ORIGINAL RESEARCH ARTICLE

SAMRIDDHI Volume 12, Special Issue 2, 2020 Print ISSN: 2229-7111 Online ISSN: 2454–5767

Histograms of Oriented Gradients-Based Gesture to Voice Conversion System for Indian Sign Language using Raspberry Pi

Rajeshri R. Itkarkar[1*], Omkar H. Darekar[1], Sahil U. Vora[1], Prachi K. Gorate[1], Nividita V. Ketkar[1], Dattataray Bormane[1], Anilkumar Nandi[2]

[1]All India Shri Shivaji Memorial Society's College of Engineering, Pune, Maharashtra, India

[2]BVB COE and Technology, Karnataka, India

Abstract

The hand gesture is one of the typical methods used in sign language. It is often very difficult for hearing-impaired people to communicate with the world. This paper presents a solution that will not only automatically recognize the hand gestures but will also convert it into speech and text output so that an impaired person can easily communicate with normal people. The system consists of a camera attached to a computer that will take images of hand gestures, histogram of gradient feature extraction is used to recognize the hand gestures of the person. Based on the recognized hand gestures, the system will produce voice output. The goal of this work is to develop a new type of human-computer interaction system to overcome the problems that users have been facing with the current system. A simple web camera is used to capture hand gesture images and recognize alphabets characters (A–Z) and numerals (0–9) using histograms of oriented gradients (HOG) features. The purpose is to implement the algorithm of extracting HOG features and these features are classified using a support vector machine (SVM) classifier for identification of Indian sign language (ISL). Hardware is developed using Raspberry Pi and Python programming.

Keywords: Histogram of gradient, Indian sign language, Raspberry Pi, Support vector machine.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2020); DOI: 10.18090/samriddhi.v12iS2.4

Introduction

Hand gesture recognition is one of the frequent and simple ways, in which humans and computers interact. Nowadays research is more focused on the recognition of gestures in real-time using artificial intelligence to process sign language. The gestures of the hand and the postures of the body are the natural means of communication. The use of hand gestures is an alternative to the awkward interface between the human-computer interaction (HCI) tools. Hand movements are typically used to express human emotions and to inform their thoughts. Visual interpretation of hand gestures, in particular, can help to achieve the ease and naturalness desired for HCI. Commonly, two approaches are widely used for decoding movements for the contact between humans and machines.

Methods using Gloves

This approach uses sensors (mechanical or optical) attached to a glove to turn the flexion of the finger into electrical signals to determine the position of the hand. Flex sensors and gyroscope are usually used to track hand and finger movements. Using gloves, users must hold a load of a cable that is connected to the device and obstructs the usability and naturalness of the interaction.

Corresponding Author: Rajeshri Rahul Itkarkar, All India Shri Shivaji Memorial Society's College of Engineering, Pune, Maharashtra, India, e-mail: rritkarkar@aissmscoe.com

How to cite his article: Itkarkar, R.R., Darekar, O., Vora, S., Gorate, P., Ketkar, N., Bormane, D.B., & Bhalke, D.G. (2020). Histograms of Oriented Gradients-Based Gesture to Voice Conversion System for Indian Sign Language using Raspberry Pi. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 12(SI-2), 16-21.

Source of support: Nil

Conflict of interest: None

Vision-Based Methods

Computer vision-based techniques are based on the way human beings perceive information about their surroundings. Although it is difficult to design a vision-based interface, still it is feasible to design the interface for a controlled environment. Detecting signs in images is a challenging task owing to their variable appearance and the wide range of poses that they can adopt. The first need is a robust feature set that allows the signs to be discriminated cleanly, even in cluttered backgrounds under difficult illumination. We will research the issue of human sign recognition feature sets for ISL, demonstrating that locally defined HOG descriptors

provide excellent performance relative to other established feature sets like wavelets. The proposed descriptors are reminiscent of edge orientation histograms, Scale-invariant feature transform (SIFT) descriptors, and shape contexts, but they are computed on a dense grid of uniformly spaced cells, and they use overlapping local contrast normalizations for improved performance.

The system comprises of Raspberry Pi, webcam, and speaker. The algorithm implanted to detect the ISL is a combination of HOG descriptor and SVM classifier with an average recognition rate of 92.38%.[1]

Literature Survey

Chen-Chiung Hsieh and Dung-Hua Liou,[2] a real-time hand gesture recognition system is developed for a complex background designed with Haar features, SVM classifier and 95% accuracy. Dalal *et al.*,[3] in this paper, locally defined gradient orientation histogram, features similar to SIFT descriptors in a complex overlapping grid provide very good results for individual detection, reducing false-positive levels by more than an order of magnitude relative to the best hair wavelet-based detector. Chih-Hung *et al.*,[4] a depth camera was used and implemented for static and dynamic depth images. Features were extracted by using an adaptive square algorithm and classified using SVM with 87.6% accuracy for dynamic. Parama *et al.*,[5] the paper presents the model of a sign language interpreter that can verbalize American sign language (ASL). This functional model is based on an HCI being generated using the user's hand movement only. The combination of hardware and software interfaces—webcam and MATLAB 2016a, performs the process of extracting features from the image captured from hand signs in real-time.

Methodology

The methodology uses HOG as feature extraction and SVM for classification. Python is used to code the HOG and SVM on Raspberry Pi.

Shape Feature Extraction using HOG

The HOG functions are widely used to detect objects. The image is divided into small square cells, the histogram of oriented gradients for each cell is determined, the result is normalized using a block-wise method, and a descriptor returned for each cell. HOG descriptor assumes the appearance and shape of the local object within an image can be described by the distribution of intensity gradients or edge directions. Implementation of these descriptors can be accomplished by dividing the image into small connected regions called cells, and a histogram of gradient directions, i.e., edge orientations for the pixels inside the cell can be computed for each cell. The descriptor then describes the combination of those histograms. The shape features are evaluated by applying color normalization on the input image, then evaluate the horizontal and vertical gradients, next is the formation of spatial blocks and then calculate orientation bin-wise and then forming a feature vector (HOG vector). The same is represented in the block schematic, as shown in Figure 1.

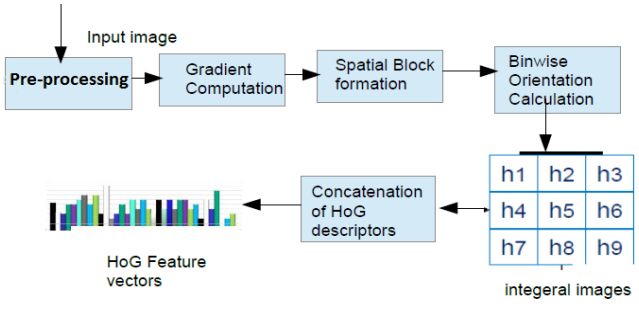

Figure 1: Methodology of HOG for shape feature extraction

The feature extraction process of HOG consists of pre-processing, gradient computation, block formation, bin-wise orientation evaluation, and finally, calculating the HOG vector/ HOG descriptor. Pre-processing consists of resizing the image and converting it to grayscale. The image is resized to 130 × 130, for the proper division of cells and blocks.

Gradient Computation

To calculate the histogram of the gradient, the vertical and horizontal gradient is calculated by filtering the image using kernel hx = [-1, 0, 1] and hy. Next, the magnitude and direction of the gradient are calculated as

$$Gr = \sqrt{gx^2 \pm gy^2} \quad (1)$$

$$\theta = \tan^{-1}\frac{gy}{gx} \quad (2)$$

The gradient has one magnitude and a direction at each pixel. Additionally, the gradient for a cell is determined. The size of the cell may be chosen as 2, 4, 8 16, and 32. The best choice is 8, i.e., the picture is broken down into 8 × 8 cells. The gradient contains two values, one for magnitude and the other for direction which adds up to 8 × 8 × 2 = 128. These 128 numbers are represented in 9 bins, which are stored as a number in an array. The histogram is a vector of 9 bins corresponding to angles 0, 20, 40, 60, 80, 100, 120, 140, and 160.[4-8]

Figure 2 shows the cell gradient with arrow length as magnitude and direction of the gradient. The figure also shows the magnitude and direction values of the gradient. The gradient values are assigned a place in the bin as per the direction. The bin is selected based on the direction and the value is selected based on magnitude. To build the 9-bin histogram, the contributions of all the pixels in the 88 cells are applied. Thus, the histogram is generated based on the image gradient. An image gradient is sensitive to the general lighting. When you make the picture darker by splitting all pixel values by two, the magnitude of the gradient will change by half, and thus, the histogram values will change by half, the descriptor must be independent of the variations of lighting. Thus, the histogram is normalized so that they

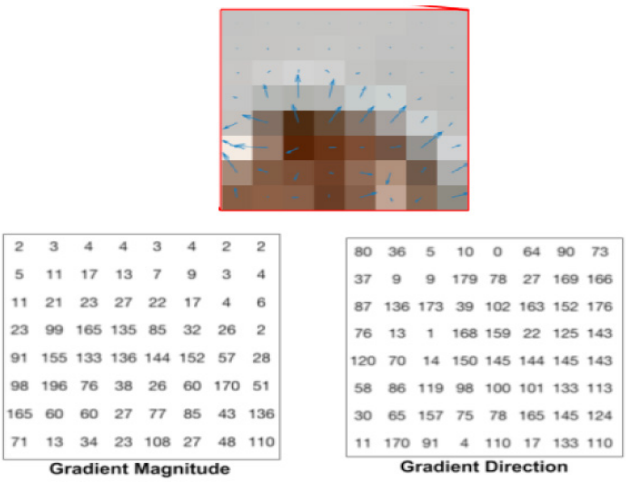

Figure 2: Cell with gradient direction and magnitude

are not affected by lightning variations. For normalization, suppose a vector [128, 64, 32], the length of this vector is = 146.64. This is called as L2 normalization of the image. Dividing each vector by length gives the normalized vector of the image. Therefore, an 8 × 8 cell forms 36 × 1 normalized vectors. This 36 × 1 normalized vector is the feature vector or called as HOG feature descriptor used for recognition by the classifier. Gradient vector length = vertical overlapping cell × horizontal cell × block size × no. of bins.

Classification by SVM

An SVM is a classifier based on the mapping of characteristics that are extracted from the feature vectors instance to space points. The points are grouped in separate groups during the training process, separated by a simple distance as wide as possible, defined by a hyperplane. In the classification phase, new feature vectors are mapped to points and, subsequently, predicted to belong to a class established during the training phase. This form of classifier provides a binary classification, creating two classes by default separated by an ideal hyperplane, but it is possible to define multiple evaluation classes by combining multiple binary issues with a multi-class problem, based on strategies, such as, one-versus-all or one-versus-one. It is a classifier which shows remarkable performance in nonlinear analysis and high-dimensional pattern recognition. As the combined features, such as, texture and shape are combined, where the dimension is high, SVM is the proposed classifier. The accuracy obtained in recognizing gestures using the Kinect sensor by SVM is 95.4%. The accuracy obtained is almost 100, i.e., 98.5% using SVM as the classifier. For static gesture recognition by SVM in is 99.4% and that with dynamic is 93%. Hence, to improve the accuracy with a high dimension, the SVM classifier is proposed.[1]

Performance Parameters

Accuracy and precision are the parameters proposed for performance measurement. With Gabor, the angle of texture direction can be varied, such as, 0, 45, or 90 degrees. With the HOG feature vector, the accuracy and precision are measured

Figure 3: Database for ISL

by changing the cell size proposed is 8 × 8 pixel, 16 × 16, 32 × 32, 64 × 64, etc., same can be measured by also varying block size (2 × 2, 3 × 3, 4 × 4 cells, etc.), with also variation in block overlapping and normalization with again different bin size of 3, 5, up to 20 bins.

Database

To implement a prototype of this work, a simple web camera is used to capture hand gesture images, and a system was developed to recognize alphabets characters (A-Z) and numerals (0-9), using HOG features. The database is created especially for ISL only. Figure 3 shows database created for Indian Sign Language Signs A to Z and Numbers 1 to 9.

System Design

Figure 4 shows the Block diagram of system.

Camera

The first functional block, i.e., the camera is used for image acquisition. A video image acquisition process is subjected to many environmental concerns, such as, the position of the camera, the number of cameras used, lighting sensitivity, and background condition. The camera is placed approximately fixed at the center of the chest on a person's body. That means the camera is supposed to be worn by the hearing impaired people. The camera will continuously capture video of the signs and the captures frames will be sent for detection of gestures. High frame rate cameras are required for fast and accurate results.

Web-Cam

Webcams are video capturing objects identified with personal computers (PCs) or PC frameworks, more often than not

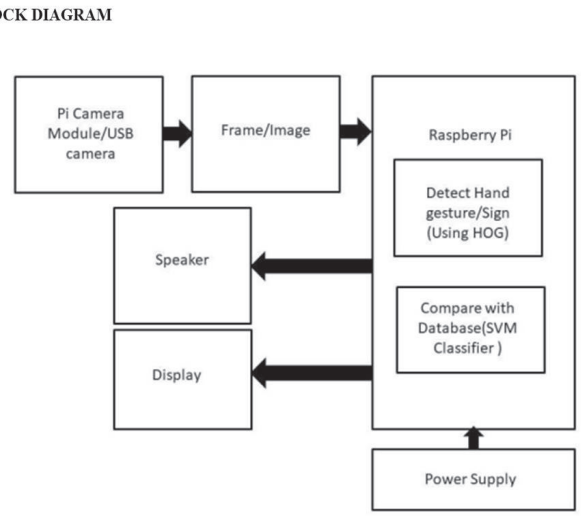

Figure 4: Block diagram

making utilization of USB or, if they interface with frameworks, ethernet or Wi-Fi. They are fabulous for low gathering charges and versatile applications. Webcams typically contain a central factor, a preview sensor, and a couple of assistance equipment. Exceptional central components are available, basically, the most extensively perceived being a plastic point of convergence that might be settled and out to set the digicam's center interest. Settled mindfulness central angles, which have no relationship for trade, are also beside open. Preview sensors can be complementary metal oxide semiconductor (CMOS) or charge-coupled device (CCD), the past being overpowering for insignificant exertion cameras, yet CCD cameras do not generally beat CMOS-based cameras inside the straightforwardness esteem run. Logitech, Microsoft Corporation, Intel, and Kinect are the manufacturers and suggestions for use. The Kinect camera's accuracy with the SVM classifier is remarkably good.

Computer

The second functional block and the main block are the personal computers where the detection of signs of fingers is done using HOG and compared with the database stored. Hence, complete image processing is carried out with the SVM classifier.

Power supply: A power supply is an electronic device that supplies electric energy to a load. The primary function of a power supply is to convert one form of electrical energy into another and as a result power supply is sometimes referred to as electric power converters. Many power supplies are separate, stand-alone devices, while others and their loads are built into larger machines. Also, the forms include set, variable, and dual control. As this is a prototype model meant to wear the power supply is a power bank of 10,000 mAh with the output of 2 Amp and 5V.

Raspberry Pi: There are no wearable sensors and connecting wires. An electronic talking contraption ends up cutting edge which incorporates a three framework system, *viz*., scene

Table 1: Specification of Logitech Quickcam pro 4000

Brand	*Logitech*
Colour	Black
Special features	640 × 480 video resolution, 1.3 megapixel, VGA CCD sensor
Item model number	Quickcam Pro 4000

capturing, data handling, and audio yield. The efficiency of communication between a normal person and speech impaired individual is improved by texting the gesture made by the speech impaired and then playing a corresponding audio file also displaying the sign meaning on screen. A camera and Raspberry Pi 3 processor has been used to capture the gesture of the speech-impaired person. The image processing program has been written using Python in the Raspbian OS. The image is compared with images available in the database and a code is generated on matching, which is HOG + SVM methodology. The generated code is used to display corresponding text on the LCD. The same is also used to generate the audio output.

The Raspberry Pi 3 demonstrate B has particularly worked with the Broadcom BCM2837 system-on-chip (SoC) incorporates four superior ARM Cortex-A53 process centers running at 1.2 GHz with 32 kB level one and 512 kB level a couple of reserve memory, a VideoCore IV illustrations processor, and is associated with a 1 GB LPDDR2 memory module on the back of the board. It also alternatives 40-pins broadly useful info yields general purpose input output (GPIO) and enhanced property with Bluetooth low energy (BLE) and BCM43143 Wi-Fi on board. It likewise has an updated control administration wellspring of 5V USB control supply up to 2.5 Amps. As of now, Raspberry Pi 3 Model B is the best of Raspberry Pi PCs. The framework handling is colossal with 1.2 GHz clock speed and 1 GB RAM Raspberry Pi can play out every single propelled process. As per the association shrewd, the board ought to be equipped for sending information to and from the board quickly. Another double band Wi-Fi underpins for 2.4 and 5 GHz 802.11b/g/n/air conditioning, which is likewise guaranteed two-fold all through the 802.11b/g/n/air conditioning Wi-Fi on the Raspberry Pi 3 Model B. With the expansion of gigabit ethernet over USB 2.0, the wired ethernet execution is additionally supported, with an extraordinary throughput of around 300 MB

Output

Speaker: Text to speech conversion—the voice processing module transforms the .txt file to an audio output using the Festival TTS, E-speak voice synthesizer. In this, we use E-Speak TTS that needs a text file. The Raspberry Pi has an onboard audio jack, a Pulse width modulation (PWM) output generates the on-board audio. Add some extra impact to your sound undertaking with these USB controlled enhancers. We tried a huge bit of a tremendous measure of different models to find ones with a better than average repeat response in

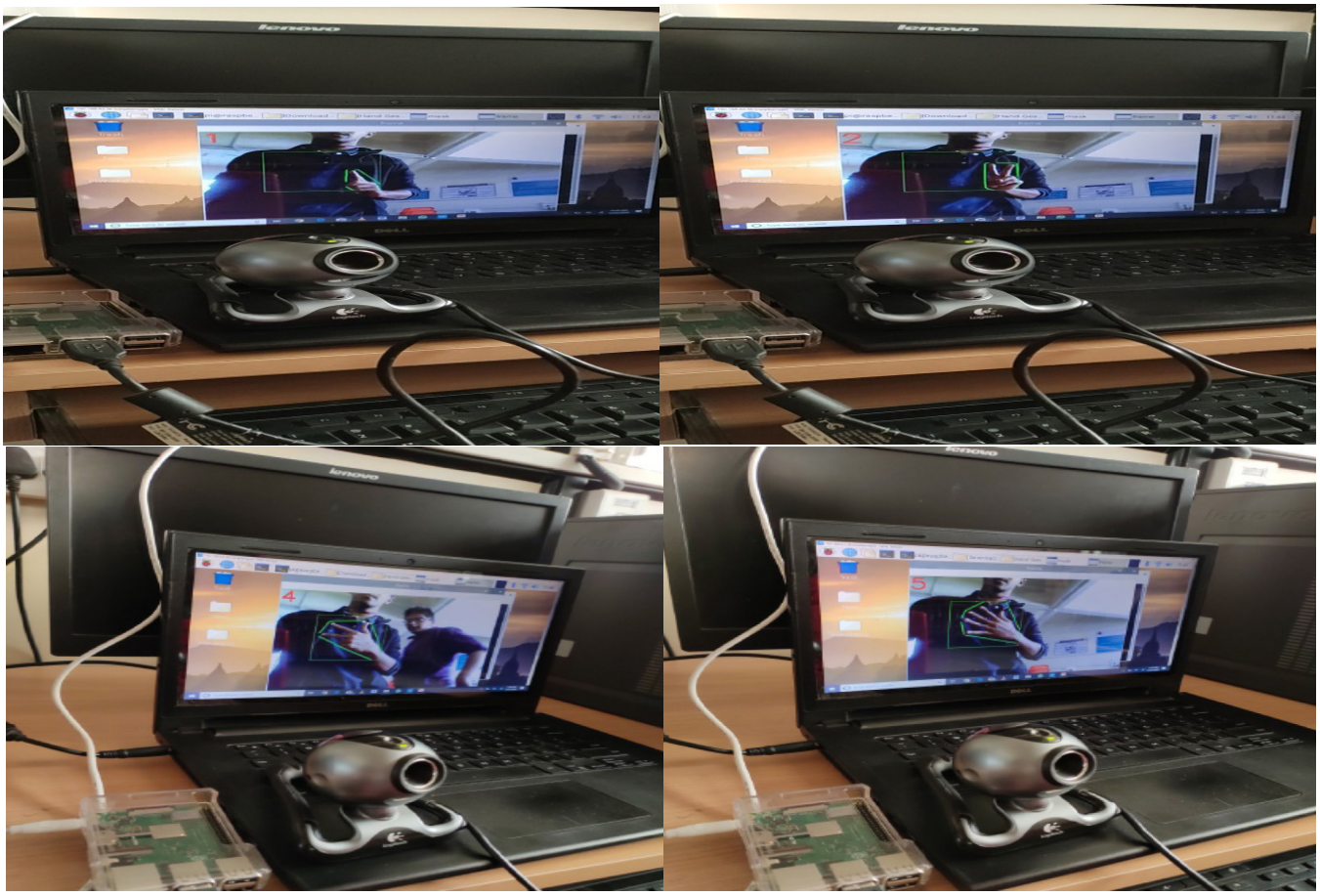

Figure 5: Identification of ISL signs

this way, you'll get quality sound yield for music playback. There is even a volume control wheel on it so you can set it up splendidly. Essentially, interface the standard 3.5 mm stereo associate with the raspberry Pi, wave shield, etc. For control, interface the USB connector to anything that can give USB control. We can normally keep the volume about halfway (which is still to a great degree riotous), where the current drawn is 200 to 400 mA. At max volume, you can end up with up to 1 Amp.

USB powered speakers feature: Completely compatible with the Raspberry Pi 100 Hz to 18 kHz recurrence reaction,

- Power necessities: 5V DC 1A crest (at max volume yield) 4 Ohm impedance speakers, 3W each.
- Speaker dimensions: 80 × 70 × 70 mm (height × width × depth)
- Link length: 1-meter; Sony, JBL, Philips are some speaker manufacturers.

Display: The purpose of attaching display is to form a complete conversation between impaired people and common people. The meaning of the signs is also displayed on the screen, as well as, the audio output is also generated.

Results

The results of this project are obtained using Raspberry pi 3B+, Logitech QuickCam pro 4000, laptop screen, and USB powered speaker. When the particular hand gesture of ISL is shown in front of the camera the screen shows the meaning of the gesture and voice output is produced with the E-speak feature of the raspberry pi. Figure 5 shows the sign recognition of ISL(Indian Sign Language) done with system.

Conclusion

The vision-based gesture recognition is more comfortable than the glove based, as the sensors attached to the gloves restrict the gestures. In vision-based, the 2D research has reached a better level in terms of accuracy and algorithm complexity, with a similar technique like HOG and SVM more accuracy can be obtained. The approach to converting ISL into voice output, as well as, displaying results on screen can improve communication between normal people and hearing/speech impaired ones. Identification of the ISL special database is also been created in this project. With the application of hand gesture recognition in the consumer and the advanced devices, the accuracy and the performance of the system may vary and the research may progress accordingly.

References

[1] Prof. Rajeshri R Itkarkar, Dr. Anil Kumar Nandi "A Survey of 2D and 3D Imaging used in Hand Gesture Recognition for Human-Computer Interaction (HCI)", IEEE International Women in Engineering (WIE) Conference on Electrical and Computer

Engineering (WIECON-ECE 2016), (19-21 December 16), DOI: 10.1109/WIECON-ECE.2016.8009115.

[2] Chen-Chiung Hsieh, Dung-Hua Liou, Novel Haar features for real-time hand gesture recognition using SVM, Real-Time Image Proceeding of Springer 2012.

[3] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[4] Chih-Hung Wu1, Wei-Lun Chen1 & Chang Hong Lin1, Depth-based hand gesture recognition, Multimedia Tools Application springer proceeding 2014.

[5] Parama Sridevi, Tahmida Islam, Celia Shahnaz, Sign Language Recognition for Speech and Hearing Impaired by Image Processing in MATLAB, IEEE Region 10 Humanitarian Technology Conference 2018, DOI:10.1109/r10-htc.2018.8629823.

[6] Reza Hassanpour, Asadollah Shahbahrami, Human-Computer Interaction Using Vision-Based Hand Gesture Recognition, Journal of Advances in Computer Research 2 (2010) 21-30.

[7] Hong-Min Zhu and Chi-Man Pun, Real-time Hand Gesture Recognition from Depth Image Sequences, 978-0-7695-4778-7/12 $26.00 © 2012 IEEE DOI 10.1109/CGIV.2012.13

[8] Nasser H. Dardas and Emil M. Petriu, Hand Gesture Detection and Recognition Using Principal Component Analysis, 978-1-61284-925-6/11©2011 IEEE.