion matrix obtained is shown in Figure 6.

Table 3: Precision and Accuracy for 25% Datasets Used for Validation.

	precision	recall	f1-score	support
kneel	0.89	0.99	0.94	117
prostrate	0.99	0.89	0.94	133
accuracy			0.94	250
macro average	0.94	0.94	0.94	250
weighted average	0.95	0.94	0.94	250

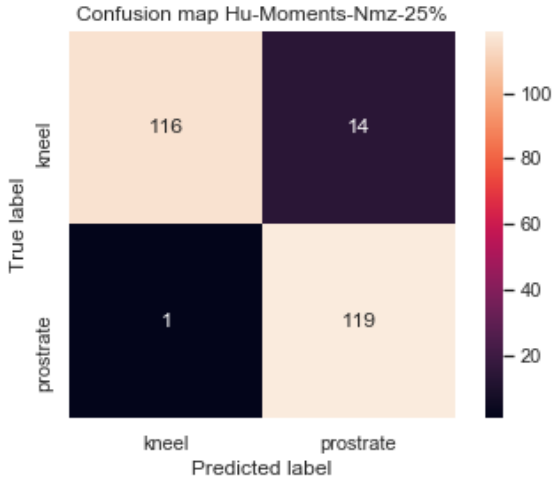

Figure 6: Confusion matrix obtained with 25% datasets used for validation.

When 80% of the dataset was used for training and 20% of the dataset was used for testing, an accuracy of 94% was obtained. Precision for the kneeling and prostrate gestures were 88% and 99% respectively as shown in Table 4 and the confusion matrix obtained is shown in Figure 7.

Table 4: Precision and Accuracy for 20% Datasets Used for Validation.

	precision	**recall**	**f1-score**	**support**
kneel	0.88	0.99	0.93	92
prostrate	0.99	0.89	0.94	108
accuracy			0.94	200
macro average	0.94	0.94	0.93	200
weighted average	0.94	0.94	0.94	200

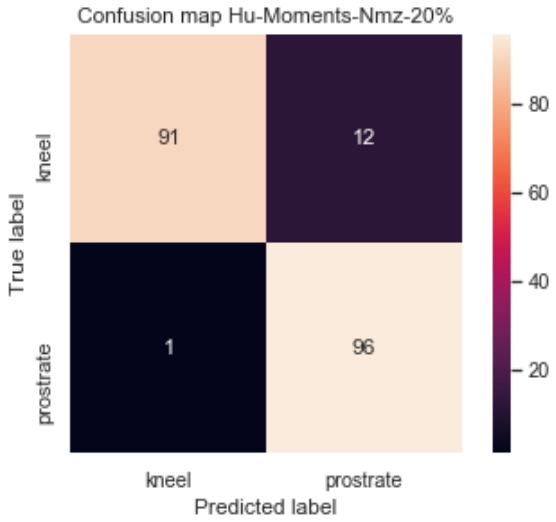

Figure 7: Confusion matrix obtained with 20% datasets used for validation.

When 70% of the dataset was used for training and 30% of the dataset was used for testing, an accuracy of 94% was obtained. Precision for the kneeling and prostrate gestures were 91% and 97% respectively as shown in Table 5 and the confusion matrix obtained is shown in Figure 8.

Table 5: Precision and Accuracy for 30% Datasets Used for Validation.

	precision	**recall**	**f1-score**	**support**
kneel	0.91	0.97	0.94	145
prostrate	0.97	0.91	0.94	155
accuracy			0.94	300
macro average	0.94	0.94	0.94	300
weighted average	0.94	0.94	0.94	300

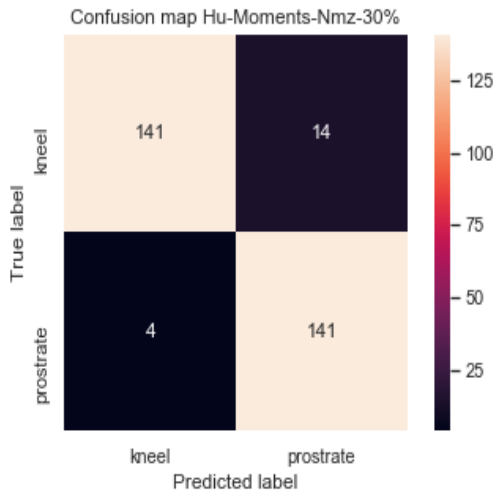

Figure 8: Confusion matrix obtained with 30% datasets used for validation.

Finally, the best parameters for the three instances were {'C' = 1000, 'γ' = 0.5} according to the heat map in Figure 9.

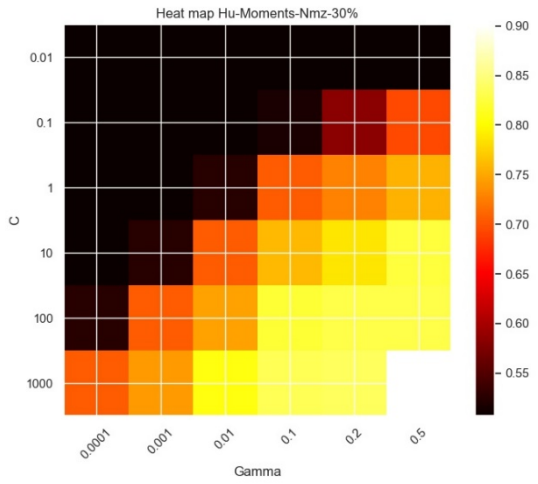

Figure 9: Map of 'C' plotted against 'γ' when 30% of the dataset was used for testing.

5. Conclusion and Recommendation

This work develops a gesture recognition system for Nigerian tribe greeting postures. Collection of video corpus for the greeting posture, processing of the corpus, extracting of relevant features and classifying of the greeting postures was carried out. The proposed Gaussian blur – SVM technique resulted in an average accuracy of 94% for the greeting posture recognition system. Gaussian blur was used to reduce image noise and for edge blurring, Moment was used for feature extraction and RBF kernel SVM was used to train the features extracted. For further works, it is recommended that the video should be collected in an enclosed space with a minimal amount of lighting in order to avoid a lot of noise in the preprocessed image. Gesture recognition for other Nigerian tribes such as Igbo and Hausa should be considered in further works.

REFERENCES

Adediran B. (1984). Yoruba Ethnic Groups or A Yoruba Ethnic Group? A review of the Problem of Ethnic Classification. *Africa*, 7(1984), 57-70.

Akindele, F. (1990). A sociolinguistic analysis of Yoruba greetings.*African Languages and Cultures*,3(1), 1-14.

Arora, S., Acharya, J., Verma, A., and Panigrahi, P. K. (2008). Multilevel thresholding for image segmentation through a fast statistical recursive algorithm. *Pattern Recognition Letters*, 29(2), 119-125.

Cortes, C., and Vapnik, V.N. (1995). Support-vector networks. *Machine learning,* 20(3),273-297.

Hu, M.K. (1962). Vision Pattern Recognition by Moment Invariants. *IEEE Transactions on Information Theory*, 8(2), 179-187.

Huang, D. Y., Hu, W. C., and Chang, S. H. (2009). Vision-Based Hand Gesture Recognition

Using PCA and Gabor Filters and SVM.*InInt. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, 2009 IEEE conference on (pp. 1 – 4), IEEE.

Hummel, R. A., Kimia, B., and Zucker, S. W. (1987). Deblurring Gaussian blur. *Computer Vision, Graphics and Image Processing*, 38(1), 66-80.

Kumar, G., and Bhatia, P. K. (2014). A Detailed Review of Feature Extraction in Image Processing Systems. *In2014 Fourth International Conference on Advanced Computing and Communication Technologies*, 2014 IEEE conference on (pp. 5 – 12), IEEE.

Ma, M., Marturi, N., Li Y., Leonardis, A., and Stolkin, R. (2018). Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76, 506-521.

Pigou, L., Dieleman, S., Kindermans, P.J., and Schrauwen, B. (2014). Sign Language Recognition Using Convolutional Neural Networks. *In European Conference on Computer Vision* (pp. 572 – 578), Springer, Cham.

Schleicher, A. F. (1997). Using greetings to teach cultural understanding. *Modern Language Journal*, 81, 334–343.

Thovuttikul, S., and Nishida, T. (2011). Handling Greeting Gesture in Simulated Crowd. *In Int. conf. on Granular Computing*, 2011 IEEE conference on (pp. 659 – 664), IEEE.

Vert, J.P., Tsuda, K., and Scholkopf, B. (2004). A Primer on Kernel Methods. In Kernel Methods in Computational Biology. MIT Press. (pp. 55–72).

Zhu, G., Zhang, L., Shen, P., and Song, J. (2017). Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM. *IEEE Access*, 5, 4517–4524.