# Neural Network Performance Analysis for Real Time Hand Gesture Tracking Based on Hu Moment and Hybrid Features

Mehdi Heidaryan[1], Fardad Farokhi[2], Kave Kangarloo[3]

[1,2,3]Electrical Engineering Department, Central Tehran Branch Islamic Azad University, Tehran, Iran. Email: Mah.heidaryan.eng@iauctb.ac.ir, Email: f_farokhi@iauctb.ac.ir. Email: Kangarloo@iauctb.ac.ir

**Abstract**

This paper presents a comparison study between the multilayer perceptron (MLP) and radial basis function (RBF) neural networks with supervised learning and back propagation algorithm to track hand gestures. Both networks have two output classes which are hand and face. Skin is detected by a regional based algorithm in the image, and then networks are applied on video sequences frame by frame in different background (simple and complex) with different illumination of environment to detect face, hand and its gesture. The number of training and testing samples in networks are equal and the set of binary images obtained from skin detection method is used to train the networks. Hand gestures are 6 cases which are tracked and they were not recognized. Both left and right hands has been trained to the network. Network features are based on the image transforms and they should not relate to deformation, size and rotation of hand. Since some of the features are in common with each other so a new method is applied to reduced calculation of input vector. Result shows that MLP has high accuracy and higher speed in tracking hand gesture in different background with minimum average error but it has a lower speed in training and convergence compare to the RBF in its final average error.

***Keywords:*** MLP, RBF, Skin Detection, Invariant moments, wavelet transform, DFT, DCT



## 1. Introduction

Hand detection and tracking is one of the communications ways between Human and Computer (HCI) and has many applications in various fields [1]. Lots of the researches have been done in this area that have detected and tracked hand. According with [2] a large number of methods have been proposed in detection of hands and the segmentation of the corresponding image regions (Vision based hand gesture recognition for human computer interaction). Extracted features were skin color, shape, motion and anatomical models of hands, but problems such as illumination variations and the size of the tracked objects, especially hand is still remained as challenges. Therefore methods must be applied in order to reduce the unwanted variations among images causes. Neural network especially MLP can be used in filtering, image segmentation, pattern recognition and object detection [3]. Neural network is noise tolerant and can learn and generalize even in a noisy environment so it can increase the obtained accuracy and have an acceptable performance together with a higher processing speed. The neural networks such as MLP and RBF was used to resolve the mentioned problem and they have acceptable performance in the cases which there is not exists a clear solution [4]. MLP with back propagation algorithm in comparison with the K-means is more accurate but a little more time-consuming in training phase [5]. In the cases which the output of the network have 2 classes, RBF and MLP networks reaches their highest accuracy but in multi class cases, when for each hand gesture one different classes is assigned, the accuracy decreases dramatically [6] so in this paper, all hand gestures are considered in a class and the face is selected as the other output classes. In RBF network, convergence

rate is very high, and the value of average error is reached to its final value, and in order to decrease the mean square error, more epochs is needed which is lead to an unacceptable increase in the elapse time. Because of the mentioned problem better accuracies were reported previously [7-8]. In the other hand feature selection in tracking hand is important because not only the features must described hand gestures and face accurately but also they should be limited in order to be implemented for real-time processing. These features need to be invariant to rotate, resize and scale. In terms of hardware implementation MLP and RBF due to the simplicity and speed of processing and learning become one of the most widely used networks in this field which require less memory [9-10]. Although the MLP and RBF network has a moderate accuracy for face and hand recognition among the other networks such as support vector machine and self-organized feature map (SVM, SOFM...) but by selecting appropriate feature set using feature selection approaches is improved the obtained overall accuracy which is not so easy for other types of the classifiers[11,12]. Fig 1 shows the block diagram of hand gesture detection method of this paper. There are lots of research papers that including hand tracking using neural network and here are some recently proposed papers: E. Stergiopoulou, N. Papamarkos in [13], first the skin is segmented with a threshold in YCbCr then the algorithm is proceeded to recognize hand gestures through the Self- Organized Neural Gas (SGONG) network. This algorithm has been implemented on image database with a fix background and it is not tested over video records. T. R. Trigo and S. Roberto in [14] recognized and segmented hand gestures and extracted a set of features. These features were examined separately and the algorithm is tested on captured binary images of hand gestures. M. K. Kiran and T. S.Vamsi in [15] detected bare hand gestures using the affine-scale invariance feature transform (ASIFT) and Extreme Learning Machine (ELM) for training purpose. Result shows that features are invariant to scale and rotation. For skin detection they used HSV color space and their test is done on a fixed background. Azadeh Kiani in [16] attempted to identify static hand gestures for Persian Sign Language and used MLP and discrete wavelet transform as its feature for classified and claimed that the obtained accuracy was 98.75%. C.T. Hsieh in [17] has attempted to detect real time hand gesture system Based on DFT and SVM. Skin detection is done with cam shift algorithm and they have claimed that they reached to 93.4% accuracy. P.V.V Kishore in [18] detected hand gestures and its shapes for Indian sign language with Gabor filter in eight different orientations and a feed forward neural network with back propagation algorithm for classification in fixed background. They have reached the average recognition rate at 98.2%.

## 2. MLP

MLP with the most widely used training algorithm named back propagation because of its simplicity and appropriate accuracy becomes one of the most widely implemented network in classification [19-21]. So in this section, a brief description of the types of MLP that is used in this approach including the weights update, output error, back propagation algorithm and activation function is given. The weights are updated in sequential mode i.e. weights are updated in every training example. The correction of weights is given by (1).

$$\Delta w_{ij}(n) = \eta \delta_j(n) y_i(n) \quad (1)$$

Where $\eta$ is learning rate and $\delta_j(n)$ is local gradient and can be obtained from (2):

$$\delta_j(n) = \varphi'_j\left(v_j(n)\right) \sum_k \delta_k(n) w_{kj}(n) \quad (2)$$

Where $\varphi'_j$ is the derivative of the activation function and finally, an error signal in iteration $n$ is obtained from (3) and (4) and after divided by the number of instances we get the average error in (5).

$$e_j(n) = d_j(n) - y_i(n) \quad (3)$$

$$E(n) = \frac{1}{2}\sum_{j=c} e_j^2(n) \quad (4)$$

$$E_{av} = \frac{1}{N}\sum_{n=1}^{N} E(n) \quad (5)$$

Where $d_j$(n) is the desire value, set C includes all the neurons in the output layer of the network and N denote the total number of examples contained in training set [22]. There are a lot of activation functions [23], but here we use hyperbolic tangent obtained from (6), because its derivative is easy to compute, error functions is less than linear and Gaussian and can be expressed directly as a function of the net input [24].

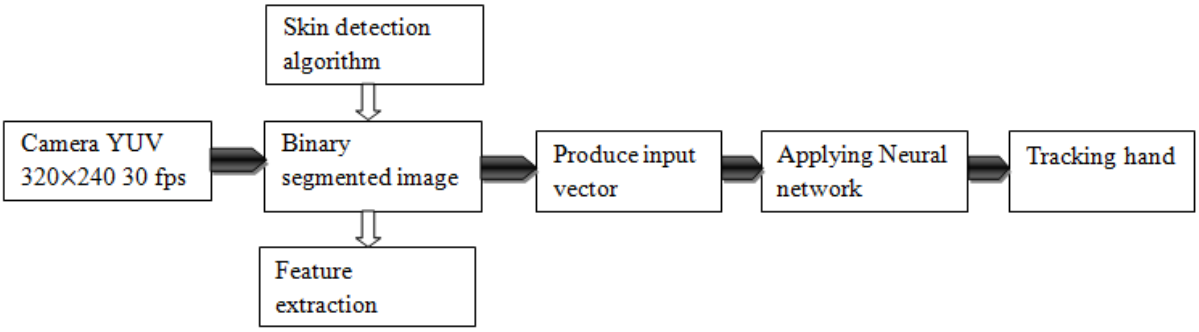


Fig.1. The block diagram of hand gesture detection method.

$$\tan h\, x = a\left(\frac{1-e^{-2bx}}{1+e^{-2bx}}\right) \quad (6)$$

Where our experience shows that the best values for a and b, are 1.716 and 0.667 respectively. The used MLP network regardless of the input has 3-layer, input, two hidden and output. Existence of the two hidden layers increases the training time compared to the RBF But on the other side also enhances accuracy. On the other hand the number of neurons in each hidden layer is 40 and network has two bit output (two neurons), first class in binary is one (01) and second is two (1 0). Chosen two outputs are because of more accurate thresholding in network, increase the accuracy and power of decision. The number of neurons in network is x-40-40-2 that x define the number of features (input vector) in network.

## 3. RBF

We used improved Radial Basis Function with supervised learning algorithm for classification [25]. The radial basis function technique consists of choosing a function that has the following form:

$$F(x) = \sum_{i=1}^{m} w_i \varphi_i(x) \quad (7)$$

Where $\{\varphi_i(x) \mid i = 1, 2,\ldots,m)\}$ is a set of m random nonlinear functions, known as radial basis functions that m≤N and N is the input vector. $\varphi_i(x)$ is obtained from (8):

$$\varphi_i(x) = G\|x - t_i\| \quad (8)$$

Where G is green function and centered at $t_i$ and the multivariate Gaussian function define by (9) is used.

$$G\|x - t_i\| = \exp\left(-\frac{\|x - t_i\|^2}{2\sigma_i^2}\right) \quad (9)$$

That $t_i$ denotes the center of the function and $\sigma_i$ denotes its width. In the supervised learning algorithm the same as MLP we have desire value $d_j(n)$ and so the error is defined by (10) and average error is obtained from (5).

$$e_j(n) = d_j(n) - \sum_{i=1}^{m} w_i G\|x - t_i\|_{C_i} \quad (10)$$

The requirement is to find the free parameters $w_i$, $t_i$ and $\sum_i^{-1}$ (the latter being related to the norm-weighting matrix$C_i$).The update equation for $w_i$, $t_i$ and $\sum_i^{-1}$ are assigned different learning rate $\eta_1$, $\eta_2$ and $\eta_3$ respectively so correction of weights, $t_i$ and $\sum_i^{-1}$ are obtained from (12), (15) and (18).

$$\frac{\partial E(n)}{\partial w_i} = \sum_{j=1}^{N} e_j(n)\, G\left\|x_j - t_i(n)\right\|_{C_i} \quad (11)$$

$$w_i(n+1) = w_i(n) - \eta_1 \frac{\partial E(n)}{\partial w_i} \quad (12)$$

$$\frac{\partial E(n)}{\partial t_i} = 2w_i(n) \sum_{j=1}^{N} e_j(n)\, G'\left\|x_j - t_i(n)\right\|_{C_i} \sum_i^{-1} R_{ij}(n) \quad (13)$$

$$R_{ij}(n) = [x_j - t_i(n)] \quad (14)$$

$$t_i(n+1) = t_i(n) - \eta_2 \frac{\partial E(n)}{\partial w t_i} \quad (15)$$

$$\frac{\partial E(n)}{\partial \sum_i^{-1}(n)} = -w_i(n) \sum_{j=1}^{N} e_j(n)\, G'\left\|x_j - t_i(n)\right\|_{C_i} Q_{ij}(n) \quad (16)$$

$$Q_{ij}(n) = [x_j - t_i(n)]\left[x_j - t_i(n)\right]^T \quad (17)$$

$$\sum_{\mathrm{i}}^{-1}(\mathrm{n}+1) = \sum_i^{-1}(n) - \eta_3 \frac{\partial E(n)}{\partial \sum_i^{-1}(n)} \quad (18)$$

## 4. Skin Color Classification

There are some color spaces that have good performance in noise and light variations and have an independent component of the light intensity such as YCbCr and YUV that one of the best color spaces which have been used in skin detection [13, 26] but changing RGB to YCbCr has arithmetic calculation that lead to reduce the speed processing in real time so we use the color space in [27] that named YrUrVr or modified YUV. According to [27] there are two kinds of colour spaces that define by JPEG2000 standard [28] the irreversible component transformation (ICT) and the reversible component transformation (RCT). These colour spaces are commonly used in image and video coding applications. Irreversible means it is impossible to reconstruct the image with integer precision. Since RCT is a lossless coding technique with the ability to reconstruct the image with integer precision. YCbCr colour space is an example of ICT and YrUrVr is an example of RCT. The forward and inverse RCT transformation (YrUrVr) obtained from (19) and (20):

$$\begin{pmatrix} Yr \\ Ur \\ Vr \end{pmatrix} = \begin{pmatrix} \frac{[R+2G+B]}{4} \\ R-G \\ B-G \end{pmatrix} \quad (19)$$

$$\begin{pmatrix} G \\ R \\ B \end{pmatrix} = \begin{pmatrix} Yr - \frac{[Ur+Vr]}{4} \\ Ur+G \\ Vr+G \end{pmatrix} \quad (20)$$

The maximum and minimum values of Ur and Vr of each skin region window were derived from a data base in [16] obtained from following relations:

$$Min(Ur) = \overline{Ur} - stdDev(Ur) \quad (21)$$

$$Max(Ur) = \overline{Ur} + stdDev(Ur) \quad (22)$$

$$Min(Vr) = \overline{Vr} - stdDev(Vr) \quad (23)$$

$$Min(Vr) = \overline{Vr} + stdDev(Vr) \quad (24)$$

Where $\overline{Ur}$ and $\overline{Vr}$ denote the mean of $Ur$ and $Vr$ and stdDev means standard deviation of $Ur$ and $Vr$.The result of testing on images are shown that $10 < Ur < 74$ and $40 < Vr < 11$. After skin pixels were converted to the modified YUV space, and be segmented based on the obtained threshold we find that the blue channel had the least contribution to human skin color additionally, leaving out the blue channel would have little impact on thresholding and skin filtering. This also implies the insignificance of the V component in the YUV format. Therefore, the skin detection algorithm using here was based on the U component only. Applying the suggested threshold for the U component would produce a binary image with raw segmentation result, as depicted in. Skin detection method have false positive due to existing color similar to skin so noise pixels appears in segmented image. Because of that in addition to use morphological filters including erosion and hole filling, after each group of detected pixels became one connected region, connected component labelling algorithm was applied. This process labelled each connected region with a number, allowing us to distinguish between different detected regions. This method performed two tasks 1: extract the area information of each labelled region 2: store the areas of all the labelled regions in the array in the order of their labels it means that for example area 1 has 120 pixels. Area-based Filtering applied to image because filtering detected regions based on their areas would successfully remove all background noise and any skin region that was not likely to be a face or hand. Fig 2 shows the segmented image with this skin detection method.

## 5. Features extraction

### *5.1. Hu invariant moments*

Moments are properties of an image and important descriptions of the region. Numerous algorithms and techniques use moments for pattern recognition, object identification, 3-D object pose estimation, robot sensing, image coding, and reconstruction and motion detection too[29]. The Hu's moments are just one of the many different types of moments that are being used in pattern recognition. The various types of moments include Zernike moments, pseudo-Zernike moments, Legendre moments, rotational moments, and complex moments. Moments are very useful because their computation is algorithmically simple and uniquely defined for any image function. Moment-based methods often yield features from an image that are invariant to translation, scaling, shift, noise and rotation [30, 31]. In recent years this descriptor besides the other features is used frequently to identify the hand gestures and can make a robust neural network with high accuracy for pattern recognition [29, 30, 32]. Moments of Order p+q and Central moments are defined by the following relations:

$$m_{pq} = \sum_x \sum_y x^p \cdot y^q \cdot f(x,y) \quad (25)$$

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p \cdot (y-\bar{y})^q \cdot f(x,y) \quad (26)$$

$$\bar{x} = \frac{m_{10}}{m_{00}} \ , \ \bar{y} = \frac{m_{01}}{m_{00}} \quad (27)$$

Where $m_{00}$ In binary image represent area and $m_{01}$ and $m_{10}$ present the center of gravity of image. Normalized central moments define with (28) and (29) and with changing the image size will not change.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad (28)$$

$$\gamma = \frac{p+q}{2} + 1 \quad (29)$$

From the central moments, Hu derives a set of seven moments, also known as Hu's moments. The following equations shows this set of seven moments given that higher moments are usually sensitive to noise, so only up to the seventh-order moments are calculated as image descriptors as following equations:

$$\varphi_1 = \eta_{20} + \eta_{02} \quad (30)$$

$$\varphi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}{}^2 \quad (31)$$


Fig.2. Binary image result of skin detection method. (A) U component threshold on video sequence. (B) Morphological filters including erosion and hole filling. (C) Area-based Filtering applied to image that remove background noise.

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \quad (32)$$

$$\varphi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (33)$$

$$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} - \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \quad (34)$$

$$\varphi_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (35)$$

$$\varphi_7 = (3\eta_{12} - \eta_{30})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[3(\eta_{03} + \eta_{12})^2 - (\eta_{12} + \eta_{03})^2\right] \quad (36)$$

*5.2. Wavelet transform*

Wavelet transform is one of the important techniques to describe the signal in time (Location) and frequency domains. Wavelet transform can be used on desired signal with different degrees and frequency content of signal be investigated and the resulting coefficients of wavelet transform also described the frequency and time (place) content. In this paper the two dimensional discrete wavelet transform (2-D DWT) with Haar wavelet is used that can be considered as a set of filters that are applied on the image with the different resolution levels and obtained from (37) and (38). This kind of two-dimensional DWT leads to a decomposition of approximation coefficients at level j in four components: the approximation at level j + 1, and the details in three orientations (horizontal, vertical, and diagonal) so four matrixes is produced that shown in Fig3. Wavelet transform coefficients are frequently used in train networks for detection hand and face [33].But in this paper unlike [34-35] coefficients are not feed to network directly but first the mean of coefficients of each matrix (LL, LH, HL, HH) is taken then these means are fed to network as an input vector. Our experience show that the accuracy of network increase with this method.

$$\psi_{m,n}(x) = a^{-\frac{m}{2}}\psi\left(\frac{x}{a^m} - nb\right) \quad (37)$$

$$\psi(x) = \begin{cases} 1 & 0 < x < 1/2 \\ -1 & 1/2 < x < 1 \\ 0 & otherwise \end{cases} \quad (38)$$

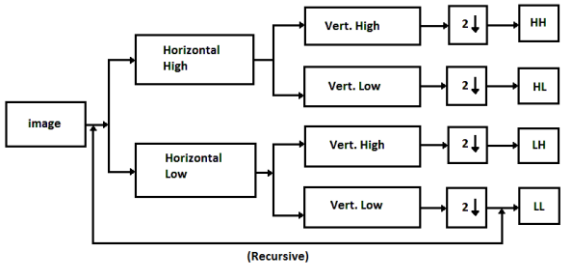


Fig.3. The matrixes of images obtained by applying the wavelet transform.

*5.3. 2-D Discrete Fourier Transform (DFT)*

Another transform is DFT that can be used for disposal of any signal [36-37]. Displacement and rotation of image will be appearing as phase delays in frequency domain [38]. However, since only the magnitude of the Fourier coefficients is considered, the phase (or equivalently, the rotation) is ignored. The benefit of this method is invariance to deformation and rotation for especially hand gesture feature [6, 17, 38]. In the other side FFT algorithm is used that fast and efficient way of calculating Discrete Fourier Transform, which reduces number of arithmetical computations from (N2) to (N*log2N). Key of the algorithm is data reorganization and further operations on it. The fallowing equations show two-dimensional Discrete Fourier Transform in an M × N image.

$$F(k,l) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi\left(\frac{kx}{M}+\frac{ly}{N}\right)} \quad (39)$$

$$f(x,y) = \frac{1}{\sqrt{MN}} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} F(k,l) e^{j2\pi\left(\frac{kx}{M}+\frac{ly}{N}\right)} \quad (40)$$

That we have: $K$, =0, 1,…, $M-1$ $l$,=0,1,…,$N-1$

20 coefficients are used in feature vector.

*5.4. Discreet Cosine Transform (DCT)*

DCT has become one of the most useful features of the image which is further used for image compression DCT transform coefficients, are similar to the real part of the DFT, as we shall see further this property reduce the computing time for the network input vector. In DCT the primary coefficients in the image contains the most image data so we can only use these coefficients as features for image and reduced input vector and increase the speed of the processing. This method is also suitable for hardware implementation [39, 40]. Direct and inverse transformations relations for 2-D DCT for M×N image size are defined as follows:

$$t(u,v) = \frac{1}{4}\alpha(u)\alpha(v)$$

$$\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\left(\frac{(2x+1)u\pi}{16}\right) \cos\left(\frac{(2y+1)v\pi}{16}\right)$$

(41)

$$f(x,y) = \frac{1}{4}\alpha(u)\alpha(v)$$

$$\sum_{u=0}^{M-1} \sum_{v=0}^{N-1} t(u,v) \cos\left(\frac{(2x+1)u\pi}{16}\right) \cos\left(\frac{(2y+1)v\pi}{16}\right)$$

(42)

That we have:

$$\alpha\,(u(v)) = \begin{cases} \sqrt{\frac{1}{M(N)}} & u\,(v) = 0 \\ \sqrt{\frac{2}{M(N)}} & u(v) = 1,2,\dots,M(N)-1 \end{cases} \quad (43)$$

The same as DFT 20 coefficients are used and now 51 features are extracted (explain in section 6).

*5.5. Relation between DCT and DFT*

As previously mentioned the magnitude of DFT due to invariant of rotation and scale as moments is used so some coefficients are common with DCT according to below relations. According to [39] if equation (41) written as (44) that we have:

$$t(u,v) = \frac{1}{2}\alpha(u) \sum_{x=0}^{N-1} [\frac{1}{2}\alpha(v)$$

$$\sum_{y=0}^{M-1} f(x,y) \cos\left(\frac{(2y+1)v\pi}{16}\right)] \cos\left(\frac{(2x+1)u\pi}{16}\right) (44)$$

So it can be written as follows:

$$t(u,v) = c(u) f c(v) \quad (45)$$

$$c(u) = \alpha(u) \cos\left(\frac{u\pi}{16}\right) \quad (46)$$

$$c(v) = c(u)^T \quad (47)$$

$$t(u,v) = c(u) f c(u)^T \quad (48)$$

An 8 * 8 Windows is considered for1-D DFT, (N = 8) and the factor of $\frac{1}{\sqrt{N}}$, skipped in direct conversion so (49) is decomposition of 1-D DFT, into real and imaginary part:

$$F(k)=\sum_{x=0}^{7} f(x)\cos\left(\frac{2\pi kx}{N}\right) - j\sum_{x=0}^{7} f(x)\sin\left(\frac{2\pi kx}{N}\right) \quad (49)$$

$$F(k)=F_r(k)-jF_i(k) \quad (50)$$

Thus the real part is:

$$F_r(0)=f(0)+f(1)+f(2)+f(3)+f(4)+f(5)+f(6)+f(7) \quad (51)$$

$$F_r(1)=f(0)+f(1)\cos\frac{\pi}{4}+f(2)\cos\frac{\pi}{2}+f(3)\cos\frac{3\pi}{4}+f(4)\cos\pi+f(5)\cos\frac{5\pi}{4}+f(6)\cos\frac{3\pi}{2}+f(7)\cos\frac{7\pi}{4} \quad (52)$$

$$F_r(7)=f(0)+f(1)\cos\frac{7\pi}{4}+f(2)\cos\frac{7\pi}{2}+f(3)\cos\frac{21\pi}{4}+f(4)\cos 7\pi+f(5)\cos\frac{35\pi}{4}+f(6)\cos\frac{21\pi}{2}+f(7)\cos\frac{49\pi}{4} \quad (53)$$

If coefficients are written based on $c_4$ then we have:

$$F_r(1)=f(0)+f(1)\,c_4+f(2)\,2c_4+f(3)3c_4+f(4)4c_4+f(5)5c_4+f(6)6c_4+f(7)7c_4 \quad (54)$$

$$F_r(7)=f(0)+f(1)\,7c_4+f(2)\,6c_4+f(3)5c_4+f(4)4c_4+f(5)3c_4+f(6)2c_4+f(7)c_4 \quad (55)$$

So the real part of the Fourier coefficients is obtained using the DCT coefficients ($c_4$) without re-calculating. This method is suitable for real time processing because calculation reduced and from viewpoint of hardware implementation because of parallel computing of DFT and DCT in same time we could skip some terms of DFT real part that common with DCT and the multipliers are reduced.

## 6. Network Input Vector

In this section an input vector or feature vector is created for neural networks. As mentioned four features from wavelet transform, 20 from cosine and Fourier transform and 7 features from Moments are fed to network. According to Fig 3, the matrix LL, LH, HL and HH for 320× 240 image size is defined as (56) and matrix S is derived from LL, finally $F_1$ is created:

$$LL=\begin{bmatrix} a_{1,1} & \cdots & a_{1,160} \\ \vdots & \ddots & \vdots \\ a_{120,1} & \cdots & a_{120,160} \end{bmatrix} \quad (56)$$

$$S=\begin{bmatrix} s_1 \\ \vdots \\ s_{120} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{160} a_{1,i} \\ \vdots \\ \sum_{i=1}^{160} a_{120,i} \end{bmatrix}, F_1=\sum_{i=1}^{120} s_i \quad (57)$$

Similarly for the matrices LH, HL and HH,$F_2$,$F_3$, $F_4$ are obtained. For the Fourier coefficient, first the FFT matrix of image is created in (58) and matrix D in (59) then matrix H contain 20 features as following:

$$FFT=\begin{bmatrix} b_{1,1} & \cdots & b_{1,240} \\ \vdots & \ddots & \vdots \\ b_{320,1} & \cdots & b_{320,240} \end{bmatrix} \quad (58)$$

Again similar with matrix S we have matrix D:

$$D=\begin{bmatrix} d_1 \\ \vdots \\ d_{320} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{240} b_{1,i} \\ \vdots \\ \sum_{i=1}^{240} b_{320,i} \end{bmatrix} \quad (59)$$

20 features is needed so 320 row of matrix D divided in 16 it means every 16 of matrix add to each other and H matrix is achieved:

$$H=\begin{bmatrix} h_1 \\ \vdots \\ h_{20} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{16} d_i \\ \vdots \\ \sum_{i=314}^{320} d_i \end{bmatrix} \quad (60)$$

Features of DCT is the similar to DFT with the difference that first coefficient bring directly to input vector because the first coefficient of DCT is contained more information about image.

$$DCT=\begin{bmatrix} c_{1,1} & \cdots & c_{1,240} \\ \vdots & \ddots & \vdots \\ c_{320,1} & \cdots & c_{320,240} \end{bmatrix}, V=\begin{bmatrix} v_1 \\ \vdots \\ v_{320} \end{bmatrix} = \begin{bmatrix} \sum_{i=2}^{240} c_{1,i} \\ \vdots \\ \sum_{i=1}^{240} c_{320,i} \end{bmatrix} \quad (61)$$

$$Z=\begin{bmatrix} z_1 \\ \vdots \\ z_{20} \end{bmatrix}=\begin{bmatrix} c_{1,1} \\ \sum_{i=17}^{32} v_i \\ \vdots \\ \sum_{i=314}^{320} v_i \end{bmatrix} \quad (62)$$

Finally with the Hu invariant moment that obtained from (30) to (36) the input vector with size of 51 for all features together is created in below:

$$Input\ Vector = [\varphi_1, \dots, \varphi_7, F_1, F_2, F_3, F_4, h_1, h_2, \dots, h_{20}, z_1, z_2, \dots, z_{20}] \quad (63)$$

## 7. Result of Networks

Training network is done by 340 images that obtained from skin detection method section 4 and show in Table.1. All directions of left and right hand is applied. The networks are trained with these images for all modes of features. Due to the best values of random initial weights obtained, each of the networks has been run several times. Two Curves in Fig 4 shows the noise and convergent of MLP and RBF neural network. RBF curve is for hand class and MLP curve for both hand and face classes. The vertical axis displays the average error ($E_{av}$) and the horizontal axis indicates the number of iterations (epoch). Against MLP, RBF network is very noisy in training and fluctuations in the value of the error show that and error value is greater than MLP and the number of iterations to converge is much more. Networks are implemented on dual core Intel CPU, 3230M, 2.6 GHz with total 400 Mb RAM usages. Table 2, 3 and 4 show the result of MLP and RBF for 12 different case of features (alone, 2 folded, 3 folded and all together). In these tables TP means true positive, TN true negative, FP false positive and FN false negative. As can be seen the average of all accuracies of MLP is higher than RBF but as individual accuracy of RBF is higher and is 100% related to wavelet, wavelet & DFT, wavelet & DCT. The amount of epoch shows that RBF has delay to Convergent with higher average error than MLP. Average error of RBF network is very fast reaches 0.3 (much faster than MLP), but to lower than that the number of repetition errors should be over. In both networks wavelet transform is one of the good features with high accuracy, when integrates with any other feature the accuracy of network is high and acceptable however in MLP, DCT, is less accurate when integrate with Hu make 74% accuracy and in RBF, 69% as well. All features together have 86% for MLP and 84% for RBF accuracy. We use video image from a webcam in YUV space and works with 1024 × 768 resolution and to speed up, resolution is reduced to 320 × 240. The practical result on the video sequence is shown in Fig.5 for two different backgrounds which are divided into simple, and complex. These results are done with MLP with 96% accuracy. As can be seen in the background, light and color are variable and color similar to skin is used for test skin detection algorithm but the results are constant and are invariant to rotation of hand in good accuracy.

## 8. Conclusion

This paper presents hand gestures tracking method with MLP and RBF neural network. Skin detection is done with YrUrVr color space and area base filtering that brings down false positive due to color similar to skin. Both networks have run 20 times to obtain best result and initial random weights and have two classes one is hand gestures and another face. The features selected in way that have invariance to rotation, size and scale and have common terms with each other like DFT and DCT coefficient so the process is faster with more frame rate. RBF achieve Average error of 0.3 very quick but to reduce this value down to 0.2 it requires much more iteration which have noise on data base but the MLP has lower average error than RBF with less iteration. The gestures are limited to 6 cases (close, open, victory, 1 finger, 3 fingers and first and forth finger) and are just tracked not detected. The best obtained accuracy for MLP is 99% and RBF 100% in same data base but different feature. (Wavelet & DFT for MLP and Wavelet & DFT, Wavelet & DCT for RBF). The MLP network is tested on video sequence with different background and illumination.

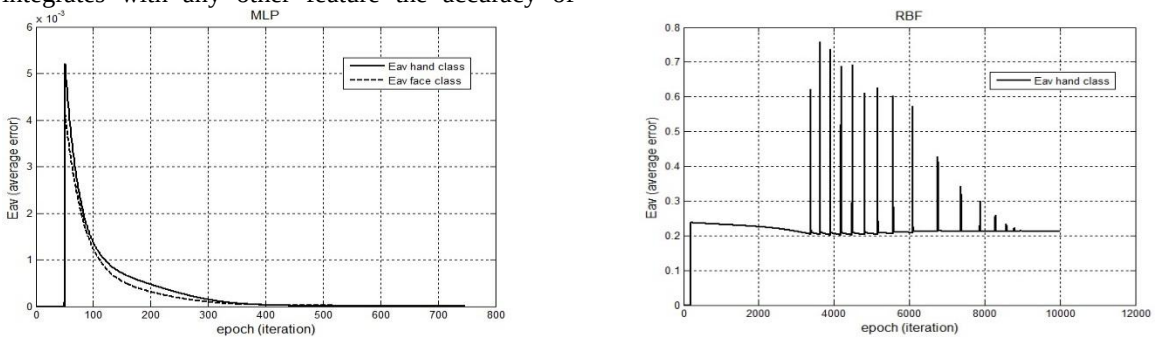


Fig 4. The RBF and MLP curves that show the Eav (average error) with number of epoch. Training in RBF is noisy that Eav has range from 0.2 to 0.75 but MLP curve shows Eav is bring down from $5 \times 10^{-3}$ to $1 \times 10^{-12}$ without any noise.

Table.2
Result of MLP and RBF with the features alone.

| | RBF | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Wavelet& DFT** | **Wavelet& DCT** | **Hu & DFT** | **Hu & DCT** | **DFT & DCT** | **Hu & DCT &Wavelet** | **Hu & DFT &Wavelet** | **Hu & DCT&DFT &Wavelet** |
| **Accuracy %** | 100 | 100 | 70 | 69 | 76 | 95 | 93 | 84 |
| **Sensitivity Hand class%** | 100 | 100 | 67 | 85 | 84 | 93 | 88 | 86 |
| **Sensitivity Face class%** | 100 | 100 | 87 | 70 | 81 | 100 | 96 | 81 |
| **Specificity Hand class%** | 100 | 100 | 75 | 47 | 58 | 100 | 96 | 81 |
| **Specificity Face class%** | 100 | 100 | 20 | 63 | 64 | 93 | 88 | 86 |
| $FN_{hand}$ | 0 | 0 | 4 | 3 | 4 | 2 | 4 | 5 |
| $FN_{face}$ | 0 | 0 | 2 | 8 | 5 | 0 | 11 | 3 |
| $FP_{hand}$ | 0 | 0 | 2 | 8 | 5 | 0 | 11 | 3 |
| $FP_{face}$ | 0 | 0 | 4 | 3 | 4 | 2 | 4 | 5 |
| $TN_{hand}$ | 24 | 22 | 6 | 7 | 7 | 13 | 9 | 13 |
| $TN_{face}$ | 22 | 24 | 1 | 5 | 7 | 31 | 4 | 31 |
| $TP_{hand}$ | 22 | 24 | 8 | 71 | 21 | 31 | 17 | 31 |
| $TP_{face}$ | 24 | 22 | 13 | 19 | 21 | 13 | 22 | 13 |
| **Number of Features** | 24 | 24 | 27 | 27 | 40 | 31 | 31 | 51 |
| **Number of neurons** | 24-20-2 | 24-20-2 | 27-20-2 | 27-20-2 | 40-20-2 | 31-20-2 | 31-20-2 | 51-20-2 |
| **Number of epoch** | 9000 (hand and face) | 9800 (hand) 9000 (face) | 9000 (hand and face) | 6700 (hand) 23000 (face) | 19000 (hand and face) | 9000 (hand 14000 (face) | 14000 (hand 9000 (face) | 14000 (hand 10000 (face) |
| $E_{av}$ | 0.2408 (hand) 0.2393 (face) | 0.11 (hand) 0.10 (face) | 0.22 (hand) 0.20(face) | 0.2359 (hand) 0.2299(face) | 0.1987(hand) 0.1944 (face) | 0.1635(hand) 0.1594 (face) | 0.2335(hand) 0.2401 (face) | 0.2235(hand) 0.2128 (face) |

Table.3
Result of RBF with the features of dual, trine and all together modes.

| | MLP | | | | RBF | | | |
|---|---|---|---|---|---|---|---|---|
| | **Hu invariant moments** | **DFT** | **DCT** | **Wavelet** | **Hu invariant moments** | **DFT** | **DCT** | **wavelet** |
| **Accuracy** | 0.81 | 0.85 | 0.75 | 0.96 | 0.77 | 0.70 | 0.77 | 100 |
| **Sensitivity Hand class** | 0.76 | 0.86 | 0.80 | 0.94 | 1 | 0.83 | 1 | 100 |
| **Sensitivity Face class** | 0.90 | 0.83 | 0.67 | 1 | 0.73 | 0.73 | 0.73 | 100 |
| **Specificity Hand class** | 0.90 | 0.83 | 0.67 | 1 | 0.5 | 0.50 | 0.5 | 100 |
| **Specificity Face class** | 0.76 | 0.86 | 0.80 | 0.94 | 1 | 0.60 | 1 | 100 |
| $FN_{hand}$ | 4 | 1 | 1 | 1 | 0 | 2 | 0 | 0 |
| $FN_{face}$ | 1 | 1 | 1 | 0 | 3 | 4 | 3 | 0 |
| $FP_{hand}$ | 1 | 1 | 1 | 0 | 3 | 4 | 3 | 0 |
| $FP_{face}$ | 4 | 1 | 1 | 1 | 0 | 2 | 0 | 0 |
| $TN_{hand}$ | 9 | 5 | 2 | 12 | 3 | 4 | 3 | 15 |
| $TN_{face}$ | 13 | 6 | 4 | 15 | 2 | 3 | 2 | 30 |
| $TP_{hand}$ | 13 | 6 | 4 | 15 | 7 | 10 | 7 | 30 |
| $TP_{face}$ | 9 | 5 | 2 | 12 | 8 | 11 | 8 | 15 |
| **Number of Features** | 7 | 20 | 20 | 4 | 7 | 20 | 20 | 4 |
| **Number of neurons** | 7-40-40-2 | 20-40-40-2 | 20-40-40-2 | 4-40-40-2 | 7-20-2 | 20-20-2 | 20-20-2 | 4-20-2 |
| **Number of epoch** | 16000 | 2000 | 2048 | 6000 | 10000 (hand) 22000 (face) | 7400 (hand) 24000 (face) | 29000 (hand) 37000 (face) | 9000 (hand and face) |
| $E_{av}$ | 0.1938 | $6\times10^{-6}$ | $2\times10^{-12}$ | $2\times10^{-4}$ | 0.2 (hand and face) | 0.2483(hand) 0.2262 (face) | 0.2 (hand and face) | 0.044 (hand and face) |

Table.4
Result of MLP with the features of dual, trine and all together modes.

| | MLP | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Wavelet& DFT** | **Wavelet& DCT** | **Hu & DFT** | **Hu & DCT** | **DFT & DCT** | **Hu & DCT &Wavelet** | **Hu & DFT &Wavelet** | **Hu & DCT&DFT &Wavelet** |
| **Accuracy %** | 99 | 95 | 85 | 74 | 74 | 96 | 96 | 86 |
| **Sensitivity Hand class%** | 100 | 100 | 86 | 87 | 75 | 100 | 100 | 92 |
| **Sensitivity Face class%** | 100 | 92 | 83 | 67 | 73 | 93 | 94 | 80 |
| **Specificity Hand class%** | 100 | 92 | 83 | 67 | 73 | 93 | 94 | 80 |
| **Specificity Face class%** | 100 | 100 | 86 | 87 | 75 | 1 | 1 | 92 |
| $FN_{hand}$ | 0 | 0 | 1 | 1 | 3 | 0 | 1 | 1 |
| $FN_{face}$ | 0 | 1 | 1 | 5 | 3 | 1 | 0 | 3 |
| $FP_{hand}$ | 0 | 1 | 1 | 5 | 3 | 1 | 0 | 3 |
| $FP_{face}$ | 0 | 0 | 1 | 1 | 3 | 0 | 1 | 1 |
| $TN_{hand}$ | 13 | 13 | 5 | 10 | 8 | 13 | 12 | 12 |
| $TN_{face}$ | 15 | 14 | 6 | 7 | 9 | 14 | 15 | 12 |
| $TP_{hand}$ | 15 | 14 | 6 | 7 | 9 | 14 | 15 | 12 |
| $TP_{face}$ | 13 | 13 | 5 | 10 | 8 | 13 | 12 | 12 |
| **Number of Features** | 24 | 24 | 27 | 27 | 40 | 31 | 31 | 51 |
| **Number of neurons** | 24-40-40-2 | 24-40-40-2 | 27-40-40-2 | 27-40-40-2 | 40-40-40-2 | 31-40-40-2 | 31-40-40-2 | 51-40-40-2 |
| **Number of epoch** | 560 | 3600 | 3000 | 1272 | 1485 | 2600 | 1800 | 650 |
| $E_{av}$ | $9\times10^{-5}$ | $2\times10^{-5}$ | $4\times10^{-6}$ | $7\times10^{-6}$ | $1\times10^{-12}$ | $2\times10^{-6}$ | $6\times10^{-6}$ | $19\times10^{-6}$ |


Fig. 5. The result of 96% accuracy MLP network on video sequence for detecting hand gestures with related to Hu & DFT &Wavelet in table 4. (A) Tracking one and two hand gestures in video frames with further light and simple background (successful case). (B) Tracking one hand with rotation and less illumination and clutter background (successful case). (C) Tracking two hands with false positive due to color and shape like hand (unsuccessful case).

## References

[1] C.K. Yang, Y.Ch. Chen, "A HCI interface based on hand gestures", Springer, Signal, Image and Video Processing, SIViP, March 2013.

[2] S. S. Rautaray, A. Agrawal, "Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey", Springer, Artificial Intelligence Review, November 2012.

[3] M. Egmont-Petersena, D. de Ridderb, H. Handelsc, "Image Processing with Neural Networks—a Review", Elsevier on Pattern Recognition, pp. 2279–2301, August 2001.

[4] M.Kumar, N. Yada, "Multilayer Perceptrons and Radial basis Function Neural Network Methods for the Solution of Differential Equations: A Survey", Elsevier, Computers and Mathematics with Applications, Vol.62, pp.3796-3811, 2011.

[5] V. Skorpil, J. Stastny, "Back-Propagation and K-Means Algorithms Comparison", IEEE, 8th International Conference on Signal Processing, 2006.

[6] H. Hikawa , S. Matsubara,"Pseudo RBF Network for Position Independent Hand Posture Recognition System", IEEE Proceedings of International Joint Conference on Neural Networks, Florida, USA,pp.1049 – 1054, August 2007.

[7] M. Qu, F. Y. Shih, J. Jing, H.Wang, "Automatic Solar Flare Detection Using MLP, RBF, AND SVM", Solar Physics, Vol.217, pp.157–172, 2003.

[8] H. Tonekabonipour, A.Emam, M. Teshnelab, M. A. Shoorehdeli, "Ischemia Prediction via ECG using MLP and RBF Predictors with ANFIS Classifiers" , IEEE, Seventh International Conference on Natural Computation,pp.776-780, 2011.

[9] J. Misra, I. Saha, "Artificial Neural Networks in Hardware: A survey of Two Decades of Progress", Elsevier, Neurocomputing, Vol.74, pp.239-255, 2010.

[10] S.Jung, S. Su. Kim, "Hardware Implementation of a Real-Time Neural Network Controller with a DSP and an FPGA for Nonlinear Systems", IEEE Transaction on Industrial Electronics, Vol.54, pp.256-271, February 2007.

[11] H. M. Ebied, K. Revett, M. F. Tolba, "Evaluation of Unsupervised Feature Extraction Neural Networks for Face Recognition", Springer, Neural Computing & Application, Vol.22, pp.1211–1222, 2013.

[12] S. Belciug, F. Gorunescu, M.Gorunescu, A.B. M. Salem, "Assessing Performances of Unsupervised and Supervised Neural Networks in Breast Cancer Detection", IEEE, 7th International Conference on Informatics and Systems (INFOS), pp.1-8, 2010.

[13] E. Stergiopoulou, N. Papamarkos, "Hand Gesture Recognition Using a Neural network Shape Fitting Technique", Elsevier, Engineering Applications of Artificial Intelligence 22, pp.1141–1158, March 2009.

[14] Thiago R. Trigo and Sergio Roberto M. Pellegrino, "An Analysis of Features for Hand- Gesture Classification", IWSSIP, 17th International Conference on Systems, Signals and Image rocessing, pp.412-415, 2010.

[15] M. K. Kiran, T. ShyamVamsi, "Hand Gesture Detection and Recognition Using Affine-Shift, Bag-of-Features and Extreme Learning Machine Techniques", Springer, Advances in Intelligent Systems and Computing 247, pp.181-187, 2014.

[16] A. K. Sarkalehl, F. Poorahangaryan, B. Zan, A. Karami, "A Neural Network Based System for Persian Sign Language Recognition", IEEE International Conference on Signal and Image Processing Applications,pp.145-149, 2009.

[17] C.T. Hsieh, C.H. Yeh, K.M. Hung, L.M. Chen, C.Y.Ke, "A Real Time Hand Gesture Recognition System Based on DFT and SVM", IEEE,8th International Conference on Information Science and Digital Content Technology (ICIDT),pp. 490 - 494, 2012 .

[18] P.V.V Kishore, S.R.C Kishore, M.V.D Prasad, "Conglomeration of Hand Shapes and Texture Information for Recognizing Gestures of Indian Sign Language Using Feed forward Neural Networks" , International Journal of Engineering and Technology (IJET), Vol.5, pp.3742-3756, Oct-Nov 2013.

[19] Rui-Feng Bo, "A New Approach to Mechanism Type Selection by Using Back-Propagation Neural Networks", IEEE International Conference on Artificial Intelligence and Computational Intelligence, pp.205-209, 2010.

[20] S.W. Lin, T.Y. Tseng, S.Y. Chou, S.C. Chen , "A simulated-annealing- based approach for simultaneous parameter optimization and feature selection of back-propagation", An International Journal of expert systems with applications, Vol.34, pp.1491-1499, February 2008.

[21] M. Krips, T. Lammert, A. Kummert, "FPGA Implementation of a Neural Network for a Real-Time Hand Tracking System", Proceedings of the First IEEE International Workshop on Electronic Design, Test and Applications, pp.7695-1453, 2002.

[22] L.M.Silva, J.Marques de S, L.A.Alexandre, "Data classification with multilayer perceptrons using a generalized error function", Elsevier, Neural Networks, Vol. 21, pp.1302-1310, 2008.

[23] V. Aeinfar, H. Mazdarani, F. Deregeh, M. Hayati, M. Payandeh, "Multilayer Perceptron Neural Network with Supervised Training Method for Diagnosis and Predicting Blood Disorder and Cancer", IEEE International Symposium on Industrial Electronics (ISIE), 2009, pp.2075-2080.

[24] Emad A. M, A. Shenouda, "A Quantitative Comparison of Different MLP Activation Functions in Classification", Springer, Advances in Neural Networks, China, Vol. 3971, pp. 849–857, 28 May 2006.

[25] M. J. Er, S. Wu, J. Lu, H. L. Toh, "Face Recognition With Radial Basis Function (RBF) Neural Networks", IEEE Transactions on Neural Networks, Vol.13, No.3, pp.697-710, May 2002.

[26] P. Vadakkepat, L. C. De Silva, L.Jing, L. L. Ling, "Multimodal Approach to Human-Face Detection and Tracking", IEEE Transaction on Industrial Electronics, Vol.55, No.3, pp.1385-1392, March 2008.

[27] Ooi M. P, "Hardware Implementation for Face Detection on Xilinx Virtex-II FPGA using the Reversible Component Transformation Colour Space", IEEE, Proceedings of the Third International Workshop on Electronic Design, Test and Applications (DELTA'06), Jan 2006.

[28] Christopoulos C., Skodras, A., Ebrahimi, T., "The JPEG 2000 Still Image Coding System: An Overview", IEEE Transaction on Consumer Electronics, Vol.46, No.4, pp.1103-1127, Nov 2000.

[29] H. Qian, Y. Mao, W. Xiang, Z. Wang, "Recognition of Human Activities using SVM Multi-class Classifier", Elsevier on Pattern Recognition Letters, pp.100-111, 2009.

[30] P. Premaratne, S.Ajaz, M. Premaratne, "Hand gesture tracking and recognition system using Lucas–Kanade algorithms for control of consumer electronics", Elsevier, journal on Neurocomputing, pp.1-8, 2012.

[31] J. Li, L. Zheng, Y. Chen, Y. Zhang, P. Lu , "A Real Time Hand Gesture Recognition System Based on the Prior Facial Knowledge and SVM", Journal of Convergence Information Technology (JCIT), Number11, Vol.8,pp.185-193, June 2013.

[32] Liu Yun, Zhang Lifeng, Zhang Shujun , "A Hand Gesture Recognition Method Based on Multi-Feature Fusion and Template Matching", Elsevier, Procedia Engineering 29, pp.1678- 1684,2012.

[33] Bui T.T.T., Phan N.H. and Spitsyn V.G, "Face and Hand Gesture Recognition Algorithm Based on Wavelet transforms and Principal Component Analysis", IEEE, 7th International Forum on Strategic Technology (IFOST), pp.1-4, 2012.

[34] B. Thi, T. Trang, V.G. Spitsyn, "Digital Image Dissection by Two- Dimensional Discrete Wavelet Transform and Fast Haar Wavelet Transform", Bulletin of the Tomsk Polytechnic University, Vol.318, No. 5, pp.73–76, 2011.

[35] Bui Thi Thu Trang, Phan Ngoc Hoang and V.G. Spitsyn, "Software and Algorithmic Support for Digital Image Classification by the Haar Wavelet Transform and Neural Networks", Bulletin of the Tomsk Polytechnic University, Vol. 319, No. 5, pp.103–106, 2011.

[36] A. Zabidi, L. Y. Khuan, W. Mansor, I. M. Yassin, R. Sahak, "Classification of Infant Cries with Asphyxia Using Multilayer Perceptron Neural Network", IEEE Second International Conference on Computer Engineering and Applications, pp.204-208, 2010.

[37] M. F. Rozali, I. M. Yassin, A. Zabidi, W. Mansor, N. M. Tahir, "Application of Orthogonal Least Square (OLS) for Selection of Mel Frequency Cepstrum Coefficients for Classification of Spoken Letters using MLP Classifier", IEEE 7thInternational Colloquium on Signal Processing and its Applications, pp.464-468, 2011.

[38] C. Chan, S. S. Mirfakhrae, " Hand Gesture Recognition using Kinect", Bchelor Thesis, Boston University Department of Electrical and Computer Engineering ,Boston, Dec 13, 2013.

[39] A.Kulshreshth, C. Zorn, J. J. LaViola, "Poster: Real-time Markerless Kinect based Finger Tracking and Hand Gesture Recognition for HCI", IEEE Symposium on 3D User Interfaces USA, pp.187-188, March 2013.

[40] Y. Ye, S.Cheng, "Implementation of 2D-DCT Based on FPGA with Verilog HDL", Springer, Electronics and Signal Processing, LNEE 97, pp.633-639, 2011.

[41] A. B. Atitallah, P. Kadionik, F. Ghozzi, P.Nouel, N. Masmoudi. , "Optimiztion and Implementation on FPGA of the DCT/IDCT Algorithm", IEEE Proceedings International Acoustics, Speech and Signal Processing ICASSP, pp.928-931, 2006.