Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 13 (2020) Pp-252-257

BBRC
Bioscience Biotechnology Research Communications

Frame Processing for Gesture Recognition Using CNN

Saba Naaz[1], K B Shiva Kumar[2] and Parameshachari B D[3]

[1]*Visvesvaraya Technological University. Belagavi, India*

[2]*Department of Electronics and Telecommunication, SIT, SSAHE, Tumkur, India*

[3]*Department of Telecommunication Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, India*

ABSTRACT

Augmented reality requires integration of real-world objects, gestures and actions with the virtual objects. Effective and realistic integration requires solving the complex task of recognition, classification, tracking of objects, gestures and actions, where gesture recognition and action mapping is an active problem in the field of augmented reality, seeking attention for optimized latency, power and throughput. This paper introduces the technique of frame processing with active tile identification to optimize the latency of Convolution neural network in the light of action mapping in augmented reality. The effectiveness of the technique being introduced is evaluated by applying it to the Bharatanatyam Mudra classification and measuring the obtained latency, power and throughput and comparing the obtained results with that of the traditional Convolution neural network. The comparison shows the technique to be effective in terms of the latency, with the best effectiveness factor of 2.30 and least being 1.25.

KEY WORDS: AUGMENTED REALITY, CONVOLUTION NEURALNETWORKS, GESTURERECOGNITION; GRAPHICALPROCESSINGUNIT, SEMANTICSEGMENTATION, STOCHASTICGRADIENTDECENTWITH MOMENTUM.

INTRODUCTION

Real-time implementation of Augmented Reality is a challenging task with the possibility of numerous applications in robotic surgery [P Pessaux et al.,2015], gaming [YAN Yukang et al.], chemical plant maintenance [V. I. Pavlovic et al.,1997)], computer control [Quek F.1994, C. Maggioni,1995], window system control [J. Kenderet al.,1995] and television control [W.T.Freeman et al.,1995] etc. The task is computationally demanding requiring optimized implementation of its integral algorithms for real time experience. Image processing and computer vision methods like semantic segmentation, blob identification, skeletal mapping is found to be the most frequently used and cited as integral part of augmented reality[Babu, R.G. et al., 2020].

The AR real time implementation is a trivial task requiring the gesture recognition and action mapping to be accomplished within the time span window of few frames of the streaming video [J. Coutaz et al.,1995]. The window is subject to constraints of rate of gestures and video frame rate. To meet this requirement of time window, algorithms need to be profiled against the parameters like latency, throughput and power. The profiling enables the designer to identify the best algorithms and their implementation strategy to meet the stringent timing requirements [S. Prabu, et al, 2019]. Unfortunately, this view of algorithm selection and strategic implementation is less treated in literature. Current paper attempts to fill this void or the gap with a case study of CNN based mudra classifier. Though the study appears to be focused on countable algorithms but the scope is generic and applicable to a wide range of algorithms and scenarios.

Later parts of the paper are organized as follows. Section II presents an overview of standard gesture recognition algorithms and neural network techniques. Section

ARTICLE INFORMATION

*Corresponding Author: hodte@gsss.edu.in
Received 8th Oct 2020 Accepted after revision 28th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)

*Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.13/37*

III discusses the experimental setup and performance parameters for evaluation of algorithms. In section IV, the strategies for optimizing algorithm implementation are presented. Section V presenting the performance parameters with optimized algorithm utility.

Gesture Recognition Algorithms: Gesture recognition has been the most studied and investigated problem in pattern recognition and image processing. A number of procedures and approaches have evolved, resulting in a different strategies and algorithms. Here a brief overview of the related works is presented.

Orientation histogram-based pattern recognition techniques has been formulated by William T Freeman and Michael Roth here histogram of local orientation is used as feature vector for recognition. The algorithm compares the feature vector of test image with the training set resulting in gesture class corresponding to the nearest match. The method and the algorithm have drawback of mismatch due to similar orientation for different gestures. Weissman and Freeman proposed a method of gesture recognition for television control applying normalized correlation [W. T. Freeman et al.,1995]. Triesch's developed Gabor filters based Elastic graphs representation of hand gestures in [C.VonDer Malsburget al.,1996], the method is quick in locating hand, but the classifier lacks generalization.

Lindberg and Lars have proposed scale-space color features technique to represent hand gestures [Lars Bretzet al.,2002]. Particle filtering was employed for detection and recognition of the hand gestures. The method is limited by scale space representation and works for uniform illumination back grounds only. [Y. Fang, etal.,2007] proposed appearance-based model for real time hand gesture recognition, Gaussian model is employed for hand region segmentation in HSV space. They employed scale-space technique for gesture recognition.

Malima and Cetin presented a simple approach through fast algorithm for vision-based hand gesture recognition. The algorithm processes the images for hand region detection applying skin color segmentation. The regions thus segmented are subject to circle construction followed by binarization and 1D signal extraction corresponding to 0 to 1 transition in the binary image, this pattern of 1D signals is used for recognizing gestures. The method works well for counting gestures but can't be generalized. [Hunteretal.,1994] presented a recursive estimation-based hand gesture interpretation method. Verri and Urras, discusses about the edge map method of hand gesture recognition [A.Verriet al.,1995]. [Chanetal.,2002] proposed curvature-based hand pose recognition.

The methods and algorithms put forth hints as, Gesture recognition to be a trivial task with the vast situations exhibiting large possible postures. Mathematically, the problem of gesture recognition in open situation corresponds to mapping an element into a superset of all possible gestures. The problem takes more typical turn with the consideration of sampling method, orientation of the object, lighting conditions and skin color variations along with object size.

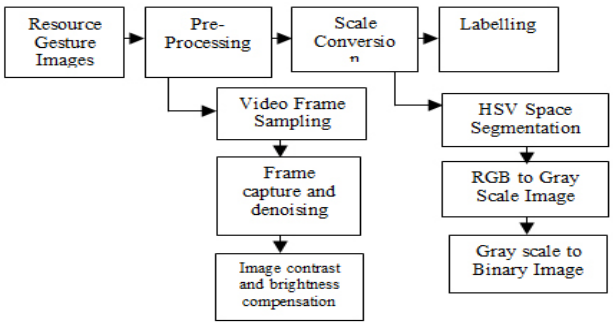
All the above-mentioned facts make the task of gesture recognition complex and open. Due to the learning and adoptability, a number of neural network-based algorithms, approaches and architectures have been proposed and employed in treating recognition and classification problems [Alex Krizhevsky et al.,2017]. Authors in [Neha Sharma et al.,2018] have presented accuracy analysis of Resnet, Googlenet, Alexnet Convolution neural networks for the CIFAR and MNIST image data set. The literature review shows a gap in performance-based analysis of the Convolution neural networks. In [Shi et al.,2017] presented a technique of integrating wavelet resolution with the Convolution neural network for the extraction of 3D feautres to improve the classification accuracy. On this line the Convolution neural network-based gesture recognition methods and algorithms can be made to perform more effective with a little attention to the inclusion of selective schemes for segmentation and processing, i.e by taking the most out of GPU processing while spending less. [T.Kaliselvi et al.,2017] shows the application of GPU in medical image analysis, while [R Aras et al.,2017] presents effective implementation of Object recognition and augmented reality applications while exploiting the GPU processing capability.

As pointed earlier, application of the traditional Convolution neural network for gesture recognition can be non-optimistic with respect to the power and latency depending on the number of sample frames processed for gesture recognition in streaming video. The next section discusses the experimental setup considered to evaluate the novel technique introduced in this paper. The techniques like frame selection scheme, active tile identification is discussed which is followed with the view of performance parameters to be considered. The points and arguments are presented through the investigation of hand gesture recognition for Bharatanatyam mudras as special case.

Experimental Setup and Performance Parameters: Bharatanatyam mudras have been taken for the case study, in particular asamyukthahastha mudras (single handed gestures) are considered. A hand focused prerecorded video with right hand mudras performed in constant background is taken for experiment. The Convolution neural network is trained for 27 mudras, each mudra is associated with 50 sample training images. The network is trained with 1350 images, the training data base is expected to be improvised with addition of more images in near future.

The gesture recognition procedure with required neural networks and algorithms are coded in MATLAB and are executed on system with core i5 CPU supported with NVIDIA GPU. Here the account of preprocessing is not discussed as the study is focused on performance optimization focusing CNN.

Figure 1: Block Diagram of Proposed System



Data sample selection and Sizing: As pointed out earlier, differential frame selection stage is added with the region of interest-based sizing to overcome the redundant processing. Frames are selected based on the difference threshold level, which is set accordance with the test analysis spread over the initial observation period (tiop), differential selection is based on equation (1). The selected frames are subject to cropping focused on ROI, with size set based on boundary of the object of interest. This is also set with reference to initial observations made.

Before computing the difference, frames are processed with histogram equalization for color and are aligned with the centroid. For successive frames with same posture the difference will be zero or negative or small positive integer as change may occur due to relative motion of hand with camera or light intensity variations. Through the experiments it is observed that the difference of>100 is good to be considered for frame selection. This observable difference is recorded almost for every 400th frames, whereas rest of frames between are same. This observation can be used to set the dynamic rate of frame selection for processing.

$$\left.\begin{aligned}&\text{Statistical Frame Difference} = \text{sum}(\text{sum}(\text{DBVFrame}))\\&\text{Frame Sample Selected} = \text{Yes SD}>\text{Threshold}\\&\qquad\qquad\qquad = \text{No SD}<\text{Threshold}\end{aligned}\right\} \quad (1)$$

DBVFrame: Difference of successive binary version of the frame, SD: Statistical Frame Difference.

Processing duplication with series of successive frames resulting in same information is redundant, leading to wastage of power and memory resources. The differential frame information can also be used to assist the classifier by considering the difference vector as training feature. Number of Frames processed every second depends on the rate of actions performed by the object of interest and the rate of video frames captured during recording or streaming. Through the successive frame analysis for motion of object interest, region of interest is identified and the sample sizing is done keeping in view of the region of interest, the resulting frame with confinement to the region of interest is referred as active tile.

Neural network setup: Convolution neural network is modeled on the basis of conceptual model presented and discussed in [Sakshi Indolia et al.,2018 and Dingjun Yu et al.,2014]. Table 1 shows the configurations and parameters of the Convolution neural network that are used for classifying the mudras. The parameters have been set with analysis of the recorded video data, this part of has been thought out to be automated later.

Table 1. Parameters of neural network.

Layer	Parametric Details
Input Image Layer	215X105 pixels
Convolution Layer	Filter Size=8X8 Number of filters=20
Activation function	ReLU
MaxPooling	2->1 with Stirde =2
Fully connected Layer	27 with 27 Classes

Training and Classification: The Convolution neural network with specifications mentioned in Table 1 is initialized with the random weights of the filter in first epoch. As discussed in the beginning of this section, prerecorded video data is used for training the Convolution neural network. The videos are sampled for training sample selection through automated script. Convolution neural network is trained with 80 epochs (set after trials). Mudras are identified and classified into 27 classes with the appropriate labels. Softmax classifier is employed to classify the mudra.

Performance Parameters: The performance of the gesture i.e., mudra recognition system discussed in this paper depends on Convolution neural network including data sampler. The performance parameters studied are power and time delay. Both the parameters are function of Convolution neural network internal parameters and size of the training feature database.

$$\text{Power} = \text{Total frames}*\text{PPF} \quad (2)$$

Equation (2) shows the power dissipated with normal approach of regular series stream processing. The differential frame selection approach results in power saving by factor of N (the differential frame skip distance, assuming to remain constant over streaming), as shown in equation (3).

$$\text{Power} = \text{Total frames}/N*\text{PPF} \quad (3)$$

The tile area i.e size in terms of pixels, selection of filter size, stride and optimization algorithm like sgdm, rsprim, Adam's has impact on the power and processing delay. A filter size with M=N=ROI/P is set and stride of M/4 is set, where P is set based on the trial of Convolution filter for the particular curve shapes, it is planned to be automated in future work. The power PPF is a function of (Filter Size, stride, epochs and OA). Time delay is also function of (Filter Size, Stride, epochs and OA). The differential

processing can be fine-tuned based on the control and actuation required, for finer actuations with continuous control less stringent difference must be selected.

RESULTS AND DISCUSSION

Figure 2(a) shows the redundancy with successive frames obtained by sampling video stream with sampling every 10th frame. The first three mudras in upper row are the same. Similarly, the next three are the copy of one mudra. The same is observed with the last two mudras. This justifies the need of differential frame selection and processing.

Figure 2 (a): Frame sampled at every 200th Stroke

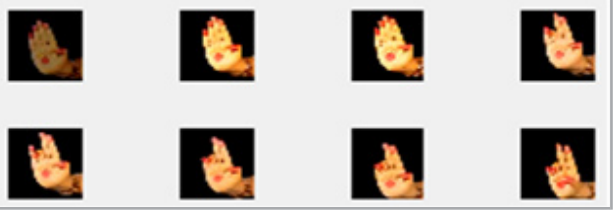
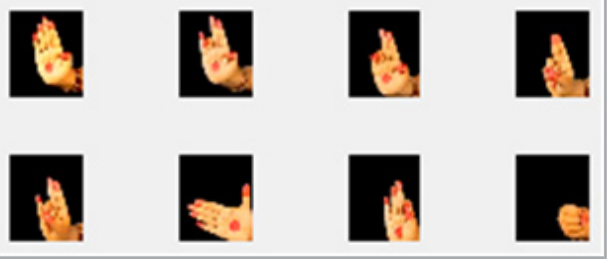


Figure 2 (b): Frame sampled at every 400th stroke.



Mudras listed in Figure 2(b) are the ones selected through the differential frame sampling. The images in both upper and lower rows clearly show, every image is unique and different mudra as opposed to the multiple copies observed in Figure 2(a). The video sample considered for study showed differential skip distance of N=400 frames. Figure 3 (a) shows the difference between two frames with distance less than skip distance, as the frames are copy of same mudras the difference is almost zero where in the less intense illuminating pixels are due to the frame color variations. Figure 3(b) shows the prominent difference with considerably illuminated pixels and shape variation, indicating two frames to be of different mudras.

Figure 4 (a), shows the normal frame and Figure 4(b) the active tile, size of the active tile is very much constrained to the region of interest with lesser number of pixels compared to the normal one.

The procedure of differential frame selection and active tile identification has been applied in the process of training as well as classification. Figure 5. Shows the graphs for accuracy of training and loss. The graphs are recorded for stochastic gradient decent with momentum optimization, the results are different for Adam optimization. stochastic gradient decent with momentum optimization had shown recorded accuracy of 83.7%, far better compared to other optimizers.

Figure 3. (a): Two frames of the same mudra with less prominent difference. (b) Two frames of different mudras with prominent difference.

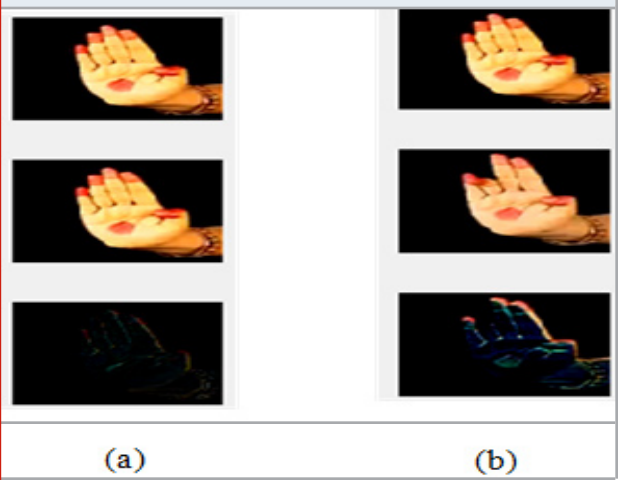


Figure 4. (a): Normal frame. (b) Active tile

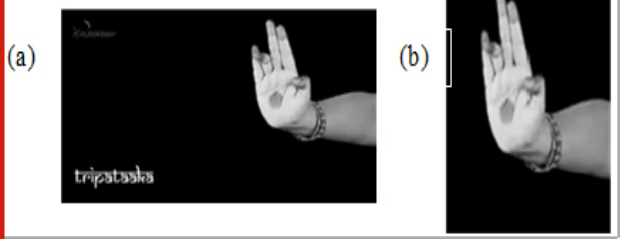


Figure 5: CNN training accuracy and loss graphs.

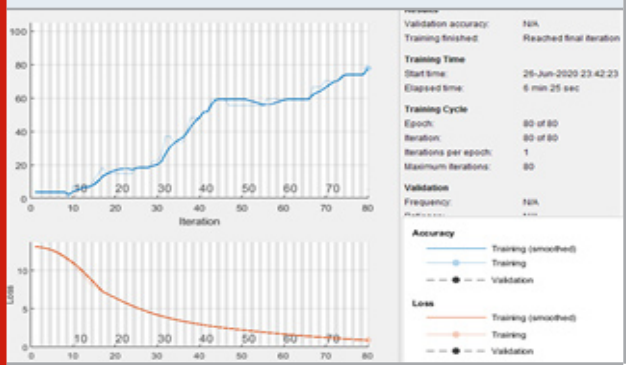


Table 3 shows the frame wise latency of the classifier in recognizing the gesture from sample frames. The latency is recorded for the two cases of with and without active tile processing. The delay is recorded for 4 instances of frames at the same stroke of time. Table 3. Mudra classification delay. The delay is in seconds.

The results of table 3, indicates that the active tile-based classification is effective in reducing latency. Though the active tile reduces time, care must be taken as it can affect accuracy of classification. With active tile, the

accuracy was observed to be around 80%, but it can be improved with increase in number of epochs.

Table 2. Classifier accuracy, mudras are labeled in order staring from D1-pathakam to D27-Trisoolam as shown in Figure 5.

Mudra	Accuracy in %
D1	89.57
D2	93.25
D15	97.32
D19	87.88

Table 3. Mudra classification delay. The delay is in seconds.

Frame number	Delay without active tile	Delay with active tile	Effectiveness
1	1.2191	0.5282	2.30
2	0.2657	0.1195	2.22
3	0.2670	0.1320	2.02
4	0.3541	0.2818	1.25

Figure 6: Asyumthahastha mudras.

Power dissipated by the approaches with and without active tile is computed applying the formula of equation (2) and (3). The data observed is for 19250 consecutive frames of the selected video stream. The average GPU Power dissipated per frame is 1.71mW[]. Power dissipated in the first scheme i.e. without active tile method is 32.918W whereas 83.79mW is the power dissipated in the second scheme with active tile method. The data recorded here is though dynamic and depends on the frame rate and rate of gestures, yet valuable power saving can be expected with active tile method.

CONCLUSION

With the integration of the differential frame and active tile technique in the Convolution neural network frame, a new perspective is brought to the gesture recognition problem. Taking the advantage of slow varying dynamic gestures and rate of video frames, redundant processing and classification delay are reduced. The experimentation results show the method to be effective with respect to latency, where the best latency factor is found be 2.30, which means the technique makes the Convolution neural network 2.30 times faster compared to the traditional neural network. Though there is improvement in latency, the study and analysis of the technique shows the gesture recognition accuracy to be function of error optimization scheme employed in training. This limitation can be due to the numerical ability of the underlying computational architecture, which can be taken as thread for the subsequent investigation. Currently the method advantageous as it is generic and can be adopted elsewhere in any of the frame processing problems.

REFERENCES

A Coates, P Baumstarck, Q Le, AY Ng, (2009) Scalable Learning for Object Detection with GPU Hardware,IEEE RSJ International Conference on Intelligent Robots and Systems,

A. Verri and C. Uras (1995) Hand Gesture Recognition From Edge Maps,Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 116-121.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, Vol 60, No.8, 2017, pp.84-90.

A.Malima, O.Erol and M.Cetin, (2017) A Fast Algorithm For Vision- Based Hand Gesture Recognition For Robot Control,

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. HintonImageNet Classification with Deep Convolutional Neural Networks, Communications of the ACM, Vol 60, No.8, pp.84-90.

C. Maggioni, (1995) GestureComputerNew Ways of Operating a Computer,Proc.Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 166-171.

C. Von Der Malsburg and J. Triesch (1996) Robust classification of hand posture against complex background,in Proceedings of Int. Conf. on Face an d Gesture Recognition. Killington, Vermont, pp. 170–175.

C.-C. Chang, I.-Y Chen, and Y.-S. Huang, (2002) Hand Pose Recognition Using Curvature Scale Space, IEEE International Conference on Pattern Recognition.

Dingjun Yu, Hanli Wang, Peiqiu Chen and Zhihua Wei, (2014) Mixed Pooling for Convolutional Neural Networks, International Conference on rough sets and knowledge technology, pp.364-375.

E. Hunter, J. Schlenzig and R. Jain, (1994) Vision-Based Hand Gesture Interpretation Using Recursive Estimation,Proc. 28th Asilomar Conf. Signals, Systems, and Computer.

J. Coutaz , F. Berard and J.L. Crowley (1995) Finger Tacking As an Input Device for Augmented Reality, Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 195-200.

J. Kender and Kjeldsen (1995) Visual Hand Gesture Recognition for Window System Control,Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 184-188.

Lars Bretzner, Ivan Laptev, and Tony Lindeberg, (2002) Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering,in Proceedings of Int. Conf. on Automatic Face and Gesture Recognition. Washington D.C. pp. 423– 428.

Neha Sharma, Vibhor Jain and Anu Mishra, (2018) An Analysis ofConvolution Neural Networks for ImageClassification, Proceedings of computer science, Volume 132, pp.377-384.

P Pessaux, M Diana, L Soler, T Piardi, D Mutter and J Marescaux,(2015)Towards cybernetic surgery: robotic and augmented reality-assisted liver Segmentectomy,Langenbeck's archives of surgery Vol 400, Issue 3, pp.381-385.

Quek F.K.H, (1994) Toward a Vision-Based Hand Gesture Interface,Virtual Reality Software and Technology Conf., pp. 17-31.

R Aras and Y Shen, (2017) GPU Accelerated Stylistic Augmented Reality, Academia.edu.

Babu, R.G., Maheswari, K.U., Zarro, C., Parameshachari, B.D. and Ullo, S.L., 2020. Land-Use and Land-Cover Classification Using a Human Group-Based Particle Swarm Optimization Algorithm with an LSTM Classifier on Hybrid Pre-Processing Remote-Sensing Images. Remote Sensing, 12(24), p.4135.

Sakshi Indolia, Anil Kumar Goswami, S.P.Mishra and Pooja Asopa, (2018) Conceptual Understanding ofConvolutionNeural Network- A Deep Learning Approach, proceedings of computer science, Volume 132, , pp.679-688.

Prabu, S., Balamurugan, V. and Vengatesan, K., 2019. Design of cognitive image filters for suppression of noise level in medical images. Measurement, 141, pp.296-301.

Shi, Cheng, Pun, Chi-Man, (2017) 3D multi-resolution wavelet convolutional neural networks for hyperspectral image classification, Information Science, Vol 420, , pp.49-65.

T.Kaliselvi, P Ramakrishnan and K.Somsundaram, (2017) Survey of using GPU CUDA programming model in medical image analysis, Informatics in Medicine Unlocked, Vol 9, pp.133- 144.

V. I. Pavlovic, R. Sharma and T. S. Huang, (1997) "Visual interpretation of hand gestures for human-computer interaction: a review," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677-695.

W. T. Freeman and C. Weissman, (1995) Television control by hand gestures, in Proceedings of International Workshop on Automatic Face and Gesture Recognition. Zurich, Switzerland,pp. 197–183.

W.T.Freeman and Micael Roth, (1995) Orientation Histograms for Hand Gesture Recognition, IEEE Workshop on Automatic Face and Gesture Recognition,.

Y. Fang, K.Wang, J Cheng and H. Lu, (2007) A Real Time Hand Gesture Recognition Method, IEEE ICME,pp.995-998.

YAN Yukang, YI Xin, YU Chun, SHI Yuanchun, Gesture-based target acquisition in virtual and augmented reality,Virtual Reality & Intelligent Hardware, Vol.1, Issue 3, pp. 276-289.