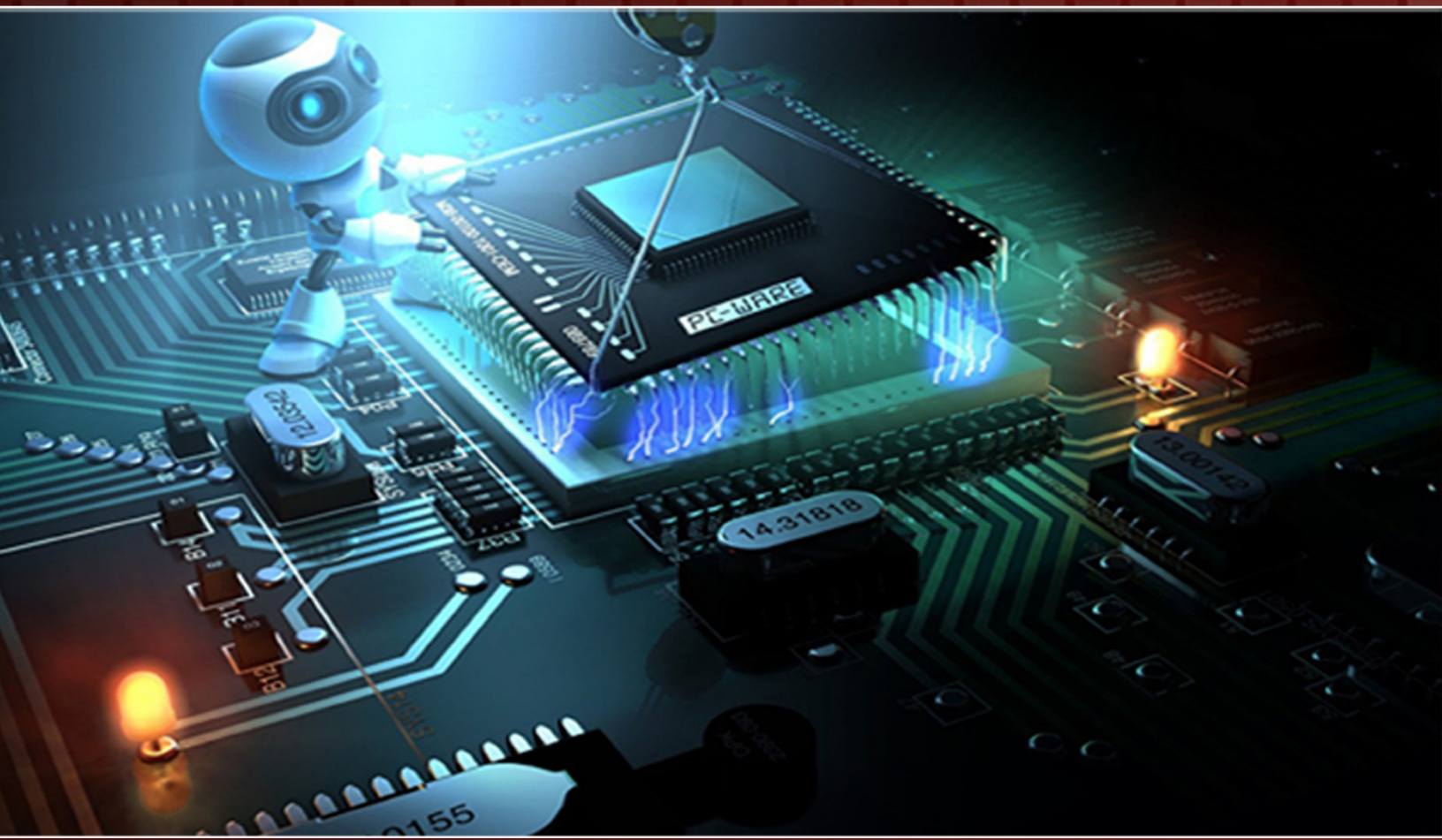
ISSN: 2054-7390

Volume 3 Issue 6, December 2015

Transactions on Machine Learning and Artificial Intelligence



Society for Science and Education
UNITED KINGDOM

TABLE OF CONTENTS

EDITORIAL ADVISORY BOARD

DISCLAIMER

All the contributions are published in good faith and intentions to promote and encourage research activities around the globe. The contributions are property of their respective authors/owners and the journal is not responsible for any content that hurts someone's views or feelings etc.

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

Absorption Spectra Analysis using Modified Self-Organizing Feature Maps

[1]Thomas Bryant, [2]Jessica Koppen and [3]Mohamed Zohdy
[1]Department of Computer Science and Engineering, Oakland University, United States;
[2]Department of Chemistry, Oakland University, United States;
[3]Department of Electrical and Computer Engineering, Oakland University, United States
tcbryant@oakland.edu; jvkoppen@oakland.edu; zohdyma@oakland.edu

ABSTRACT

This research demonstrates an application of a modified self-organizing feature map (SOFM) algorithm to analyze and discover the quality of chemical absorption spectrum data. By forming an *NxN* neural array from input features and varying the essential parameters of the algorithm, map recognition quality is increased at the expense of more computation. The features of this SOFM are based on absorption intensity variations with excitation wavelength. SOFMs are used to discern pattern similarities and differences between spectral data. A context tree allows individual features, or key numbers in the data, to be input and classifies the vector to the type of data that is most similar. This research also use the self-organizing map to enhance as well as visualize resultant classification efficiency through the use of watershed transformation.

Keywords: Self-Organizing Feature Map, Unsupervised Learning, Energies, Intensities, Chemical Absorption

1 Introduction

UV/Vis spectroscopy is a routinely used tool to probe the electronic energy levels of atoms and molecules. Changing the levels by an electron, if symmetry allowed, is associated with emission or absorption of a photon of W/Vis light.

The interest in the interactions between gold nanoparticles and light dates back to Michael Faraday's observation of ruby-red colored colloidal solutions of gold. The property that gives the gold nanoparticles their ability to absorb visible light is known as surface plasmon, i.e., the collective oscillation of conduction electrons. As the size of the nanoparticle decreases and the electrons become confined, the metallic band structure becomes discretized. In other words, the discrete level structure emerges from continuous bands. This process can be explained by a quantum mechanical model of a particle in a box except for the "box" being spherical in shape. At this point a nanoparticle becomes a molecule with defined energy level structure. This species is referred to as a nanocluster (or sometimes a quantum dot).

Energy levels of gold nanoclusters have been probed by UV spectroscopy, which is uniquely suited to discern the electronic level structure [1]. UV spectroscopy shows that the plasmons disappear and the

DOI: 10.14738/tmlai.36.1155
Publication Date: 11th November, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1155

UV spectrum displays distinct, molecular-like transitions. The first study to demonstrate these distinct molecule-like transitions involved photodissociation spectroscopy of Aun+-Xe clusters by Collings et al. [2]. A recent study by Lecoultre et al. recorded the high resolution UV spectra of small Aun nanoclusters (n=1-5, 7-9) in low-temperature Ne matrices [3]. These spectra reveal a number of sharp electronic transitions which vary with the cluster size. The theoretical calculations of energy levels that give rise to these electronic transitions are necessary for a number of reasons including:

(i) assignments of excited states
(ii) determination of optical gaps
(iii) finding positions of the lowest excited state, etc.

Theoretical predictions of excited states rely on wave function methods among which the most trustworthy is Equation-of-Motion Coupled Cluster with single and double excitations (EOM-CCSD) [4]. This method is reliable but scales prohibitively with the number of electrons (as N^7 where N is the number of electrons) and therefore cannot be used for large molecules. For this reason methods based on linear response time-dependent density functional theory (TD-DFT) which scale as N^4-N^5 are gaining in popularity [4]. However, these methods show a strong dependence on the density functional used, that is often hard to rationalize, in that that seemingly "better" functionals may results in worse predictions of excited states. In gold clusters one is faced with additional obstacles in excited states calculations: large number of electrons and the fact that valence d electrons form large number of molecular orbitals with nearly identical energies (practically a d-band).

In assessing the performance of functionals used in TD-DFT one typically chooses a priori fixed number of transitions to excited states and then performs a statistical analysis on the basis of comparison with either more accurate method such as EOM-CCSD or experiments. Each excitation is characterized by the excitation energy (in eV) and oscillator strength f which is related to the intensity of the transition. A transition may be allowed or forbidden by symmetry. In addition, some symmetry-allowed transitions may show zero f.

In gold, the approach based on assignment and comparison is impossible to follow for all functionals because TD-DFT leads, at least for some functionals, to too many transitions with nonzero oscillator intensity. This effect may be compared to noise. It is not clear which element of density functional theory causes this failure. It is crucial at this point to identify the functionals that have this problem and the ones that do not.

In this paper, we use the UV spectrum of the cluster Au to assess the performance of 9 popularly used density functionals in the prediction of this spectrum. The corresponding EOM-CCSD results for transition energies and associated oscillator strengths, f, serve as the reference spectrum. TD-DFT calculations were performed for 200 excited states, each for the following list of functionals:

Table 1. The list of functional, each representing its own data set

B3PLYP
CAM-B3LYP
LC-LDA
LC-WPBE
LDA
PBE
TPSS
WB97X

These functional are listed in alphabetical order. They were obtained through computational chemistry study.

Each data set also includes excitation energies, which is on the x-axis and oscillator strengths (intensities) which is on the y-axis.

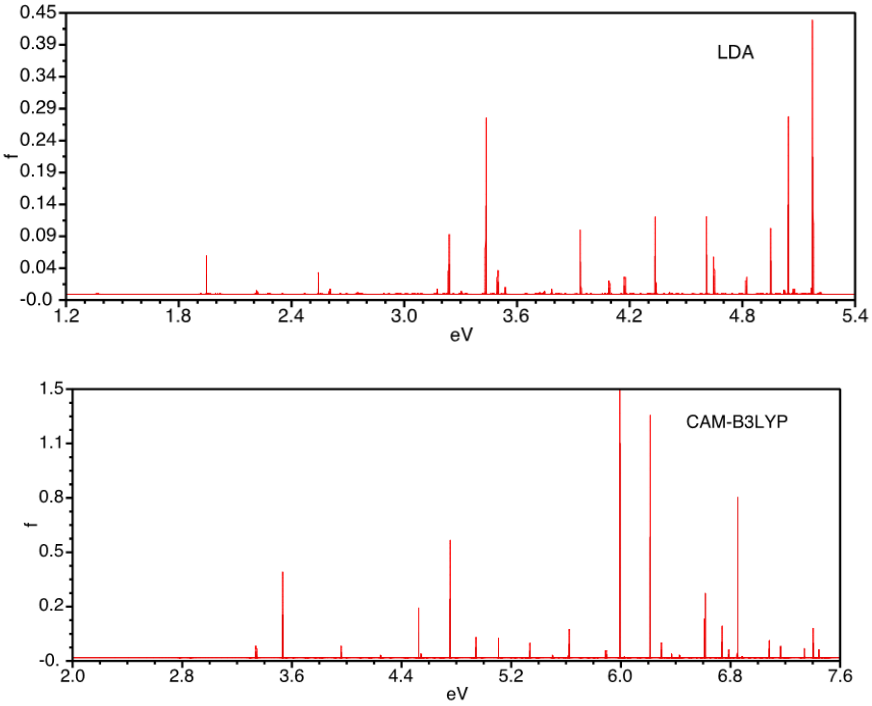


Figure 1. Calculated TD-DFT spectra of Au8 with two different functional (a) LDA; (b) CAM-B3LYP. Note different scales for transition energies and intensities.

The difficulties with assignment and statistical analysis spectra resulting from different functionals is illustrated in Fig. 1 (a) and (b) showing the LDA and CAM-B3LYP spectra, respectively. The spectra differ in three major ways:

(i) number of active transitions (i.e. with nonzero f),
(ii) intensities with CAM-B3LYP resulting with about three-fold higher intensities,
(iii) energy range of the first 200 transitions: in LDA this range is (1.37 eV – 5.2 eV) whereas is CAM-B3LYP the range is (2.79 eV - 7.52 eV). Clearly, the task of evaluating these spectra requires novel methods.

The method presented in this paper can qualitatively distinguish between bad and good spectrum. It is based on the following ideas. A typical result of the application of this method is presented in Fig. 2 below.

By extracting suitable features from the spectra, the unsupervised learning neural network, or modified self-organizing feature map (SOFM) is able to produce the discriminant clustering that occurs and also visualize it. Cluster analysis in terms of the SOFM is the assignment of observations into subsets, or clusters, so that observations within the same densely populated region are compact and separate. Signals can be mathematically represented in time domain, frequency domain, and better still joint time-frequency domain. The main focus of this study will deal with the frequency domain. The key

features extracted from the frequency domain are the resonances observed, which are obtained by analyzing the peaks and frequencies of the spectra.

In the last step of the feature extraction process, features are entered as “input”. The preprocessed data is divided into quadrants and then used to obtain the salient features. Once the “input” is given, the top node of a context feature tree changes to the color of the spectrum it most closely resembles. This comparison is based solely on the information obtained from the inputs. The matrix of the magnitudes from the joint time-frequency diagram was relatively small here; the largest matrix was 200 rows by 2 columns; all of the data had similar sizes except for the training data which was 28 rows by 2 columns. As a result, the entire matrix could not be used efficiently when comparing against new data. Thus, an existing algorithm was modified to select features for the self-organizing feature map (SOFM).

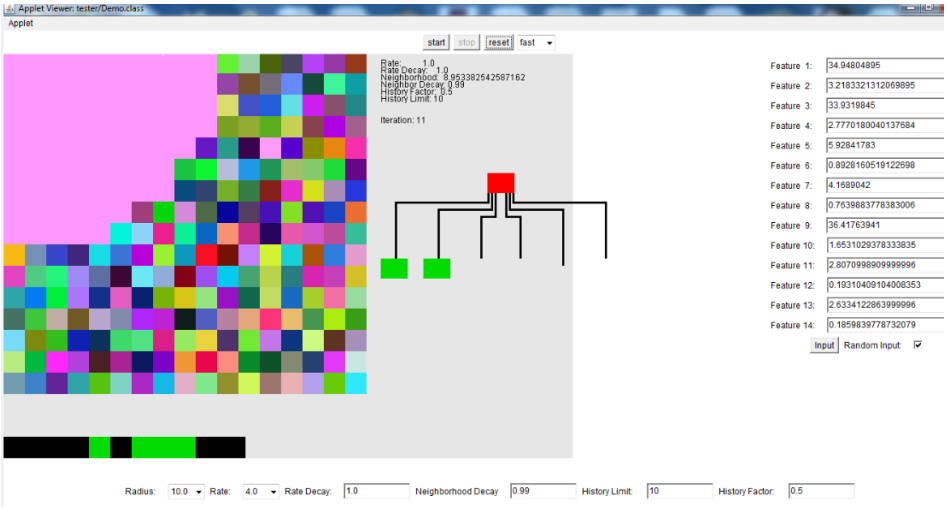


Figure 2: The self-organizing map being trained to recognize clusters

The image above shows the self-organizing map, the context tree, and the features that represent input. Preprocessed data is divided into quadrants and then used to obtain the salient features; the optimal amount is fourteen. Distinct features are obtained through the frequency domain, and then the program learns the patterns hidden in the data through clustering the SOFM [5]. To classify the spectra, the winning node is calculated using the Euclidean distance formula to determine which node is closest. The output is visualized and displayed on a two-dimensional map, and the vector is colored according to the spectra that has the closest node. It visualizes spectra as similar colors that act as competing agents. The “winners” are the vectors that weigh the least [6].

The map was initialized with random weights once the features were obtained. Nodes were created to form map anchors with these pseudo-random entities. The algorithm (mentioned later) reads the data and compares it to the random numbers. This determines which pseudo-random number is closest to using the earlier-mentioned Euclidean distance formula. This stage of updating the weights to ideally adapt the network’s behavior is known as the training phase [6].

Before the SOFM algorithm begins to cluster the nodes, it accepts ten parameters: the neighborhood radius, learning rate, rate decay, neighborhood decay, history limit, and history factor. The default values are best depending on the type of data, but the user can specify any or all of the values to obtain

more effective clustering. The colors of the previous several iterations are displayed underneath the map, therefore increasing the accuracy of the clustering [5].

2 Feature Extraction Algorithm

To obtain features from frequency domain, spectrum data was examined. Then, all peaks and resonance frequencies are analyzed to determine key features. W_i, or the frequency, is obtained from the x-axis and M_i, or intensity (or peak), is obtained from the y-axis.

These features are input by the user through fourteen text boxes that are used to manipulate the data. Once the "input" is given, the top node of a context tree changes to the color of the most closely resembled data. This comparison is based on the information obtained from the input file; it does not affect the SOFM.

3 The Learning Algorithm

3.1 Training Stage

The primary goal of SOFM is to transform input patterns, of arbitrary dimensionality, into the responses of one or two-dimensional output arrays of neurons. These three steps are required [6]:

(i) an array of neurons that compute simple functions of the incoming inputs;
(ii) a mechanism for selecting the neuron (or neurons) with the largest produced output called winners;
(iii) an adaptive mechanism that updates the weights of the selected neuron and its neighbors.

This algorithm is used to train SOFM by Kohonen [5] and summarized as follows:

Initially, the weight vectors of the map's neurons are randomized. The SOFM takes an input vector and then traverses the map by calculating the Euclidean distance:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_1 - p_1)^2} \quad (1)$$

where p, q are two vectors being compared, other distances such as p-norm may also be used.

The similarities between the input vectors and neuron weight vector correspond to the distance between them. The algorithm then tracks the smallest distance (this node is called a winner, best matching unit, or BMU). The neighborhood function has the BMU as center node along with its neighborhood making them even more similar to presented input vector by using the equation below, as referenced from a previous paper [6]:

$$W_i(t+1) = W_i(t) + \alpha(t) + \beta(t)[k(t) - W_i(t)] \quad (2)$$

Repeat t, which is a positive constant and iteration and repeat from 2 while $t < \lambda$:

t denotes the current iteration

λ is the limit on iterations of time
W_v is the current weight vector of neuron v
dist(t) is the target input data vector
$\alpha(t)$ is learning rate due to time
$\beta(t)$ is the neighborhood function depending on distance from BMU

The above parameters determine the performance efficiency of the SOFM algorithm. The set of input vectors will each have three components, corresponding to a color space. The target output will correspond to a color space and settle on the RGB output map. Weight vectors convergence to the target in finite number of iterations. The current iteration must always be less than the limit on time iteration. A history limit is also embedded, which shows vectors that have been input previously in color format, and a history factor to update previous winning nodes has been implemented to decrease the time necessary to complete the training phase.

The neurons in the SOFM are initialized with small random weights. These initial weights determine the learning speed of the map. It will take more iterations depending on how large the learning rate is. However, the colors will not group together seamlessly on the visual display if the weights are too small. The next step included clustering and classifying the "ideal", or training data using the learning algorithm. This data trained a 17x17 neural network.

4 Simulation Results

4.1 Spectra Selection

The data was extracted into a feature vector of input length fourteen and shown on a context tree. After testing of the parameters, they were fine-tuned using sequential input, a radius of 10.0, a learning rate of 4.0, a rate decay of 1.0, a neighborhood decay of 0.99, a history limit of 10.0, and a history factor of 0.5. The neighborhood value has a two parameters of value: size and a distance function to influence.

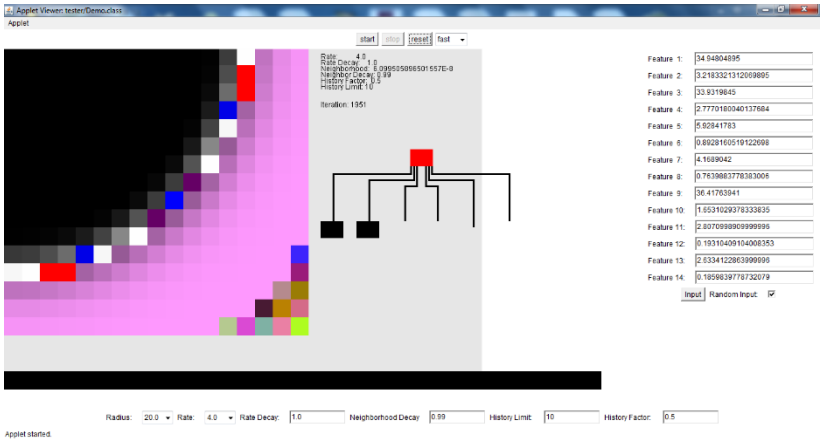


Figure. 3: Training data in comparison to the best absorption spectrum

The self-organizing map above shows training data, or ideal data, compared to the best spectrum, wB97X. The ideal data and spectrum data both have clearly defined clusters.

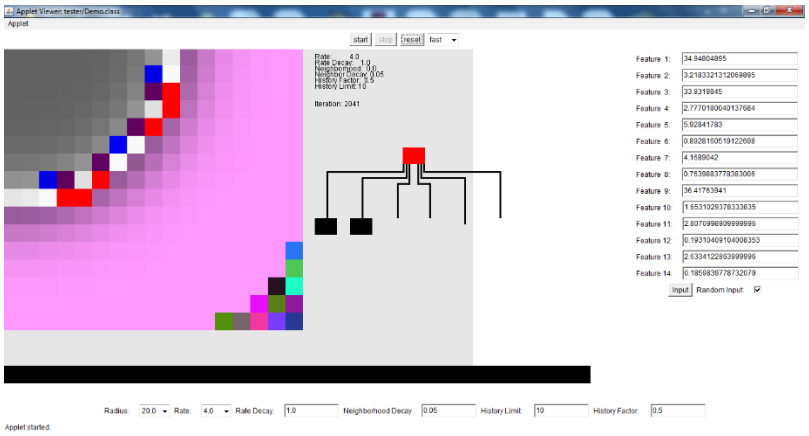


Figure. 4: Absorption spectra B2PLYP and wB97X used in correlation

The above figure shows the best pair of spectrum data, B2PLYP and wB97X. These two spectra have been entered into the self-organizing map in conjunction. The clear purple and silver clustering indicates that both clusters were clearly defined here as well.

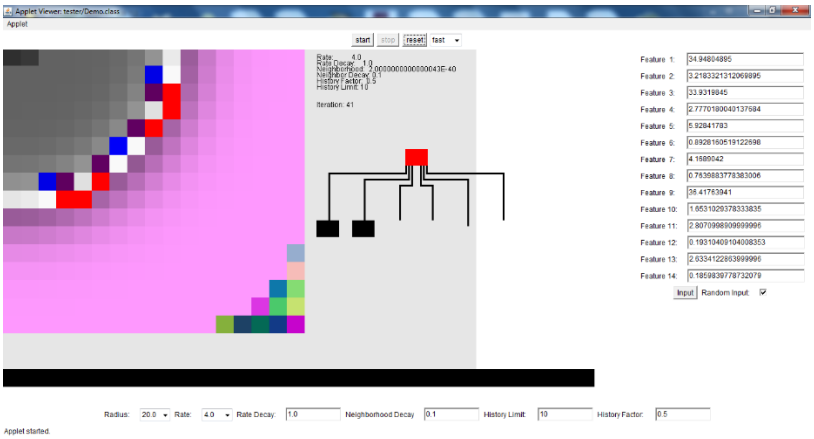


Figure 5. Absorption spectra B2PLYP and wB97X used to compare both energies and intensities

The self-organizing map in Figure 5 above shows the best pair of spectra grouped together with both energies and intensities. Both of the spectra clustered in a very high quality once again, with a few colors (blue, white, dark purple, red) serving as transition colors from purple to silver.

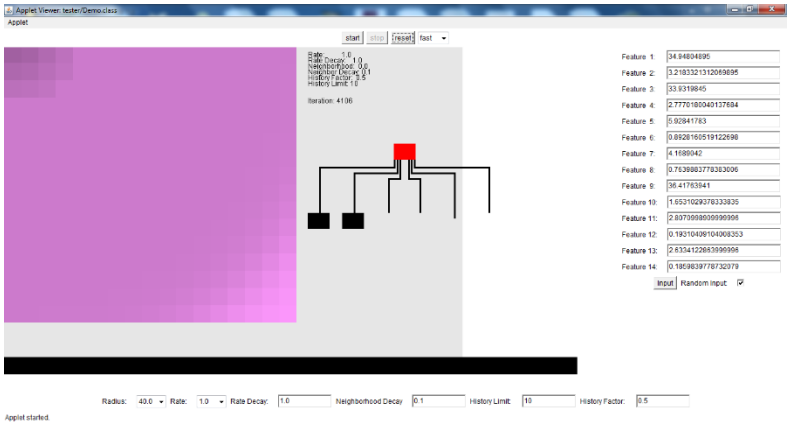


Figure. 6: SOFM correlation of spectra wB97X and LDA

The above image showed the best spectra (wB97X), grouped with the worst (LDA), using both energies and intensities. The large amount of purple indicates that the clusters conglomerated together or that one spectra completely "won" over the other.

4.2 Range of Absorption Spectra

Table 2. List of functionals listed from the ones containing the best to worst spectra

wB97X
B2PLYP
CAMB-3LYP
LCLDA
LCWPBE
B3PLYP
TPSS
PBEPBE
LDA

This table above lists the absorption spectra from the best spectrum to the worst spectrum. They were also determined to have the highest to lowest numbers when the fourteen features were averaged.

Of the spectra that were used in the relevant chemistry testing, wB97X was revealed to be the best, while TPSS was revealed to be the worst. This coincides with the results the self-organizing map found since wB97X had the lowest average of features by number, and while PBE had the highest, it was not used in the chemistry testing, and TPSS, which was used, was found to be the second worst (it had the second highest median of the eight spectra when they were reduced to fourteen features). In addition, it was found that the better they cluster, the better the spectra was. When comparing wB97X and the training data, they clustered almost seamlessly.

5 Conclusion

The SOFM (self-organizing feature map) has successfully classified UV chemical spectra inputs according to properly selected features and accurately clustered and visualized the data presented. It also provided a context for input features that served as training data with an associated context tree diagram and displayed the color that most resembled the spectra. After testing, the colors of the SOFM clustered best with sequential input and the following parameters: a neighborhood radius of 10.0, a learning rate of 4.0, a rate decay of 1.0, a neighborhood decay of 0.99, a history limit of 10.0, and a history factor of 0.5. SOFM performance was determined by the degree to which the colors clustered together and the number of squares of random color were left on the SOFM.

The history function that serves as one of the algorithm parameters improves the performance during the training phase substantially. Correlation usage increases the success rate of identifying unknown input by 15-20%. Blended distance extending Euclidean not only has a significant effect in improvement of computational time in our self-organizing feature maps, but it can also be used in other applications such as recognizing speech patterns, genetics, and music analysis. We finally added watershed transform as a good preprocessor to enhance the fidelity of visualizing SOFM concepts.

REFERENCES

[1] T. G. Schaaff, R. L. Whetten, "Giant gold-glutathione cluster compounds: Intense optical activity in metal-based transitions," J. Phys. Chem. B, 104, 2630-2641 (2000).

[2] A. Collings, K. Athanassenas, D. Lacombe, D. M. Rayner, and P. A. Hackett, "Optical absorption spectra of Au_7, Au_9, Au_{11}, and Au_{13}, and their cations: Gold clusters with 6, 7, 8, 9, 10, 11, 12, and 13 *s*-electrons," J. Chem. Phys. 101, 3506 (1994).

[3] S. Lecoultre, A. Rydlo, C. Felix, J. Buttet, S. Gilb, and W. Harbich, "UV-visible absorption of small gold clusters in neon: Au-n (n=1-5 and 7-9), " J. Chem. Phys. 134, 074302 (2011).

[4] A. I. Krylov, Annu. "Equation-of-motion coupled-cluster methods for open-shell and electronically excited species: The Hitchhiker's guide to Fock space ," Rev. Phys. Chem. 59, 433 (2008).

[5] T. Kohonen, *Self-Organizing Maps.* Springer, 2001, BerlinWhetten RL.

[6] J. Vesanto and Alhoniemi, E. "Clustering of the self-Organizing Map." IEEE Transactions on Neural Networks, vol. 11, No. 3, May 2000. Neural Networks Res. Centre, Helsinki Univ. of Technol., Esposo, Finland.

[7] Bradley, Matthew, Kay Jantharasorn, Keith Jones, and Dr. M. Zohdy. "Self Organized Neural Networks Applied to Animal Communications." REU Report, 2008.

[8] Su, Mu-Chun and Chung, Hsiao-Te. "Fast Self-Organizing Feature Map Algorithm." IEEE Transactions on Neural Networks. Vol. 11, no. 3, May 2000, pp. 721-733.

[9] T. Bryant, M. Hodges, and M. Zohdy. "Modified Self-Organizing Maps for Engine Health Diagnostics", International Journal of Computing and Information Technology. Vol. 3, no. 2, March 2014.

[10] T. Bryant and M. Zohdy, "Noise Signal identification by Modified Self-Organizing Maps", International Journal of Computing and Information Technology. vol. 2, no. 6, November 2013.

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

Defect Detection in Fabric using Wavelet Transform and Genetic Algorithm

Depavath Harinath[1], K. Ramesh Babu[2] ,P.Satyanarayana[3], M.V. Ramana Murthy[3]
[1]*Dept. of Computer Science, HRD Degree and P.G College Narayanaguda, Hyderabad, Telangana, India.*
[2]*Dept. of Mathematics, M.V.S.R Engineering College, Nadergul, R.R.District, Hyderabad, Telangana, India.*
[3]*Dept. of Computer Science, Osmania University, Hyderabad,Telangana,India.*
harinath.depavath@gmail.com

ABSTRACT

Fabric defect detection is one of the indispensable units in the manufacturing industry to maintain the quality of the end product. Wavelet transform is well suited for quality inspection application due to its multi-resolution representation and to extract fabric features. In this paper a new scheme is proposed for fabric defect detection in textile industry. For this purpose, all coefficients were extracted from perfect fabric. These coefficients can defect main fabric image & indicate defects of fabric textile by optimal subset of these coefficients. For finding defects a suitable subset of Genetic Algorithm is used in this process. The Shannon entropy is used as evaluation function in Genetic Algorithm By using two separable sets of wavelet coefficients for horizontal and vertical defects, it was seen that we get better results for defect detection. The advantage of this approach is that it improves accuracy of fabric defect detection as well decreases computation time.

Keywords— Wavelet transform, Genetic Algorithm, Fabric Defect Detection.

1 Introduction

In textile industry, quality control is very important as it directly affects the final product of the industry. Hence to improve quality of textile industry, inspection of textile is very important. In ancient period the inspections in textile industry were done manually which took long time for detection & it become highly expensive due to labor costing. But now inspections are done with the help of automatic inspector machineries. Failure in selection of textile directly affects on quality of production in textile industry & manual inspection requires large time, hence it increases time for manufacturing of product. To improve inspection, modern computer vision inspection system came into existence. It improves accuracy of finding defects and also reduces labor power.

There are numerous reports that shows computer vision based inspection has become very effective in detecting defects in textile industry [1]. Automatic inspections are based on computer vision method & they inspect the defects automatically. Defect in textile industry can be detected by various methods. Research areas in detecting defects are textile, surface of wood, tile, aluminum, distance difference, etc [2]. Uniform, random or patterned textures are the various areas for detecting defects in textile [3]. In this proposed method, fabric defect detection is done by distance difference. In this

DOI: 10.14738/tmlai.36.1551
Publication Date: 12th November, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1551

proposed system, wavelet transformer is used as filter & Genetic algorithm as defect detector. It improves accuracy as well reduces computational time.

A fabric defect is any abnormality in the fabric that hinders its acceptability by the consumer. The textile processing does not eliminate variability incurred during different steps in textile manufacturing. As materials flow from one stage of processing to another, components of variability are added and the final product may involve a cumulative variability that is much higher than the variability of the input fibers and thus it cause a defect in the fabric. The main factors that lead to fabric defects are failure of opening and cleaning the machines that completely eliminate contaminants and trash particles, and it may leads to spinning, weaving and knitting related defects. So the fabric inspection has to identify all types of defects with minimum effort.

Several methods have been proposed to address the problem of detecting defects in textile fabrics, including statistical, spectral and model based approaches.

The wavelet transform provides a solid and unified mathematical framework for the analysis and characterization of an image at different scales . It provides both time and frequency information, and can be successfully applied for textile defect detection. Fabric defect detection based on wavelet transform performs better with less computation than the traditional statistical texture analysis approaches in identifying defects.

The flow of proposed methods is explained in three main parts. First part consists of extracting coefficient through wavelet transform & obtaining coefficient of subset. In second part the suitable subsets are selected through genetic algorithm. Finally in third part suitable subsets are applied to other images and then defects are detected through genetic algorithm.

In first part, the proposed system uses wavelet filter for extracting coefficient of subset. This filter puts special information in each quarter of filtered image. The initial image is divided into four parts by two dimensional wavelet transform. The first part is an approximation of initial image and the other parts include components with high frequencies & low frequencies. By using wavelet transform as filter the wavelet coefficients are obtained. Then genetic algorithm is applied to find suitable subset. Then defects are obtained by fabric defect detection.

2 Literature Survey

Wavelet Transform is decomposition of a signal into constituent parts. This decomposition can done in many ways,

1. Fourier series: Harmonic sinusoids; Single integer index.
2. Fourier transform: Non- harmonic sinusoids; single real index.
3. Walsh Decomposition: “harmonic” square waves; single integer index.
4. Kahunen – Loeve decomposition: Eigen function of covariance; single real index.
5. Short – Time FT (STFT): windowed, non-harmonic sinusoids; double index.

Wavelet transform provide waves in frequency & time domain. Duration of decomposition of signal is limited & has limited band-pass components. High frequency components have short duration & have wide band. Whereas, low frequency components have longer duration & have narrow band.

The wavelet transform is used as filter in this proposed system. The wavelet transform uses a family of wavelet functions and its associated scaling function is used for decomposition of original image into different sub-bands, providing both frequency & spatial function.

Wavelet transform analyses the signal by a group of orthogonal functions in the form of Ψm, n(X), that is computed from translation & dilation of the mother wavelet Ψ. These orthogonal functions are shown below,

$$\Psi_{m,n}(x) = 2^{\frac{-m}{2}} \Psi(2^{-m} x^{-n}) \quad (1)$$

m and *n* are real numbers in this equation.

Equation 2 & 3 are used as analysis & synthesis formulas respectively.

$$C_{m,n} = \int_{-\infty}^{+\infty} f(x) \psi_{m,n} n(x)\, dx \quad (2)$$

$$f(x) = \sum_{m,n} C_{m,n} \Psi_{m,n}(x) \quad (3)$$

Coefficients of this transform are obtained recursively. A synthesis of *j* is shown in equation 4.

$$f(x) = \sum_k C_{0,k} \phi_{0,k}(x) = \sum k C j_{+1,} k \Phi j_{-D}\, d_{j+1,k}\, \Psi_{j+1,k}(x) \quad (4)$$

Two dimensional wavelet transform is used as filter.

The two dimensional wavelet transform divides image into four part .Other parts include components with high frequencies of image & edge points' shows vertical, horizontal & diagonal direction. The wavelet transform uses two types of filter, high frequency filter & low frequency filter. When first order of DWT is applied on image, the image is divided into four part is I_{HH}, I_{LL}, I_{LH} & I_{HL}. Where, H, L is the high & low bound filter respectively

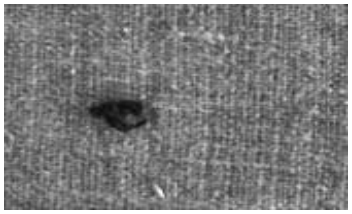


Figure (1): Original image.

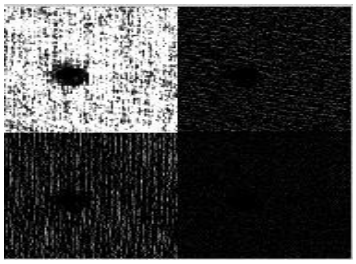


Figure (2): Wavelet Decomposed image after first order

Wavelet transform evaluates the difference in information in two different resolutions & creates a multi-resolution view of image [2]. Hence wavelet transform is used for fabric defect detection. After decomposing image by wavelet filter, the proposed system gets coefficients of subset. For finding suitable subset the proposed system uses the genetic algorithm.

3 The Genetic Algorithm

In a genetic algorithm, a population of candidate solutions to an optimization problem is evolved towards better solutions [4]. The genetic algorithm is iterative process, the evolution starts from a population of randomly generated individual, with the population in iteration is called a generation.

In each generation the fitness function of every individual is calculated. The fitness function is usually the value of objective function which optimizes the better solution to the problem. The fit individuals are selected from the current population & each genome of individual is modified for formation of next generation. In next iteration, new generation of candidate solutions is used in next iteration of the algorithm. The algorithm is repeated until satisfactory fitness level is not reached. The generation of second population is done through the genetic operation such as crossover & mutation.

The simplest form of genetic algorithm involves three types of operators [13]:

1. ***Selection***: This operator selects chromo-somes from population for reproduction. It fitter the chromosome more times so that it is likely to be selected to reproduce.
1. ***Crossover***: The operator in which the subsequences of two chromosomes are exchanged to create new child is called as crossover. This operator roughly mimics biological recombination between two single chromosome organisms.
2. ***Mutation***: This operator randomly flips some bits in a chromosome. Mutations occur at each bit position in a string with some probability.

The Genetic algorithm work as follows [13]

1. Start with a randomly generated population of N L-bit chromosome.
2. Calculate the fitness function F(x) of each chromosome x in the population.
3. Repeat the following steps (a)-(c) until N offspring have been created:

    a) Select a pair of parent chromosome from the current population, with the probability of selection being an increasing function of fitness. Selection is done with replacement," meaning that the same chromosome can be selected more than once to become parents."

    b) With probability Pc (the crossover probability) is used to form two offspring. If no crossover takes place, form two offspring that are exact copies of their respective parents.

    c) Mutate the two offspring at each locus with probability Pm (the mutation probability), and pace the resulting chromosomes in the new population.
4. Replace the current population with the new population.
5. If the stopping condition has not satisfied, go to step 2.

4 Fabric Defect Detection

In second part, the propose method obtains suitable subset through genetic algorithm after obtaining coefficients of subset by using wavelet transform. To obtain defect in fabric the proposed system has to

apply the suitable subset to the images. This suitable subset detects the main image of fabric & indicates the defects in the fabric image. If the optimization problem in Equation 5 can be solved, then suitable subset of wavelet coefficient can be

$$\text{Min } J = C^* P^* P^{T*} C^T$$

Where,

$$C_{j(e)} = \sum_{k=0}^{n-1} C_k C_{k-2J} = 0 \quad (5)$$

In this equation J is the normal form in high frequency components of image, n is the number of coefficients, C is subset of wavelet coefficient as a solution, p is an array of image's pixel values. To minimize J function through c subset, the proposed system uses the genetic algorithm. In this algorithm, each solution c is chromosome & different function can be calculated. As the proposed system uses the wavelet transform as filter, this filter divides image into vertical & horizontal directions. Due this it does not give good result when fabric is not symmetric. So, to obtain better result, two separate filter should be used(horizontal & vertical filter).If W1,W2 is horizontal & vertical filers & u is a perfect image, then filtered image F can be obtained by,

$$F = W1 * U * W_2^T \quad (6)$$

In this proposed system the fitness function for filtered image F is calculated by using sum of squares & entropy functions. Eq. 7 & 8 shows these functions respectively.

$$\text{Sum} = \sum_{i=1}^{n} \sum_{j=1}^{m} F(i,j)^2 \quad (7)$$

Where, n = number of lines, m = column in images. & F (i, j) is the pixel value of the image, in line umi & column j.

$$H = -\sum_{l=1}^{n} d(j) * log_2 (d(j)) \quad (8)$$

Where,

n = number of grey levels, & d(i) is probability of the grey levels in i in the filtered image F. This probability is the real number between 0 & 1. With the help of Eq.7 & 8, optimization problem can be obtained by minimizing J function in Eq. 9 & 10 resp.

$$J = \sum_{l=1}^{n} \sum_{j=1}^{m} (W1 * U * W_2^T)^2 \quad (9)$$

$$J = \text{Entropy}\,(W_1 * U * W_2^T) \quad (10)$$

The wavelet coefficients are obtained on monotony background of entropy function. So, if the entropy function is very small, the image will be more monotonic. Whereas, sum of the square function of image is optimized by minimizing its pixel values. Entropy has simpler computation as compare to sum of squares. So, in this proposed system, entropy function is used for calculation of fitness function of Genetic Algorithm.

5 Expected Result and Conclusion

In this proposed system, the database of fabric is used for finding defect in it. First, the grey image is converted into contrast enhanced image. After conversion the histogram of image is obtained. This

histogram is compared with the histogram of the processed image. After comparison, the wavelet transform is used as filter. In this session, the two dimensional wavelet filter is used. This filter divides the image into four parts & the coefficients of subsets are extracted through wavelet filter.

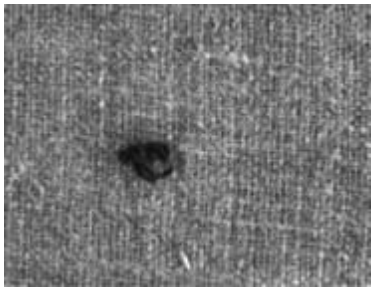


Figure (3): Original grey image

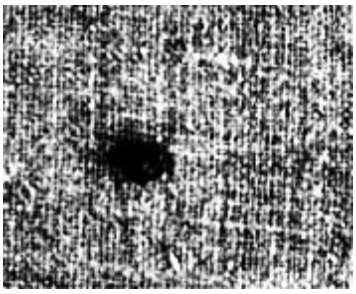


Figure (4): Contrast Enhanced image

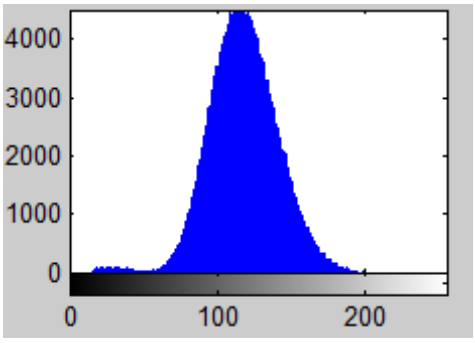


Figure (5): Histogram of original image

After comprising the histogram with processed image, wavelet decomposition is performed.

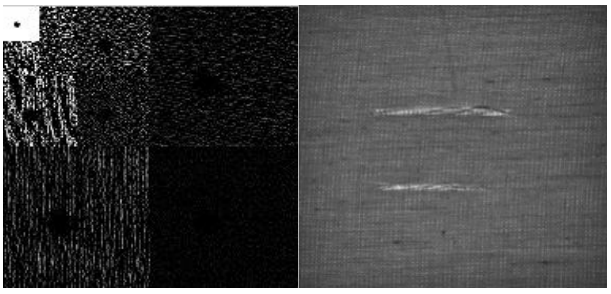


Figure (6): Wavelet Decompose image

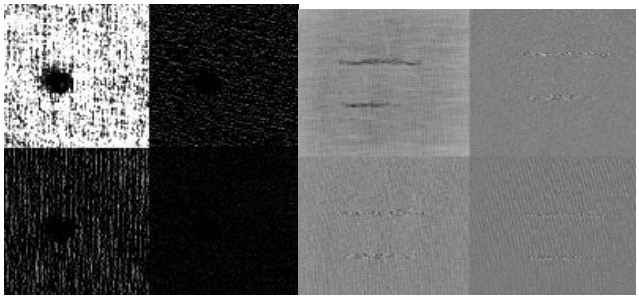


Figure (7): Decomposed image after first stage

After that the proposed system uses genetic algorithm to find suitable subset, then this suitable subset is applied to other images. Original image is deleted & the defects are gained.

REFERENCES

[1] Mahajan P. M., Kolhe S. R and Patil P.M"A review of automated fabric defect detection".

[2] Narges Heidari, Reza Azmi & Boshra Pishgoo "Fabric Textile Defect Detection, By Selecting A Suitable Subset of Wavelet Coefficients, through Genetic Algorithm", International Journal of Image Processing (IJIP), Volume (5): Issue (1): 2011

[3] Lijing Wang, Zhongmin Deng and Xungai Wang "Application of Wavelet Transform Method for Textile Material Feature Extraction".

[4] B.V. Ramana Reddy1, A. Suresh2, M. Radhika Mani3, and V.Vijaya Kumar4 "Classification of Textures Based on Features Extracted from Preprocessing Images on Random Windows", International Journal of Advanced Science and Technology Volume 9, August, 2009.

[5] www.wikipedia.com

[6] Gonzalez R C & Woods R E, Digital Image Processing, 1982,194.

[7] Image Processing Tool Box for use with Matlab (The Math Works Inc. USA), 2005.

[8] Duda,"Pattern Classification."Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, 2009, pp-18-29.

[9] Mahajan P.M.1, Kolhe S.R.2 and Patil P.M.3"A review of automatic fabric defect detection techniques" Bioinfo Publications, Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, 2009.

[10] Xianghua Xie"A Review of Recent Advances in Surface Defect Detection using Texture analysis Techniques" Electronic Letters on Computer Vision and Image Analysis 7(3):1-22, 2008

[11] K.Srinivasan and P.H. Dastor and P. Radhakrishnaihan, and S. Jayaraman "FDAS: A Knowledge-based Frame Detection Work for Analysis of Defects in Woven TextilStructures", Journal of Textile Institute, vol. 83, pp. 431-447, (1992).

[12] R. Chin, "Automated Visual Inspection Techniques and Applications: A Bibliography", Pattern Recognition, 15(4): pp. 343–357, (1982).

[13] Z. Guanng and W. Jianxia, "Fabric Defect Detection Method Based on Image Distance Difference", Electronic Measurement and Instruments, (2007), pp. 822 -825.

[14] D. Chetverikov and A. Hanbury, "Finding Defects in Texture Using Regularity and Local Orientation." Pattern Recognition. 35(10), pp. 2165–2180, (2002).

[15] X.Z. Yang and G. Pang and N. Yung, "Discriminative Fabric Defect Detection Using Adaptive Wavelets", Optical Engineering, pp. 3116-3126, December (2002).

[16] X.Z. Yang and G. Pang and N. Yung, "Robust Fabric Defect Detection and Classification Using Multiple Adaptive Wavelets", Image and Signal Processing, IEEE Proceedings, volume 152, PP. 712-723, (2005)

[17] D. Chetverikov " Measuring the Degree of Texture Regularity", In Proceedings of International Conference on Pattern Recognition, 1984, volume 1, PP. 80–82, (1984).

[18] Yu.zhang, zhaoyang Lu, Jing Li,"Fabric Defect Detection & classification using Gabor filters & Gaussian Mixture Model", Springer-LNCS, pp: 635-644,(2010).

[19] HAN leil, LI zhong,"quick defect detection based on structure character of woven fabric image & wavelet transform", computer engineering & design",(2009)

[20] Rong Fu, Meihong Shi, Hongli Wei, Huijuan chen,"fabric defect detection based on adaptive local binary patterns", international conference on robotics & bi omimetics, pp: 1336-1340,(2009)

[21] S.Mallat, "A Wavelet Tour of Signal Processing", Academic Press, 2nd ed., San Diego,(1999).

[22] S.Mallat,"A Theory for Multiresolution Signal Decomposition: the Wavelet Representation", IEEE Transactions on Pattern Anal. and Mach. Intell., vol. 11, no. 7, pp. 674-693, (1989)

[23] Z. Michalewicz, "Genetic Algorithms + Data Structures =Evolution Programs", AI Series.Springer-Verlag, New York, 3rd edition, (1996)

[24] M. Mitchell, "Genetic Algorithm: An Overview", pp: 31-39, (1995).

[25] L.W. Leung and B. King an, "Comparison of Image Data Fusion Techniques Using Entropy and INI". 22nd Asian Conference on Remote Sensing, November (2001)

[26] Workgroup on "Texture Analysis of DFG, "TILDA Textile Texture Database", http://lmb.informatik.uni-freiburg.de/ research/dfg-texture/tildeNarges Heidari, Reza Azmi & Boshra Pishgoo.

[27] P.Brodatz, Textures: A Photographic Album for Artists and Designers, Dover, New York, (1966).

[28] Depavath Harinath, "Enhancing Security through Steganography by using Wavelet Transformation and Encryption" in International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, Issue 08, [August– 2015]

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

Telecommunications Subscription Fraud Detection using Artificial Neural Networks

[1]Ledisi G. Kabari, [2]Domaka N. Nanwin and [3]Edikan Uduak Nquoh
[1]School of Applied Sciences, Ken Saro-Wiwa Polytechnic, Bori, Nigeria;
[2,3]Faculty of Natural and Applied Sciences, Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Nigeria;
ledisigiokkabari@yahoo.com; kakusman@yahoo.com; nquedt@gmail.com

ABSTRACT

Telecommunications Companies are facing a lot of problems due to fraud; hence the need for an effective fraud detection system for the telecommunications companies. This paper presents a design and implements of a subscription fraud detection system using Artificial Neural Networks. Neurosolutions for Excel was used to implement the Artificial Neural Network. The system was tested and found to be user friendly, effective and 85.7% success rate achieved.

Keywords: Telecommunications; Subscription; Artificial Neural Networks; Neurosolutions; Fraud; detection; Subscription Fraud.

1 Introduction

Telecommunications is fast turning the world into a global village. Globally, the development of telecommunications industry is rapidly increasing with one innovation replacing another in a matter of years, months, and even weeks. Without doubt telecommunication is a key driver of any nation's economy. Telecommunication is the communication of information by electronic means usually over some distance [1]. It involves the transmission and receipt of information, messages, graphics, images, voice, video and data between or among telephones, internet, satellites and radio.

The telecommunications sector is a wide sector with millions of users. This sector is broadly categorized into two categories based on its users; Domestic users and Commercial users.

The domestic users are provided with connections at an affordable rate while the commercial users are provided with connections at a comparatively higher rate as their usage scale is higher. But it has been discovered that there are cases where the subscription at the commercial level is fraudulently brought under the domestic level hence causing a significant loss to the sector. This kind of subscriptions if brought under the right category would have yielded a greater income to the sector. Fraud is defined as the deliberate and premeditated act perpetrated to achieve gain on false ground [2]. Fraud is also seen as any transmission of voice data across a telecommunications network, where the intent of the sender is to avoid or reduce legitimate call charges [3]. Telecommunication fraud is the theft of services or deliberate abuse of voice or data networks [4]. Telecommunications fraud can be broken down into

DOI: 10.14738/tmlai.36.1695
Publication Date: 14th December, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1695

several generic classes which describe the mode operators are defrauded but this paper focuses on **"SUBSCRIPTION FRAUD"**.

Subscription fraud is a contractual fraud. In these kinds of fraud revenue is generated through the normal use of a service without having to pay. In this scenario, the fraudster operates at level of phone numbers where all transactions from this number is fraudulent and all activities in such cases are further abnormal throughout the active period of the account. Subscription fraud can be divided into two categories:-

- Subscription fraud for the purpose of personal usage by the fraudster.
- Subscription fraud for profit. In this category the fraudster opens a small outfit where he starts up a call center. The fraudster has no intensions of paying his bills but he sells the airtime to people who intend to make cheap long distance calls for cash.

Fraud detection problems are found in many sectors of lives endeavor and the telecoms sector is not an exception. Hence fraud detection is referred to as the attempt engaged in discovering illegitimate usage of a communication network by identifying fraud as quickly as possible once it has been perpetrated [5].

A Neural Network consists of hardware or software that attempts to emulate the processing patterns of the biological brain [1]. It is a type of artificial intelligence system modeled after the neurons (nerve cells) in the biological nervous system and intended to simulate the way a brain processes information, learns and remembers. A neural network is designed as an interconnected system of processing elements, each with a limited number of inputs and outputs. These processing elements are able to "learn" by receiving weighted inputs that with adjustment, time and repetition can be made to produce appropriate outputs. Neural Networks are used in areas such as fraud detection, pattern recognition, speech analysis and speech synthesis.

The radical changes in the terrain of the telecommunications sector have made it difficult to control and detect fraudulent activities. Thus, to achieve positive results the problem of fraud requires to be handled with rapt and effective attention. Artificial Neural Networks has been found to be very useful in fraud detection, hence our usage.

The aim of this paper is to present a design and implement of a subscription fraud detection system via artificial neural networks. To achieve it aim the paper is set out to:-

I. Identify the different subscription services provided by the telecommunications sector.
II. Identify the different ways telecommunications fraud is perpetrated.
III. Utilize the artificial neural network model to detect subscription fraud in the telecommunications sector.

In terms of significance, the paper will benefit the private telecoms sector because they will be able to detect subscription fraud and hence reduce the losses incurred. The paper brought to limelight a novel means of combating telecommunications subscription fraud with the utilization of the intelligent agent like artificial neural networks. The paper will also benefit other researchers who intend to go into this area as this is the birthing of innovative ways of solving the problem of subscription fraud in the telecoms sector.

2 Telecommunication Services and Frauds

2.1 Telecommunications Mobile Subscription Services

Telecommunication companies in Nigeria provide different services to win the heart of their subscribers because of the market competition. However, some of these mobile subscription services are common to all the mobile operators in Nigeria, these includes: Prepaid Services and Postpaid Services.

2.1.1 Prepaid Services

This is the most popular of the services provided by mobile operators. As the name implies 'Pre-paid', all transaction in this service is pay-as-you-go. This service is easy for the mobile operator to maintain in the event of fraud and it is less susceptible to fraud as compared to postpaid services.

2.1.2 Postpaid Services

This is the most conventional service offered by mobile operators all over the world. As the name implies 'Post-paid', credit facilities is given for services used for some period of time, usually between 1-6 months. Though, this service is not common in Nigeria because there is no proper means of identification in case a subscriber defaults, yet, all mobile operators in Nigeria still render the service as a result of the stiff competition in the industry.

Other services provided by Nigeria's mobile telecommunication companies that are susceptible to fraud includes: Roaming Services, Value Added Features and Service (VAS), Premium Rate Services (PRS).

2.2 Telecommunication Fraud

Telecommunication industry has expanded dramatically in the last few years with the development of affordable mobile phone technology [6]. With the increasing number of mobile phone subscribers, global mobile phone fraud is also set to rise. It is a worldwide problem with substantial annual revenue losses of many companies. Telecommunication fraud which is the focus is appealing particularly to fraudsters as calling from the mobile terminal is not bound to a physical location and it is easy to get a subscription. This provides a means for illegal high profit business for fraudsters requiring minimal investment and relatively low risk of getting caught. Telecommunication fraud is defined as the unauthorized use, tampering or manipulation of a mobile phone or service.

Telecommunication fraud can be simply described as any activity by which telecommunications service is obtained without intention of paying. This kind of fraud has certain characteristics that make it particularly attractive to fraudsters. The main one is that the danger of localization is small. This is because all actions are performed from a distance which in conjunction with the mesh topology and the size of network makes the process of localization time consuming and expensive. Additionally no particularly sophisticated equipment is needed if one is needed at all. The simple knowledge of an access code, which can be acquired even with methods of social engineering, makes the implementation of fraud feasible. Finally, in the product of telecommunication fraud, a phone call is directly convertible to money.

2.3 Types of Telecommunication Fraud

The telecom industry suffers major losses due to fraud [7]and [8]. There are many different types of telecommunications fraud and these can occur at various levels. The two most common types of fraud are subscription fraud and superimposed fraud.

2.3.1 Subscription Fraud

In subscription fraud, fraudsters obtain an account without intention to pay the bill. Thus at the level of a phone number, all transactions from this number will be fraudulent. In such cases abnormal usage occurs throughout the active period of the account. The account is usually used for call selling or intensive self-usage.

2.3.2 Superimposed Fraud

In Superimposed fraud, fraudsters take over a legitimate account. In such cases the abnormal usage is superimposed upon the normal usage of the legitimate customers. There are several ways to carry out superimposed fraud, including mobile phone cloning and obtaining calling card authorization details. Examples of such cases include cellular cloning, calling card theft and cellular handset theft. Superimposed fraud will generally occur at the level of individual calls; the fraudulent calls will be mixed with the justified ones.

2.3.3 Intrusion fraud

This occurs when an existing, otherwise legitimate account, typically a business, is compromised in some way by an intruder, who subsequently makes or sells calls on this account. In contrast to subscription calls, the legitimate calls may be interspersed with fraudulent calls, calling for an anomaly detection algorithm.

2.3.4 Fraud based on loopholes in technology

Consider voice mail systems as an example. Voice mail can be configured in such a way that calls can be made out of the voice mail system (e.g., to return a call after listening to a message), as a convenience for the user. However, if inadequate passwords are used to secure the mailboxes, it creates vulnerability. The fraudster looks for a way into a corporate voice mail system, compromises a mailbox (perhaps by guessing a weak password), and then uses the system to make outgoing calls. Legally, the owner of the voice mail system is liable for the fraudulent calls; after all, it is the owner that sets the security policy for the voice mail system.

2.3.5 Social engineering

Instead of exploiting technological loopholes, social engineering exploits human interaction with the system. In this case the fraudster pretends to be someone he or she is not, such as the account holder, or a phone repair person, to access a customer's account. Recently, this technique has been used by "pre-texters" in some high-profile cases of accessing phone records to spy on fellow board members and reporters [9].

2.3.6 Fraud based on new technology

New technology, such as Voice over Internet Protocol (VoIP), enables international telephony at very low cost and allows users to carry their US-based phone number to other countries. Fraudsters realized that they could purchase the service at a low price and then resell it illegally at a higher price to consumers who were unaware of the new service, unable to get it themselves, or technologically unsophisticated. Detecting this requires monitoring and correlating telephony usage, IP traffic and ordering systems.

2.3.7 Fraud based on new regulation

Occasionally, regulations intended to promote fairness end up spawning new types of fraud. In 1996, the Federal Communications Commission (FCC) modified payphone compensation rules, requiring payphone operators to be compensated by the telecommunication providers. This allowed these operators to help cover the cost of providing access to phone lines, such as toll-free numbers, which do not generate revenue for the payphone operator. This spawned a new type of fraud—payphone owners or their associates placing spurious calls from payphones to toll-free numbers simply to bring in compensation income from the carriers.

2.3.8 Masquerading as another user

Credit card numbers can be stolen by various means (e.g., “shoulder surfing” looking over someone’s shoulder at a bank of payphones) and used to place calls masquerading as the cardholder. There are many more fraud techniques, some of which are quite sophisticated and combine more than one known method.

Telecommunications fraud is not static; new techniques evolve as the telecom companies put up defenses against existing ones. The fraudsters are smart opponents, continually looking for exploitable weaknesses in the telecom infrastructure. Part of their motivation is accounted for by the fact that once an exploit is defined, there are thousands (or millions) of potential targets. New types of fraud appear regularly, and these schemes evolve and adapt to attempts to stop them.

2.4 Telecommunication Fraud Detection

Fraud is a multi-billions problem around the globe. The problem with telecommunication fraud is the huge loss of revenue and it can affect the credibility and performance of telecommunication companies. The most difficult problem that faces the industry is the fact that fraud is dynamic. This means that whenever fraudster’s feel that they will be detected they find other ways to circumvent security measures. Telecommunication fraud also involves the theft of services and deliberate abuse of voice and data networks. In such cases the perpetrator’s intention is to completely avoid or at least reduce the charges for using the services. Over the years, fraud has increased to the extent that losses to telephone companies are measured in terms of billions of American dollars. Fraud negatively impacts on the telephone company in 4 ways such as financially, marketing, customer relations and shareholder perceptions.

There are various **techniques** available for managing and detecting telephone fraud these include:

2.4.1 Manual review of data

The problem with this technique is the fact that there are too many data records for a team to filter the

fraudulent data. Typically a telecom company will have in order of 1 million or more records of telephone calls generated by their customers for a single month within a specific region. As a result this is a time consuming and laborious technique for detecting fraud.

2.4.2 Conventional analysis

This is the fixed rule based expert system together with statistical analysis. A rule based system is a set of rules that take into account the normal calling hours, the called destinations as well as the normal duration of the call.

2.4.3 Adaptive flexible techniques

This is using advanced data analysis like artificial neural networks (ANNs). Fed with raw data, a neural network can quickly learn to pick up patterns of unusual variations that may suggest instances of fraud on a particular account [10].

2.5 Approaches to Face Fraud

Dealing with the fraud is a very complex task mainly due to its transversal nature to the operator's structure [11]. Traditional fraud techniques are evolving and adapting to the new network infrastructure. The fraud techniques are considered because basic ideas remain despite the underlying technology. Deceptions in telecommunications include subscription frauds where the cheater accesses the services without being subscribed. User can also suffer line or identity theft being charged for services used by others. Telecommunication operators can oversee users that exceed their download quote and rate performing illegal service redistribution, sometimes for an economic profit. Finally cloning or unauthorized access to services may lead to compromising privacy.

Anyway the most common types of fraud on telecommunications are subscription fraud and identity theft. After that voice mail fraud and calling card fraud prevail. The analysis of different fraud techniques points out that the tendency is a convergence of the fraud which increases the complexity of its detection.

Fraud management systems have proved to be a suitable tool to detect fraud in different networks with diverse techniques such as self-organizing maps (SOM), general data mining, rule based systems profiling through Artificial intelligence techniques like neural networks or decision trees based on the hierarchical regime switching models, Bayesian networks, fuzzy rules or other data mining techniques. There also exist works on how to discover new rules to detect fraud in telecommunications and on the privacy concerns of applying detection techniques to user's data.

Fraud detection can also be done at 2 levels call or behavior and with two different approaches user profile or signature based [12]. Most of the techniques use the CDR data to create a user profile and to detect anomalies based on these profiles. The mined large amounts of CDR have in order to find patterns and scenarios of normal usage and typical fraud situations. These scenarios were then used to configure monitors that observe the user behavior with relation to that type of fraud. These monitors are then combined in a neural network which raises an alarm when sufficient support of fraud exists. This type of system can be classified in a rule based approach since it relies in the triggering of certain rules due to abnormal usage. The rule based system has the drawback of requiring expensive

management of rules. Rules need to be precise (avoid false positive alarms) and constantly evolving (detect new scenarios) which result in very time consuming programming.

The most common and best succeeded methods for fraud analysis are signature based. These methods detect the fraud based on deviation detection by comparing the recent activity with the user behavior data which is expressed through the user signature. The work can be adapted and extended by reformulating the notion of signature and by introducing the notion of statistical based distances to detect anomalies. Furthermore the Computational cost can be reduced by using simple statistical functions avoiding processing costly histograms. A clear problem with a histogram approach is that discretization intervals or bucket must be clean and what is right for one customer may be wrong for another. Other approaches have also been widely applied to fraud analysis like neural networks. Another applied technique is link analysis. Here the client links are updates over time establishing a graph of called communities of interest that can easily reveal networks of fraudster's. These methods are based on the observation that fraudsters seldom change their calling habits but are often linked to other fraudsters.

3 Methodology

The proposed system is a system based on the artificial neural network technology. This system is supposed to handle the challenges encountered by the rule based system of detecting fraud. Artificial Neural Networks is a more efficient mechanism when it comes to telecommunications fraud detection because it can accommodate large amounts of data, analyze it and possibly prompt the network of the fraud to be perpetrated.

The inputs are the data of a subscriber such as the name, age, sex, registration duration, subscription bundle and amount of data used. The data is processed so as to ascertain if the subscriber is fraudulent or non-fraudulent. It outputs the state of a particular subscriber. That is "fraudulent" or "non-fraudulent".

3.1 Design of the Proposed System

We took five basic steps during the design process: Collecting data, preprocessing data, Building the network, Train, and Test performance of the model.

Collecting and preparing sample data is the first step in designing ANN models. The data is gotten from historical data of fraudulent and non-fraudulent subscriptions.

After data collection, three data preprocessing procedures are conducted to train the ANNs more efficiently. These procedures are: (1) solve the problem of missing data, (2) normalize data and (3) randomize data. The missing data are replaced by the average of neighboring values during the same week. Normalization procedure before presenting the input data to the network is generally a good practice, since mixing variables with large magnitudes and small magnitudes will confuse the learning algorithm on the importance of each variable and may force it to finally reject the variable with the smaller magnitude.

We then decided on the number of hidden layers, neurons in each layer, transfer function in each layer, training function, weight/bias learning function, and performance function. In this work, multilayer perceptron (MLP) and radial basis function (RBF) networks are used.

During the training process, the weights are adjusted in order to make the actual outputs (predicated) close to the target (measured) outputs of the network**.**

3.2 Knowledge Base

We used Microsoft excel to design our database because it is simple to access and use. It contains the data in the form of fields, records and attributes. It contains all that is needed to train the network so that it can be able to detect the fraud intuitively. It is seen as the store of all the domain and fraud specific knowledge that is required by the system.

3.3 Knowledge Base of Rules

Knowledge base of rules was constructed. This could also be referred as the inference engine. They are the rules that will be used to train the network so that when a record is to be tested to ascertain if it is fraudulent or not it can do that effectively. This is where the actual fraud detection methods and techniques are stored. This is done in the form of "if-then-else" rules that reason over the knowledge contained in the knowledge base. The rules for detection and the knowledge base are interrelated as the representation of the rules depends on the contents of the knowledge base.

3.4 Store of Nature of Fraud

This is a store of all the types of fraudulent outputs gotten. This store will be fed into the self-learning system.

3.5 Self-Learning System

This provides the overall system with the capabilities to learn new rules about fraud from the submission of fraudulent and non-fraudulent domain specific data and to automatically detect irregular observations in the data thus providing a feedback mechanism for enriching and updating the knowledge base of rules.

3.6 Architecture of the Proposed System

The architectural design of the proposed system is as shown in figure 1.

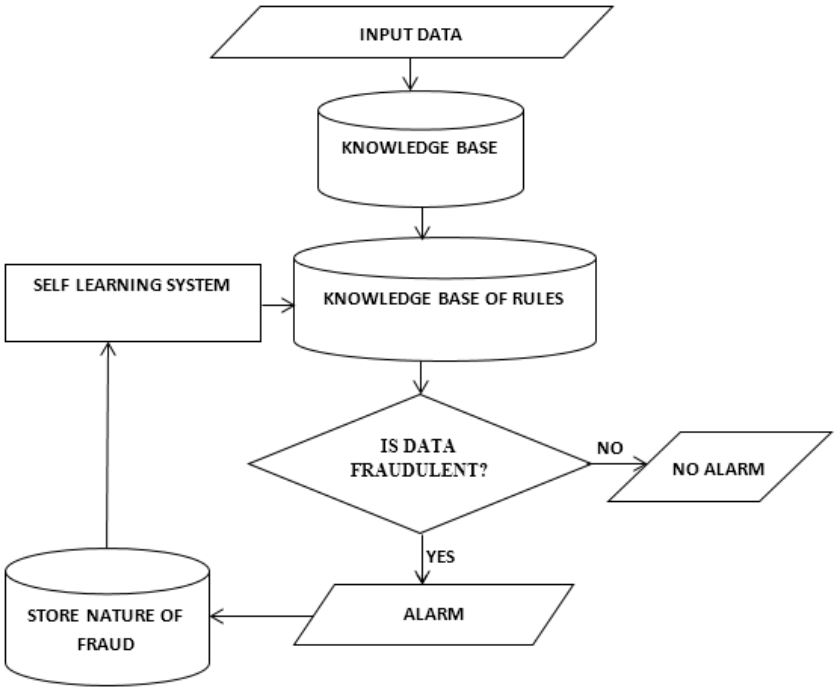


Figure 1: Architectural Design of the System

4 Experiments and Analysis

To implement the proposed system, an Artificial Neural Network (ANN) software called NeuroSolutions was used which is considered one of the leading software available for implementing neural networks. Neural networks are long, complicated mathematical equations and NeuroSolutions is designed to make the technology easy and accessible to both novice and advanced neural network developers. There are three basic phases in neural network analysis: training the network on your data, testing the network for accuracy and making predictions/classifications from new data. Only the Express Builder in the NeuroSolutions Excel interface can accomplish all of this automatically in one simple step. With the NeuroSolutions for Excel interface, it has never been easier to get started quickly in solving your problem. It provides an easy-to-use and intuitive interface for users to easily setup a simulation that automatically builds, trains and tests multiple neural network topologies and generates an easy-to-read report of the results including the best-performing model.

Neurosolutions is a numerical computing environment and also a programming language. It allows easy matrix manipulation, plotting of functions and data, implementation of algorithms, creating user interfaces and interfacing with programs in other languages. The Neural Network Toolbox contains the Neurosolutions tools for designing, implementing, visualizing and simulating neural networks. It also provides comprehensive support for many proven network paradigms, as well as graphical user interfaces (GUIs) that enable the user to design and manage neural networks in a very simple way. NeuroSolutions provides an easy-to-use and intuitive user interface for Microsoft Excel. It simplifies and enhances the process of getting data into and out of a NeuroSolutions neural network. It also benefits both novice and advanced neural network developers by offering an easy-to-use, yet extremely powerful features (Neurosolutions, 2014).

The collected dataset is divided into three data subsets: first data subset (60% of the whole dataset) was used as inputs for the ANN training phase, and second data subset (15% of the dataset) was used as inputs for the ANN cross validation. The last data subset (25% of the dataset) was used as inputs for the ANN Testing. The system uses the training dataset to predict cases that are fraudulent or not. If there is a big gap (deviation) between the output *parameters* actual values and the predicted values, there is a high probability of fraud in the fraud report, otherwise there is none.

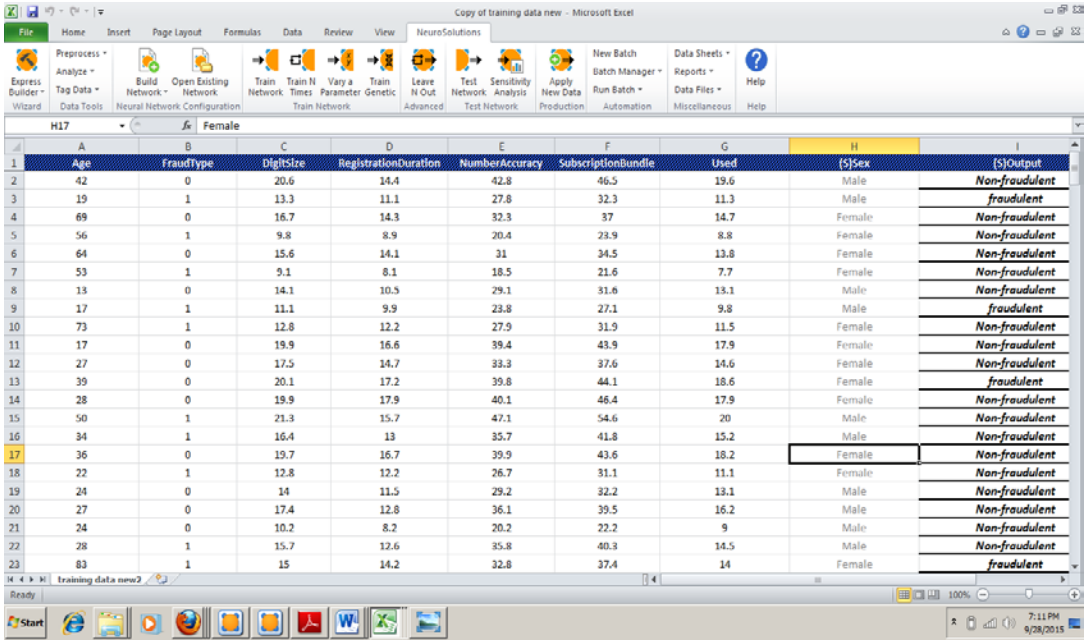
4.1 System Requirement

All computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as system requirements and are often used as a guideline.

For an effective operation, the system can be implemented provided the following hardware and software components are at least met. Processor: Pentium 4, Intel Core Duo or higher, RAM: 3 GB, HDD: 650 MB and above, Keyboard: Enhanced keyboard, Mouse: Enhanced serial or parallel mouse, CRT: 15" colored monitor, Printer: Optimal (Colored/black and white), Operating System: Windows XP, VISTA, 7, 8.1 either 32-bits or 64-bits, Microsoft Excel: 97, 2000, 2002, 2003, 2005, 2007, 2010 and 2013, Neurosolutions 7.0.

4.2 Launching of the Network and Results

We choose the default setup for "Classification PNN (< 10,000 Rows)" since our data has less than 1,000 total samples. A sample for training data is shown in figure 2. Specifying Training, Cross Validation and Testing Data Segments is shown in figure 3. Building the network and result after building are shown in figure 4 and figure 5 respectively. After training and Network testing are shown in figure 6 and figure 7 respectively. The output is as shown in figure 8.



	A	B	C	D	E	F	G	H	I
1	Age	FraudType	DigitSize	RegistrationDuration	NumberAccuracy	SubscriptionBundle	Used	(S)Sex	(S)Output
2	42	0	20.6	14.4	42.8	46.5	19.6	Male	Non-fraudulent
3	19	1	13.3	11.1	27.8	32.3	11.3	Male	fraudulent
4	69	0	16.7	14.3	32.3	37	14.7	Female	Non-fraudulent
5	56	1	9.8	8.9	20.4	23.9	8.8	Female	Non-fraudulent
6	64	0	15.6	14.1	31	34.5	13.8	Female	Non-fraudulent
7	53	1	9.1	8.1	18.5	21.6	7.7	Female	Non-fraudulent
8	13	0	14.1	10.5	29.1	31.6	13.1	Male	Non-fraudulent
9	17	1	11.1	9.9	23.8	27.1	9.8	Male	fraudulent
10	73	1	12.8	12.2	27.9	31.9	11.5	Female	Non-fraudulent
11	17	0	19.9	16.6	39.4	43.9	17.9	Female	Non-fraudulent
12	27	0	17.5	14.7	33.3	37.6	14.6	Female	Non-fraudulent
13	39	0	20.1	17.2	39.8	44.1	18.6	Female	fraudulent
14	28	0	19.9	17.9	40.1	46.4	17.9	Female	Non-fraudulent
15	50	1	21.3	15.7	47.1	54.6	20	Male	Non-fraudulent
16	34	1	16.4	13	35.7	41.8	15.2	Male	Non-fraudulent
17	36	0	19.7	16.7	39.9	43.6	18.2	Female	Non-fraudulent
18	22	1	12.8	12.2	26.7	31.1	11.1	Female	Non-fraudulent
19	24	0	14	11.5	29.2	32.2	13.1	Male	Non-fraudulent
20	27	0	17.4	12.8	36.1	39.5	16.2	Male	Non-fraudulent
21	24	0	10.2	8.2	20.2	22.2	9	Male	Non-fraudulent
22	28	1	15.7	12.6	35.8	40.3	14.5	Male	Non-fraudulent
23	83	1	15	14.2	32.8	37.4	14	Female	fraudulent

Figure 2: Sample of training data from excel

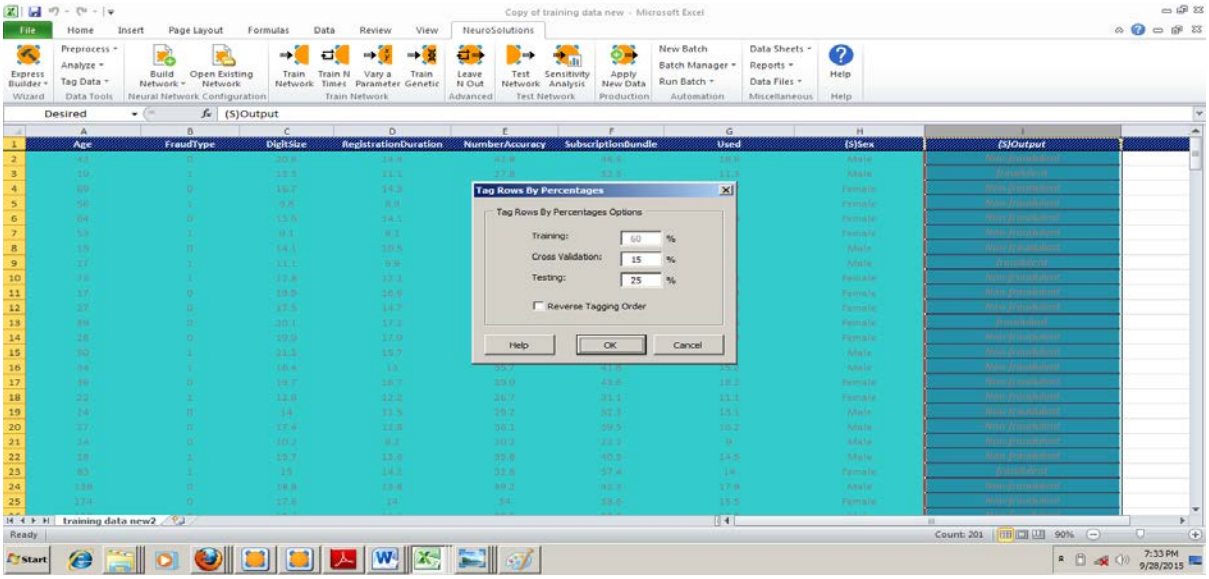


Figure 3: Specifying Training, Cross Validation and Testing Data Segments

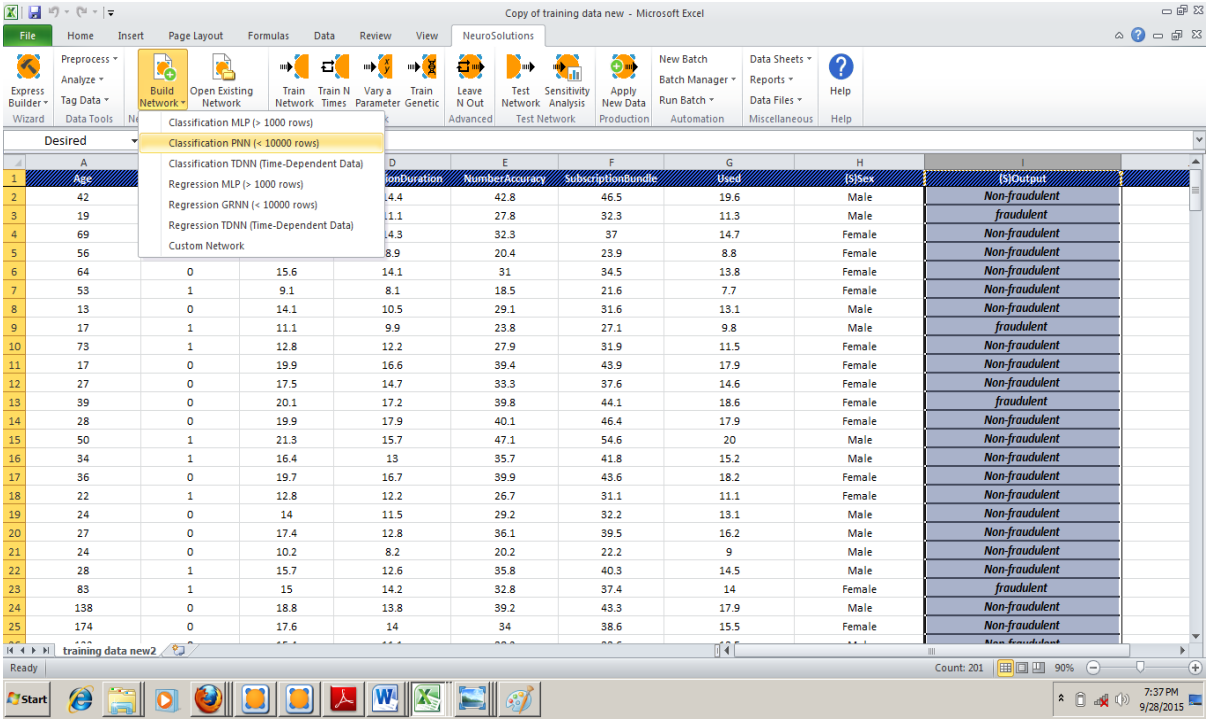


Figure 4: Building the Network

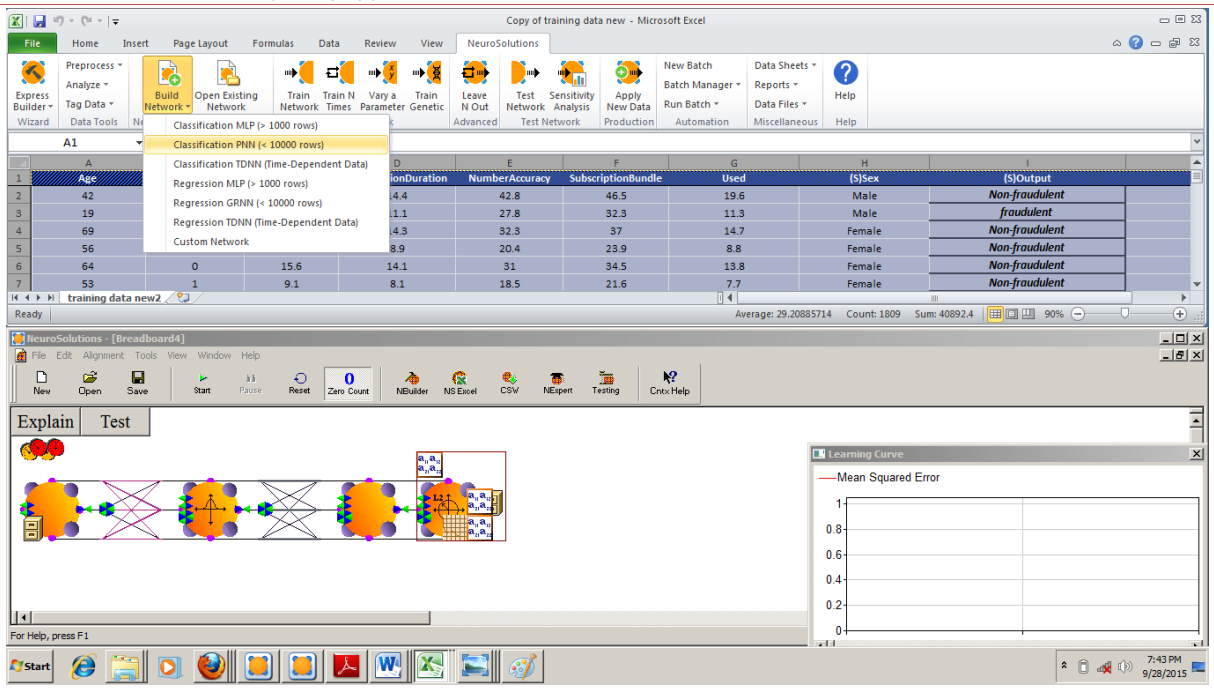


Figure 5: Result after Building Network

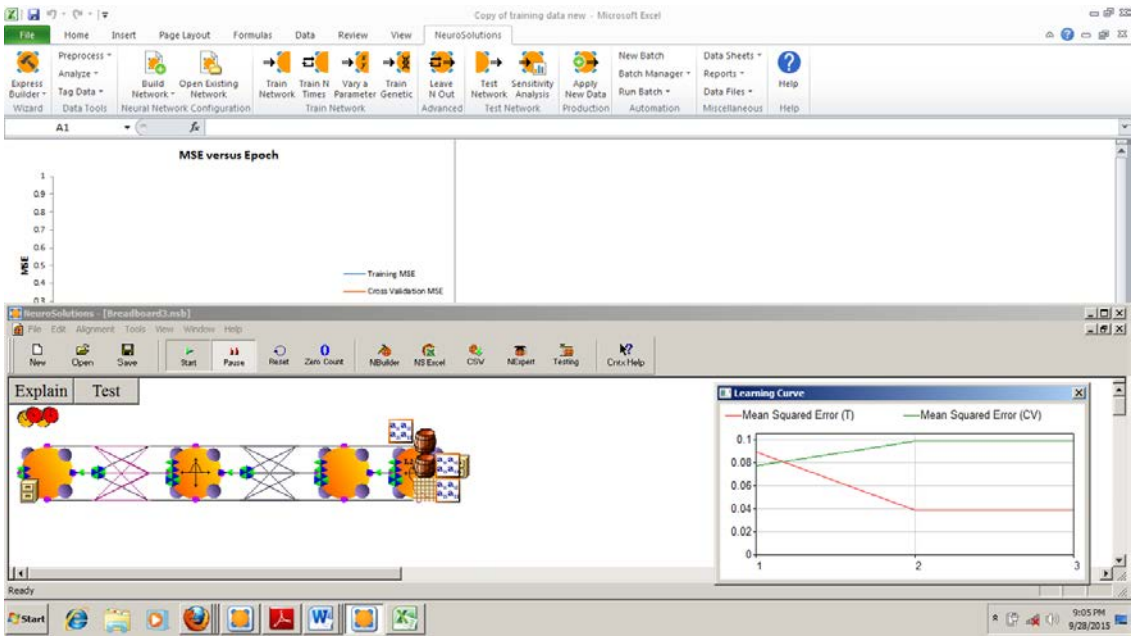


Figure 6: After Training

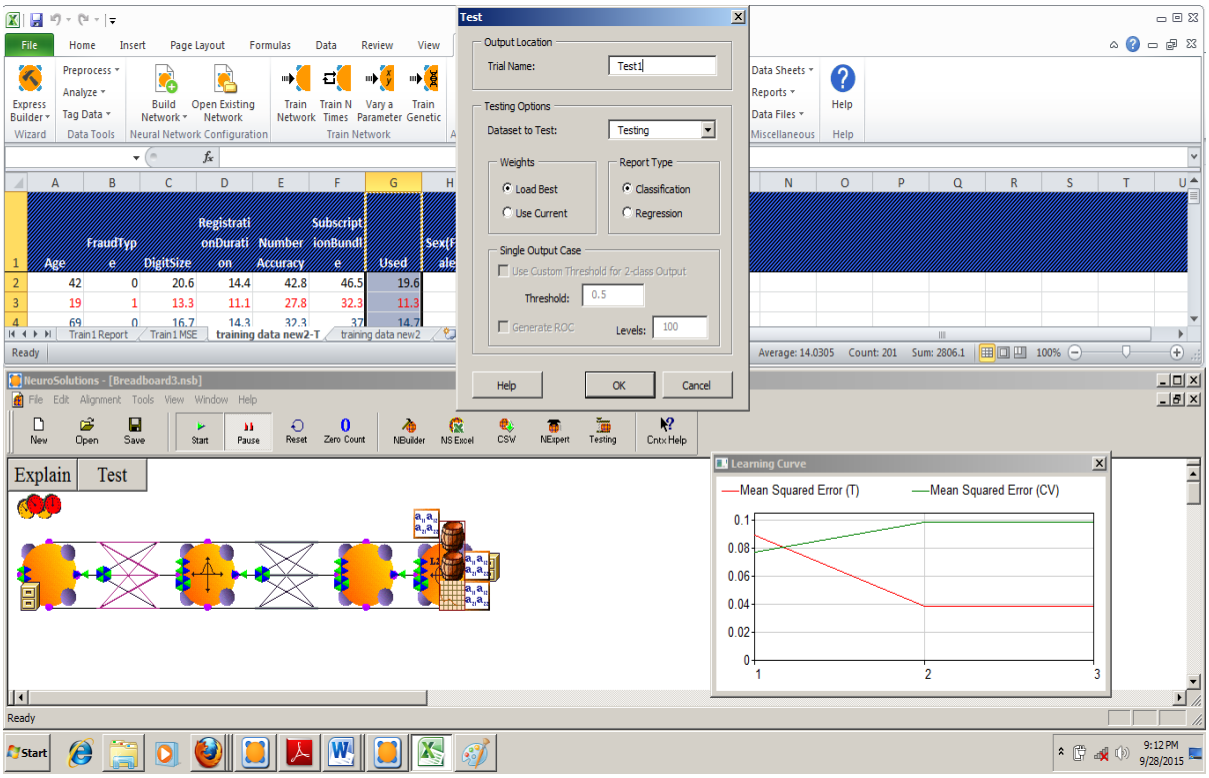


Figure 7: Neural Network Testing

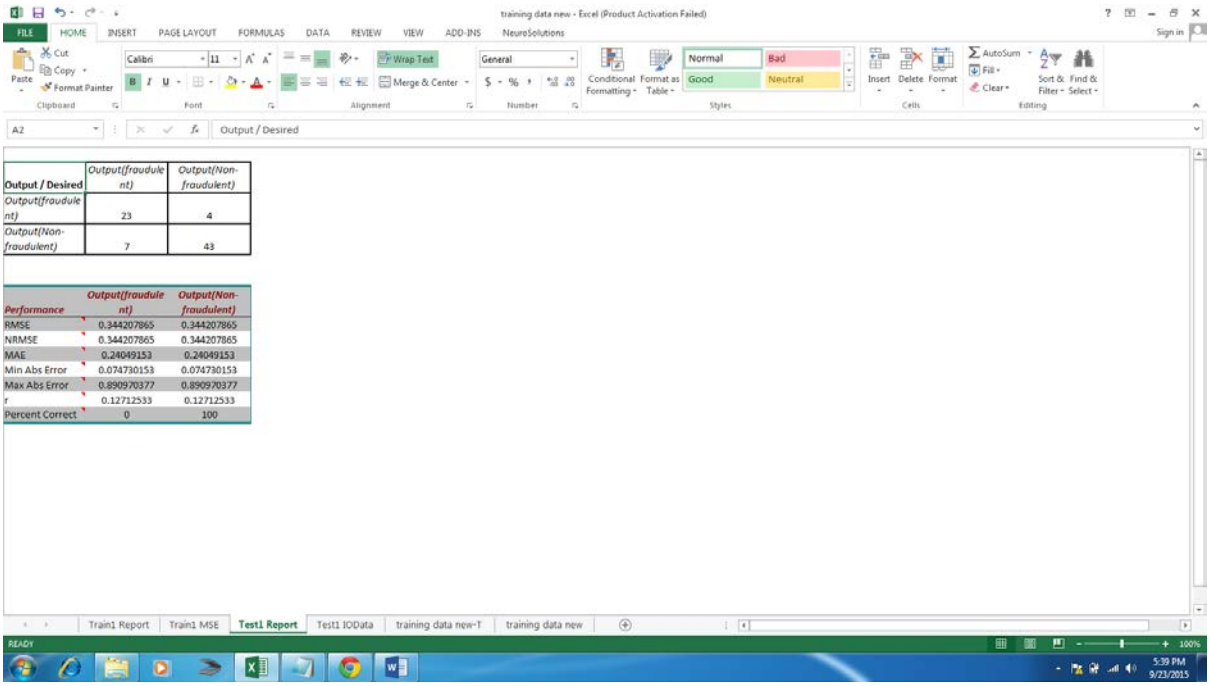


Figure 8: Desired Output Report

The end result is an easy to understand report displaying a confusion matrix of the results as well as some general statistics about the performance of the model on the output of the sample data. The confusion matrix is read from the top left and goes diagonally to the right displaying the correct predictions. So for this particular training run we classified 23 Fraudulent and 43 Non-Fraudulent

correctly with 7 Fraudulent and 4 Non-Fraudulent misclassified. Total number of sample data detected is 77, number classified correctly is 66 and 11 was misclassified. Hence, the percentage of success is $66/77 * 100 = 85.7\%$, and the failure rate is $11/77 * 100 = 14.3\%$. This makes a total of 100%.

5 Conclusion

Fraud detection problems are found in many sectors of lives endeavor and the telecoms sector is not an exception. Hence fraud detection is referred to as the attempt engaged in discovering illegitimate usage of a communication network by identifying fraud as quickly as possible once it has been perpetrated.

The radical changes in the terrain of the telecommunications sector have made it difficult to control and detect fraudulent activities. Thus, to achieve positive results the problem of fraud requires to be handled with rapt and effective attention. Artificial Neural Networks has been found to be very useful in fraud detection, hence its usage.

The work thus identified different subscription services provided by the telecommunications sector, identify the different ways telecommunications fraud is perpetrated and utilized the artificial neural network model to design and implement a subscription fraud detection system for the telecommunications sector.

A lot of telecommunications companies are not ready to give reports as it concerns their existing fraud detection system as they deem it confidential, hence analysis of their system was not possible. Thus, we generate the possible data and simulate the process of fraud detection since we were unable to access real data sets.

REFERENCES

[1]. Laudon K. C., Laudon, J. P. Brabston, M. E, *Management Information System*: Managing the Digital Firm: Pearson Education Canada Inc; Toronto, Ontario, 2002.

[2]. Alexopoulos, P. and Kafentzis, K., *Towards a Generic Fraud Ontology in E Government,* ICE-B*,* 2007. p. 269-276.

[3]. Hollmen, J., *User Profiling and Classification for Fraud Detection in Mobile Communication Networks*, PhD thesis, Helsinki University of Technology, Department of Cognitive and Computer Science and Engineering. Espoo, Finland. 2000.

[4]. Hiyam, A. E. Tawashi, *Detecting Fraud in Cellular Telephone Networks Jawwal Case Study*. MBA thesis Islamic University, Faculty of Commerce, Department of Business Administration, Gaza. 2010.

[5]. Bolton, R. J. and Hand, D. J., *Statistical Fraud Detection, A review*, Institute of Mathematical Statistics*,* 2002. 17(3), p. 235–255.

[6]. Pieprzyk, J., Ghodosi, H. and Dawson, E., *Information Security and Privacy, 12th Australasian Conference*, ACISP 2007, Townsville, Australia, July 2-4, 2007: Proceedings, Springer, Germany, p. 446-447.

[7]. Prasad, S. K., Routray, S. and Khurana, R., *Information Systems, Technology and Management*. *Third International Conference, ICISTM 2009, Ghaziabad, India, March 12-13, 2009*, Proceedings, Springer, Germany, p. 259-260

[8]. Żytkow, J. M. and Rauch, J., *Principles of Data Mining and Knowledge Discovery*: *Third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999: Proceedings*, Springer, USA, p. 251.

[9]. Kaplan, D. A., *Intrigue in High Places*, *To Catch a Leaker*, Hewlett– Packard's Chairwoman Spied on the Home–Phone Records of Its Board of Directors, *Newsweek* (September), 2006.

[10]. Liatsis, P., *Recent Trends in Multimedia Information Processing*, Proceedings of the 9th International Workshop on Systems, Signals and Image Processing, *World Scientific Publishing*, London, 2002. P. 474-475.

[11]. Samarati, P., *Information Security Theory and Practices*: *Security and Privacy of Pervasive Systems and Smart Devices*: *4th IFIP WG 11.2 International Workshop, WISTP 2010*, Passau, Germany, April 12-14, 2010, Proceedings, Springer, USA, p. 201.

[12]. Perner, P., *Advances in Data Mining, Applications in Medicine, Web Mining, Marketing, Image and Signal Mining*: *6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006:* Proceedings, Springer, Germany, p. 535.

[13]. Neurosolutions for Excel. *http://www.neurosolutions.com/documentation/NeuroSolutionsforExcel.pdf* Retrieved, September 28, 2015.

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

An Unsupervised Neural Network Method for Age Group Estimation using Facial Features

[1]Oladele, M.O. [2]Omidiora, E.O. and [2]Adepoju, T.M.
[1]Department of Computer Engineering, The Federal Polytechnic, Ede, Nigeria
[2]Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
m_omotayo@yahoo.com; eoomidiora@lautech.edu.ng; atemilola@gmail.com

ABSTRACT

Age estimation from facial features is an important research subject in the field of face recognition. It is an active research area that can be used in wide range of applications such as surveillance and security, telecommunication and digital libraries, human-computer intelligent interaction, and smart environment. This paper developed an unsupervised neural network by using a Self – Organizing Feature Map (SOFM) to estimate age group from facial features.

The face images were divided into eight different age groups ranging from babies, young teenagers, mid teenagers, teenagers, young adults, mid adults, young old and old and SOFM was used to estimate the age group from the input face image. Principal Component Analysis (PCA) was used to extract the facial features and the extracted features were presented to the SOFM for training and testing. The developed system was experimented with 630 face images with different ages from the FG-NET database. 450 samples were used for training while 180 were used for testing. The results showed a training time of 116.333 seconds and an accuracy of 92.2%.

Keywords: Age estimation, Facial Features, Unsupervised Neural Network, Principal Component analysis, Self-Organizing Feature Map.

1 Introduction

The face is one of the most important biometric features of the human being which is normally used for identification. Each person has their own innate face and mostly a different face. As a human, to recognize the different faces without any difficulty is easier but it is difficult for systems to recognize human faces [6]. Face recognition system can be used in various research applications such as age estimation, gender determination, emotion detection, head orientation etc [7].

Human age classification is one of the most challenging problems in computer vision and pattern recognition. Estimating human age from his or her face is a hard problem not only for the existing computer vision systems but also for humans in some circumstances [1]. Age classification is concerned with the use of a training set to train a model that can estimate the age of human from face images.

The developed system was able to group the ages into corresponding clusters or groups, then use the extracted facial features to classify the input face image into the corresponding age group using the

DOI: 10.14738/tmlai.36.1689
Publication Date: 17th December, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1689

SOFM. Principal Component Analysis (PCA) was used to extract the facial features. It is an analytical tool used in identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences (Oladele, Omidiora and Afolabi, 2015). Principal Component Analysis is a suitable strategy for feature extraction because it identifies variability between human faces, which may not be immediately obvious [8].

2 Existing Age Group Methods

Various age estimation methods has been developed so far in the field of face recognition. In human computer interaction, aging effects in human faces has been studied from two main reasons: automatic age estimation for face image classification and automatic age progression for face recognition [9].

Kwon and Lobo uses two main types of features: Geometrical ratios calculated based on the distance and the size of certain facial characteristics and an estimation of the amount of wrinkles detected by deformable contours (snakes) in facial areas where wrinkles are usually encountered [5].

In 2001, Horng, Lee and Chen proposed an approach for classification of age groups based on facial features [4]. The process of the system was mainly composed of three phases: Location, feature extraction and age classification. Two back propagation neural networks were constructed. The first one employs the geometric features to distinguish whether a facial image is a baby or not. If it is not, then the second network uses the wrinkles features to classify the image into one of three adult groups.

Yang and Ai used Real AdaBoost algorithm to train a classifier by composing a sequence of Local Binary Pattern (LBP) features as a representation of face texture. Age is classified into only three periods: child, adult and oldness [10]. In 2009, Guo, Fu, Dyer and Huang presented a locally adjusted regressor which uses age manifold learning to map pixel intensity of the original face images into a low dimensional subspace for the learning and the prediction of the aging patterns [3].

In 2015, Oladele, Omidiora and Afolabi used principal component analysis to extract facial features and back propagation neural network to classify age into age groups: babies, young teenagers, mid teenagers, teenagers, young adults, mid adults, young old and old [7]. The experiment produced an accuracy of 82.2% and Mean Absolute Error of 3.88 when experimented with 180 testing samples.

3 Design Approach

The block diagram for the age group estimation system is described in Figure 1. The first stage is the image pre-processing stage which comprises of the grayscale conversion, image cropping and resize. The second stage is the facial feature extraction stage where the facial features were extracted using PCA and finally the age estimation stage which uses SOFM to estimate the age group.

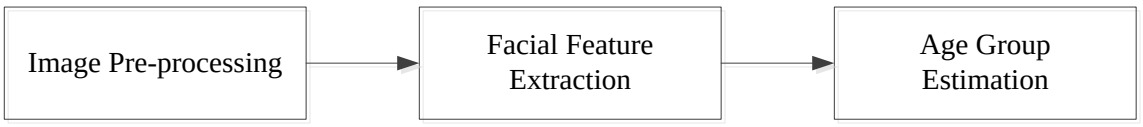


Figure 1: Block diagram of the Age group Estimation System

3.1 Image Database

Images from the FG-NET database was used for the system. From the FG-NET database, 630 face images of different ages of individuals were used. The images were divided into training and testing sets. 450 samples were used for the training set while 180 samples were used for the testing set. The images ranges from ages 0 – 69 years. Figure 2 shows some of the sample face image from the FG-NET database.

Figure 2: Sample images from the FG-NET database [2]

3.2 Image Pre-processing

Image pre-processing stage comprises of the grayscale conversion, image cropping and image resizing. This stage helps to get rid of unwanted information that would have been extracted as features and reduces the work to be done during dimensionality reduction (feature extraction). Grayscale conversion was used to reduce the number of pixels. Cropping was done using the Viola-Jones algorithm so as to remove the irrelevant features of the face image. The image was then resized to 40 by 40. Figure 3 shows the flowchart of the image pre-processing stage.

3.2.1 Grayscale Conversion

FG-NET database consists of both coloured image and gray image. The coloured image were converted into grayscale using the MATLAB function rgb2gray so as to reduce processing time being a two-dimensional matrix. After the image was converted to grayscale, the facial part of the image was detected so as to remove unwanted information such as the image background.

3.2.2 Image Cropping

The detected face image was cropped such that the image was left only with the facial region which contains the necessary features needed for the age group estimation. Cropping helps to remove unnecessary features such as the background from the image, leaving just the portion that is needed to provide the required age – relevant face features. The facial part of the images was detected using the Viola-Jones algorithm. The body parts detected by the Viola-Jones algorithm were frontal face, a single or pair of eyes, Nose and Mouth.

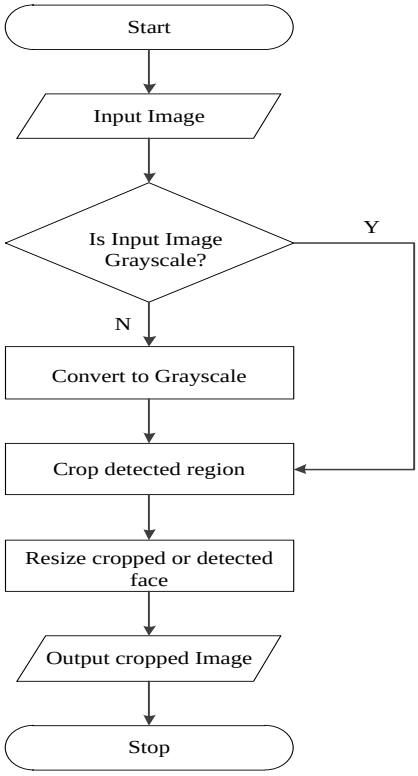


Figure 3: Flowchart of the image pre-processing stage

3.2.3 Image Resize

Having obtained the required portion of a given facial image it is important to ensure that the cropped portion of the facial image is neither too small nor too big for further processing. Therefore, it is important to choose an appropriate size to which images will be scaled to avoid image distortion. The images used in this research were resized to 40 by 40 pixels which contains only the facial part of the original image. The MATLAB function imresize was used to resize the image.

3.3 Facial Feature Extraction

Principal Component Analysis (PCA) was used to extract the facial features from the images. PCA is a suitable strategy for feature extraction because it identifies variability between human faces, which may not be immediately obvious. In facial feature, these variables are called eigenfaces because when plotted they display a ghostly resemblance to human faces.

In PCA, Eigenface finds the principal components of the distribution of faces, or the eigenvectors of the covariance matrix of the set of face images. These eigenvectors is a set of features that together characterize the variations between face images. The highest 20 eigenvectors were used in this work. Facial features including Eyes, Nose, Lips and Chin were extracted using the PCA.

3.4 Age Group Estimation

The final stage is the age group estimation using the unsupervised neural network. This stage consists of two phases: training and testing phases. In the training phase, the extracted PCA features were presented to the SOFM for training. SOFM is a clustering algorithm. During training, it classifies a set of input vectors of face parameters in a number of clusters corresponding to different age groups. For the

testing stage, when given a new vector of face parameters, i.e. the input face image, the trained networks will determine the age group of the person corresponding to the face image and output the age group. Figure 4(a) and Figure 4(b) shows the training and testing stages.

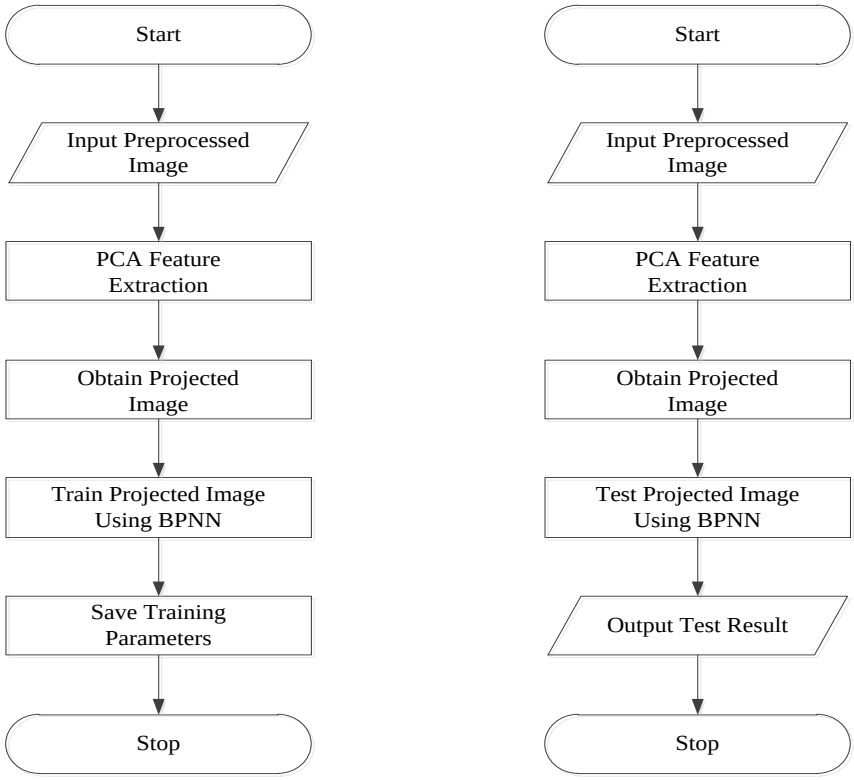


Figure 4(a): Training stage for Age Group Estimation

The results of the developed system were evaluated using total training time (total time to train the face images), and percentage accuracy. The metric that was used for the evaluation of the percentage accuracy is:

$$Percentage\ Accuracy = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples} * 100$$

4 Results and Discussion

The results obtained from the developed system with a training sample of 450 images and testing sample of 180 images (40% of the training images) shows a training time of 116.333 seconds. A total of 156 images were classified correctly, 10 images were near correct classification and 14 images was wrongly classified which shows an accuracy of 92.2%. Near-correct classification is the estimated age that fall above or below the next age group with an estimated error of ±2 years. Table 1 shows the results of SOFM on the different age groups.

Table 1: Results of Self Organizing Feature Map on the Age groups

Age Group	Number of Samples	Number of Correctly Classified	Number of Incorrectly Classified	Number of Near Correct Classification	Average Testing Time (Seconds)	Accuracy
(Baby) 0 – 5	25	25	0	0	17.512	100%
Young Teenager (6 – 10)	30	28	0	2	23.649	93.3%
Mid Teenager (11 – 15)	30	26	0	4	25.152	100%
Teenager (16 – 20)	25	17	4	4	18.744	84%
Young Adult (21 – 30)	25	22	3	0	19.263	88%
Mid Adult (31 – 50)	30	25	5	0	24.399	83.3%
Young Old (51 – 60)	10	9	1	0	17.860	90%
Old (Above 60)	5	4	1	0	12.876	80%

From Table 1 it was observed that the age group above 60 has the least accuracy of 80% while age ranges 0 – 5 and 11 – 15 have the highest accuracy of 100%. Also, a total of 10 samples fall to the near-correct classification category. This category lies between young teenagers and teenagers i.e. age group 6 – 20 because this age range reveals the similarities and distinguishing characteristics that constitutes facial textures during the formative years noticeable within the age group.

The developed system was further tested with some black faces and the results in Table 2 was gotten. From the table, it was observed that the performance was not as accurate as desired when compared with the test results from white faces that was used for training. This could be attributed to different skin textures peculiar to both black and white skins.

It was also observed that illumination affects the results when tested with the black faces, which can as well affect the performance. Therefore, accurate result can be gotten when the system is trained with black faces.

Table 2: Results of developed system when tested with black faces

S/N	Index Number	Real Age	Real Age Group	Estimated Age Group	Classification Result
1	090A04	4	Baby	Baby	CC
2	090A50	50	Mid Adult	Mid Adult	CC
3	090A30	30	Young Adult	Young Adult	CC
4	090A07	7	Young Teenager	Mid Teenager	NC
5	090A12	12	Mid Teenager	Baby	IC
6	090A14	14	Mid Teenager	Baby	IC
7	090A52	52	Young Old	Teenager	IC
8	090A15	15	Mid Teenager	Young Adult	IC
9	090A48	48	Mid Adult	Mid Adult	CC
10	090A00	0	Baby	Baby	CC

Figure 5 shows the output of the developed age group estimation system

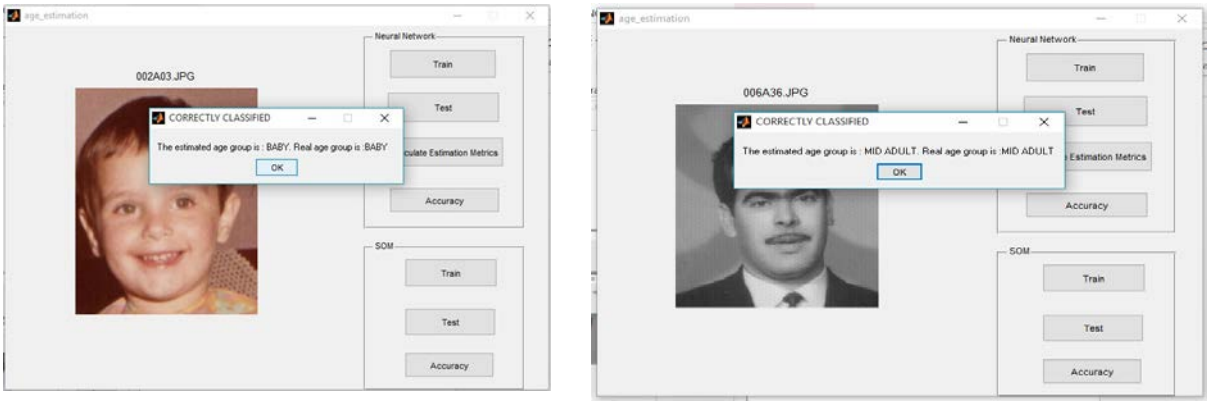


Figure 5: Output of the developed Age group Estimation System

5 Conclusion

An unsupervised neural network system for age group estimation with an improved accuracy was developed to extract facial features and determine age group from face images using the extracted facial features. The face images were classified into eight groups: babies, young teenagers, mid teenagers, teenagers, young adults, mid adults, young old and old. An accuracy of 92.2% was achieved and thus provided an improved method for age group estimation in terms of accuracy.

REFERENCES

[1] Cootes, T., Edwards, G and Taylor, C. (2001), "Active Appearance Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 6, pp. 681-685.

[2] FGNET (2015), "The FG-NET Aging Database", [Online]. Available: http://www.fgnet.rsunit.com/, 2015.

[3] Guo Guodong, Mu Guowang, Fu Yun and Huang S. Thomas (2009), "Human Age Estimation Using Bio-inspired Features", IEEE Conference on Computer Vision and Pattern Recognition, pp.112-119.

[4] Horng, W., Lee, C. and Chen, C. (2001), "Classification of Age Groups Based on Facial Features", Journal of Science and Engineering, Vol. 4, No. 3, pp. 183-192.

[5] Kwon, Y. H. and Lobo, N. Da Vitoria (1999), "Age classification from facial images", Computer Vision Image Understanding, Vol. 74, No. 1, pp. 1-21

[6] Nagi Jawad (2007), "Design and Development of an Efficient High-speed Face Recognition System" Final Thesis, College of Engineering, Universiti Tenaga Nasional, Selangor, Malaysia.

[7] Oladele, M.O., Omidiora, E.O., and Afolabi, A.O. (2015), "A Face-based Age Estimation System using Back Propagation Neural Network Technique", British Journal of Mathematics and Computer Science (In Press).

[8] Omidiora, E.O. (2006), "A Prototype of Knowledge-Based System for Black Face Recognition System using Principal Component Analysis and Fisher Disriminant Algorithms", Unpublished Ph. D. (Computer Science) Thesis, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

[9] Ramesha, K., Raja, K.B., Venugopal, K.R. and Patnaik, L.M. (2010), "Feature Extraction based Face Recognition, Gender and Age Classification" International Journal on Computer Science and Engineering, Vol. 2, No.1, pp. 14 – 23.

[10] Yang, Z.G. and Ai, H.Z. (2007), "Demographic Classification with Local Binary Pattern", pp. 464 – 473.

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

3D HMM-based Facial Expression Recognition using Histogram of Oriented Optical Flow

[1]Sheng H. Kung, [2]Mohamed A. Zohdy and [3]Djamel Bouchaffra
[1,2]Electrical and Computer Engineering, Oakland University, Rochester, MI, USA;
[3]Center for Development of Advanced Technology (CDTA), Baba-Hassen, Algiers, Algeria
shkung@gmail.com; zohdyma@oakland.edu; dbouchaffra@ieee.org

ABSTRACT

In this paper, we propose a 3D HMM (Three-dimensional Hidden Markov Models) approach to recognizing human facial expressions and associated emotions. Human emotion is usually classified by psychologists into six categories: Happiness, Sadness, Anger, Fear, Disgust and Surprise. Further, psychologists categorize facial movements based on the muscles that produce those movements using a Facial Action Coding System (FACS). We look beyond pure muscle movements and investigate facial features – brow, mouth, nose, eye height and facial shape – as a means of determining associated emotions. Histogram of Optical Flow is used as the descriptor for extracting and describing the key features, while training and testing are performed on 3D Hidden Markov Models. Experiments on datasets show our approach is promising and robust.

Keywords: Human Computer Interaction (HCI); Facial Expression Recognition; Feature Extraction, Optical Flow; Hidden Markov Model; HMM; Three -dimensional Hidden Markov Model; 3D HMM.

1 Introduction

Human Computer Interaction (HCI) is a fast-growing field propelled by advances in computer technology and machine learning theory. It is the study of how we communicate with (or through) machines, be it computers, robots, vehicles, home devices, the internet or artificial limbs. At a physical level, HCI has progressed from a low level of interaction (keyboard, mouse and pen) to a high level one (touch screen, sensor, voice, video and nerve). Beyond the mere physical aspect, however, HCI has also advanced into higher cognitive and affective functions [18]. Looking towards the future, HCI will assuredly evolve one step further to a predictive level, bringing it ever closer to resembling a true human-to-human interaction – with machines mimicking basic human senses of sight, hearing, taste, smell and touch as well as human learning abilities and emotions. Facial expression recognition is a step toward understanding of human emotions.

In simple terms, HCI is about the design of an intelligent interface that eliminates the need for third-party involvement during operation of the machine and expands the capability of that interface. The influence of HCI is far-reaching, finding critical application in areas such as surveillance, biometric security, video games, assistive devices for people with disabilities, robot-assisted surgery, accidence-avoidance and driver-assistance in automobiles, as well as in brain-controlled prosthetic limbs that

DOI: 10.14738/tmlai.36.1661
Publication Date: 27th December, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1661

explore the neural-machine interface. Clearly, HCI can lead to vast improvement in the quality of life for humans. Facial expression and emotion recognition will raise the compassion level of HCI application.

Much like other pattern recognition problems, the key to determining how difficult facial expression recognition problems are is in the intra-class and inter-class features variations. Intra-class variations, such as age and facial paraphernalia (materials either added-on or attached to faces creating occlusion, such as hair, glasses, beards, moustaches, cosmetics) are not small; inter-class variations, such as appearance, ethnicity, culture background and gender are also quite subtle. This combination of variations will complicate facial expression recognition. In addition, factors such as illumination, pose, viewpoint, scale, shade and noise add even further complexity to the problem. Here, facial expression recognition is defined as using computers in an attempt to automatically identify human facial expressions and infer their underlying meaning (emotion or intention). Facial expression recognition generally consists of facial expression analysis, representation, classification and interpretation.

This research is focused on HCI in the recognition of human facial expression and emotion analysis. We propose a 3D Hidden Markov Model (HMM) approach to recognizing human facial expressions and associated emotions. To the best of our knowledge, this is the first application of 3D HMM and Histogram of Oriented Optical Flow to facial expression recognition. Psychologists classify human emotion as displayed through facial expression into six categories: Happiness, Sadness, Anger, Fear, Disgust and Surprise. They further categorize facial movements based on the muscles that produce those movements using a Facial Action Coding System (FACS). We look beyond muscle movement and investigate facial features – brow, mouth, nose, eye height and facial shape – as demonstrated by motion, and use a Histogram of Optical Flow as a descriptor for extracting and describing those features.

This paper is organized as follows: Section 2 reviews the related work, Section 3 reviews feature representation, Section 4 describes our methodology, based on optical flow and 3D HMM, Section 5 details experiments and analysis, and we conclude with a summary.

2 Related Works

Facial expression research traces back to 1872 when Darwin suggested that facial expressions were innate in his famous book, The Expression of the Emotions in Man and Animals. Over the years, there remains much debate among scientists and psychologist about whether facial expressions are nature or nurture. We think both factors have influences on human facial expressions. There are six universal basic emotional states (e.g., happy, angry) with corresponding facial expressions varying in intensity from culture to culture (i.e. Asian cultures usually suppress facial expressions). There are other built-up and social activity-derived emotional states (e.g. shame, anxiety and embarrassment) whose corresponding facial expressions are heavily influenced by culture and the surrounding environment. These cultural contexts make facial expression recognition even more interesting and challenging.

Not until 1998 when Paul Ekman and Wallace Friesen adopted a system from a Swedish anatomist and published Facial Action Coding System (FACS) [9] did we have a method for measuring and scoring facial behavior. FACS identifies how various facial muscle movements made individually or in groups, affect facial expression. The movements in face and the one or more muscles that cause these movements are described and coded in Action Units (AU). For example, AU 1 is “Raising Inner Brow” and AU 26 is “Dropping the Jaw”, each with its own designated muscle(s) names. AUs are divided between the upper and lower face and also include some non-facial muscle movements concerning the eye and tongue.

AUs can be additive, if they are independent to one other, and ideally all facial expressions can be decomposed into their constituent AUs. FACS quickly became the de facto standard for characterizing facial expressions. Ekman and Friesen later also developed EMFACS (Emotional Facial Action Coding System) and FACSAID (Facial Action Coding System Affect Interpretation Dictionary) to interpret FACS scores in terms of emotion categories. Almost in parallel, Facial Animation Parameters (FAP) was developed in 1998 by Moving Pictures Experts Group (MPEG) as a facial animation specification in the MPEG-4 international standard that provided an alternative way of modeling facial expressions. FAPs – much like AUs – are closely related to muscle actions. There are 68 FAPs (e.g. FAP 3 for "open_jaw", FAP 5 for "raise_b_midlip and FAP 7 for "stretch_t_cornerlip").

There are shortcomings in FACS, such as ambiguity and subjectivity in intensity between AUs, to the complexity of representing emotions through multiple AUs (forest/tree syndrome) and reliability issues (agreement between observers). But FACS' advantage lies in the fact that "a human rater can encode facial actions without necessarily inferring the emotional state of a subject, and therefore one can encode ambiguous and subtle facial expressions that are not categorizable into one of the universal emotions, such as fake smiles" [12]. Researchers including Tian et al. [37], Valstar et al. [41] and Mahoor et al. [24] have had success in using computers to overcome shortcomings in FACS to automatically identify AUs and thus quickly identify emotions.

Ekman and Friesen [9] in essence provided us the framework and the means to analyze facial expressions anatomically.

3 Feature Representation

3.1 Preprocessing

Feature extraction is a starting and important task in any recognition process [44], whether it is a single image-based object, scene, face or facial expression or a video (frame) based facial or human action recognition. It consists of extracting the object or motion cues that are salient and discriminative among the recognized objects or actions. A good feature should be invariant to illumination, occlusion, scale, viewpoint, deformation, and clutter background, as well as affine, rotational and translational transformation of images. In order to reduce sensor noise, the image is usually smoothed by convolving the image with a Gaussian kernel. Illumination variations can be effectively minimized in preprocessing by normalizing the image using a filter; for example, histogram equalization or a discrete cosine transform (DCT) after truncating an appropriate number of coefficients to minimize variations under different lighting conditions.

3.2 Global Feature Extraction

Global feature is also called holistic feature in contrast to component-based local feature. It is a feature which represents image as a whole, for example, the shape, geometric dimension, distance, color etc. Global Feature has the advantage of being simple and low computational complexity. However, it has disadvantage of being generally low accuracy. Depending on the application, a fusion strategy to combine global and local feature may be necessary.

3.2.1 Model-based Approach - Decomposition

One of the model-based approaches is to find a model of bases for images of variable objects and a set of model parameters for each image. It involves conducting Principal Component Analysis (PCA), or, recently, Independent Component Analysis (ICA) to decompose images into a small set of characteristic feature images. It was originated in face recognition [14]. A face image is treated in PCA as a point (or vector) in a very high dimensional space (256x256 = 65,536). Eigenfaces, developed by Sirovich and Kirby [34] and applied to face recognition by Turk and Pentland [39], are eigenvectors derived from PCA and act as a set of orthogonal face basis usually represented in lower dimension. An individual face is then a weighted sum of the eigenfaces by projecting it into the space of face basis. To recognize a face is to compare weights of known individuals. It is generally an approach of nonlinear, generative and parametric model. Since it is based on whole image pixel intensities, the first three eigenfaces are usually discarded to reduce the illumination effect on image variations.

Eigenfaces have shortcomings in robustness to shape, pose and expression variations. To overcome them, Active Appearance Model (AAM), first introduced by Edwards, Taylor and Cootes [7], combined PCA shape model from landmark points and PCA grey-level appearance model from texture (pixel intensities). It can generate almost any face. Matching a face image is to find model parameters to minimize the difference between the image and a synthesized face – an optimization problem similar to template matching.

The basis vectors that ICA produces are more spatially localized and statistically independent than the ones in PCA. The figure below shows the first 8 eigenvectors computed on 500 randomly selected images from FERET gallery (top) and 8 of 200 ICA basis vectors computed (bottom). [9] Uddin et al. [40] proposed an enhanced ICA for facial expression recognition.

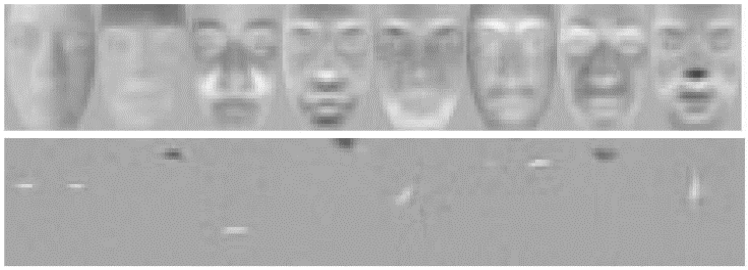


Figure 1. (Top) The first 8 eigenvectors computed on 500 randomly selected images from FERET gallery (Bottom) The 8 of 200 ICA basis vectors computed. (Draper et al. [9])

3.2.2 Model-based Approach - Transformation

Another model-based approach is to transform the image into bases of completely different domain - frequency domain. The most popular transformations are Fourier Transform, its related and faster DCT (Discrete Cosine Transform), and DWT (Discrete Wavelet Transform), all in 1D to 3D. The drawback of Fourier Transform is that it assumes the signal or image frequency is smooth across time or space (i.e., stationary) and, in transforming into the frequency domain the temporal or spatial information - when or where a discontinuous event occurs (frequency pattern changes) - is lost. In an effort to correct this deficiency, STFT (Short-Time Fourier Transform) or Gabor Transform uses a fixed sliding window to localize in time or space some frequency changes. However the dilemma is determining the size of the

window – a narrow window resulting in poor frequency resolution and a wide window resulting in poor time/space resolution. Wavelet overcomes the resolution dilemma by trying different frequencies at different windows. Unlike sinusoidal bases of Fourier Transform, Wavelet Transform uses wavelets as bases, which are irregular and asymmetric and, thereby, can very well model the signal or image with bumps, dips and humps. Features are in terms of the coefficients from transform. Pinto et al. [28] used 3D and 3D Wavelet Transforms to extract 3D facial expression features. Hough Transform is to map imperfect geometric shape into parameters space, e.g. a line is mapped to its slope and interception space as a point. After all the feature points are mapped into a space called accumulator, finding the maxima (vote) will determine the original image shape.

Tsai et al. [38] combines Harr wavelet and PCA for human-robot interactive emotion recognition. Global feature extraction is to extract global variation pattern for classification. It could also be used on sub-images for regional feature extraction. Following shows an example of log-Gabor responses extracted from a normalized expressive face. [11]

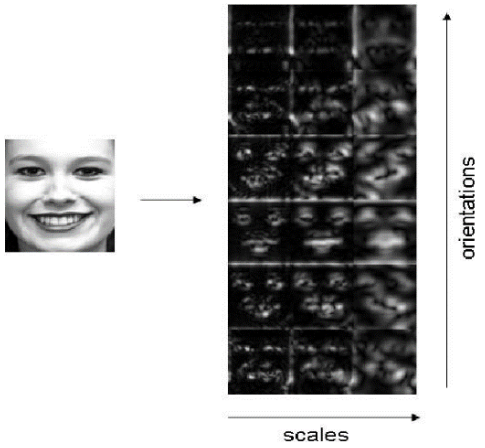


Figure 2. An example of log-Gabor responses (by scales and orientations) extracted from a normalized expressive face on the left. (Fanelli et al. [11])

3.2.3 Template-based Approach

Template is mainly used to detect the components of objects, for example, parts of the body, components of the face - eyes, nose and mouth regions. The template of face is generated by cropping all the facial regions in the training set from the ground truth data and averaging them. It is used to match the facial regions of testing objects in order to detect faces. The template can also be developed by describing the contour or shape of the objects for 2D correlation or distance matchings. The template can be algorithmically deformed to achieve efficient recognition of objects with shape deformation, occlusion and background clutter.

3.3 Local Feature Extraction

Local feature extraction consists of the design of feature detector and feature descriptor on image space. Image space can be spatial (x, y) for single image and spatial temporal (x, y, t) for video. Detector refers to detection of interest points or keypoints in image space for subsequent processing of descriptor. The interest point needs to be clearly defined mathematically in image space, reproducible and rich in

salient information. Given the interest point found at a location, descriptor is to describe the image structure in a neighborhood of that location. It needs to be strongly inter-class discriminative, while tolerant of intra-class variation and robust to various image variations.

3.3.1 Detector

Gradient Detector

Gradient is the first order derivative detector. Operators like Prewitt, Sobel, Robinson and Kirsch are all first order derivative filter for detecting edges. Harris corner uses covariance matrix of gradients to determine the corner.

Laplacian Detector

The Laplacian $L(x,y)$ of an image with pixel intensity values $I(x,y)$ is given by

$$L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} .$$

It is a second order derivative mask, which is also called "zero crossing detector" due to its highlighting of zero passing edges. It is combined with Gaussian smoothing filter of (0, σ) to becomes a Laplacian of Gaussian (LoG) detector. LoG can be approximated by A Difference of two Gaussians (DoG) having different standard deviations. DoG works like a bandpass filter by subtracting one blurred image from another to preserve the spatial information corresponding to some range of frequencies while discarding all others. The extrema (Maxima/Minima) of the Laplacian detector was used by SIFT (Scale-Invariant Feature Transform) to detect interest points in a local patch. However, there are following steps to filter out weak and unstable keypoints, which are low contrast and along the edge.

Hessian Detector

Hessian is also a second order derivative detector. However, it is in a matrix form. It is a square matrix of second-order partial derivatives of a scalar-valued function. The 2X2 Hessian Matrix can be expressed as

$$H(x,y) = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix}$$ for an image $I(x,y)$.

It describes the local curvature of a function of many variables. The eigenvectors of the matrix give the directions for minimum and maximum curvature while the eigenvalues correspond to the amount of curvature in those directions. SURF (Speed-up Robust Features) [2] use extrema of the determinant of Hessian matrix in a local neighborhood of image to detect interest points. SURF is more efficient than detector and descriptor of SIFT due to the advantage of speed and accuracy of Hessian matrix. Wu et al. compared SIFT with its variants including SURF [42].

Blob Detector

Maximally Stable Extremal Regions (MSER) [8] detect blobs in images, which are robust under perspective transformations. It explores distinguished regions [25], regions that possess some distinguishing, invariant and stable property. It is also called extremal regions because all pixels inside have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels

on its outer boundary. MSER is based on the idea that regions, which stay nearly the same through a wide range of thresholds, must be maximally stable. It is used in face detection and tracking.

Combined Detector

The Harris-Laplace detector combines the traditional 2D Harris corner detector with the idea of a Gaussian scale-space representation in order to create a scale-invariant detector. Harris-Affine detector combined the Harris corner detector with the idea of iteratively applying affine shape adaption algorithm in order to create an affine-invariant detector. Hessian-Laplace and Hessian-Affine detectors are the same. In order to create a scale-invariant detector, a Gaussian pyramid is constructed in SIFT and SURF from the input image by repeated smoothing and subsampling, then a DoG (for SIFT) or Hessian matrix (for SURF) is computed from the differences between the adjacent levels in the Gaussian pyramid.

3.3.2 Descriptor

Local Binary Patterns (LBP)

LBP is a non-parametric local descriptor to capture the texture pattern of an image in a small neighborhood around a pixel. It labels the surrounding pixels with binary 1 and 0 depending on whether the intensity of the pixel is greater than the central pixel. Due to its discriminative power and computational simplicity, LBP has become a popular approach in various applications including facial expression analysis. It could be quantized and grouped into local histograms for classification. Shan et al. [33] proposed a Boosted-LBP for facial expression recognition. Below shows a basic LBP operator.

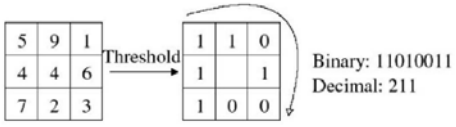


Figure 3. A basic LBP operator. (Shan et al. [33])

Histograms

Histogram of Oriented Gradients (HOG) is used by SIFT [22] for descriptor of the interest points. After SIFT keypoints are detected, a 16x16 neighborhood around the keypoint is taken. As the following diagram shows, it is devided into 16 sub-blocks of 4x4 size. For each sub-block, 8 bin orientation histogram is created. Together all 16 sub-blocks of 8-bin orientation histogram are concatenated to obtain 128 (16*8) dimensional feature vector – the keypoint descriptor. In addition to this, several measures are taken to achieve robustness against illumination changes, rotation etc. The advantage of quantization of gradient locations and orientations is to make the descriptor robust to small geometric distortions and small errors in the region detection.

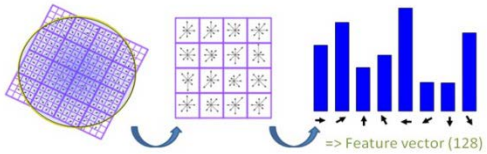


Figure 4. A SIFT descriptor generation process diagram.
(Source: https://gilscvblog.wordpress.com/2013/08/18/a-short-introduction-to-descriptors/)

Soyel et al. [35] proposed a discriminative SIFT (D-SIFT) to match the SIFT keypoints detected from two facial expression images as shown below. On the left are matches from the regular SIFT and on the right from D-SIFT.

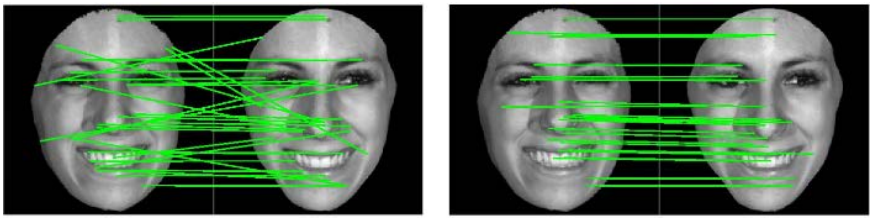


Figure 5. (Left) Face descriptors detected and matched by standard SIFT algorithm. (Right) Face descriptors obtained by KLD-based matching under varying expression intensities. (Soyel et al. [35])

Spatial-frequency

Gabor filter is defined as the product of a Gaussian kernel and a complex sinusoid. It receives attention because the frequency and orientation representations of the Gabor filter are similar to those of mammal's visual cortex. Gabor filter is often used as a filter bank because it has various combination of parameters - θ represents the orientation, λ is the wavelength, and σ1 and σ2 represent scale at orthogonal directions (or σ if the Gaussian is symmetric) [26]. When it is applied the output is equal in size with the original image, and the features extracted have great redundancy. Therefore, dimension reduction is needed. Praseeda Lekshmi et al. [29] and Bashyal et al. [1] used Gabor filter in facial expression recognition.

Gabor wavelets are related to Gabor filter. The family of Gabor wavelets are created by dilation (scale) and shift from the mother wavelet. When Gabor wavelet transformation is applied, Gabor wavelet coefficients are output as features for the neighborhood pixels of the interest point. It is much more powerful than geometric positions. Tian et al. [36] used Gabor Wavelets in facial expression recognition.

SURF [2] calculates Haar wavelet response for descriptor. A 20X20 region is split up into smaller 4×4 square sub-regions. As the following picture shows, for each sub-region, SURF compute a few simple features at 5×5 regularly spaced sample points. The horizontal and vertical Haar wavelet responses dx and dy are calculated and summed. The absolute values of the responses $|dx|$ and $|dy|$ are also calculated and summed for each sub-region. For all 4×4 sub-regions, it forms a vector of length 64 as keypoint descriptor.

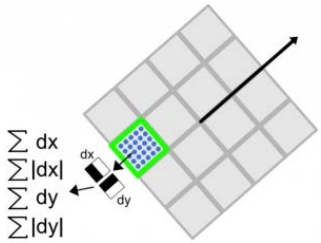


Figure 6. A SURF descriptor generation process diagram.
(Source: http:/www.juergenwiki.de/work/wiki/doku.php?id=public:surf)

4 Methodology

The proposed framework and workflow is shown below.

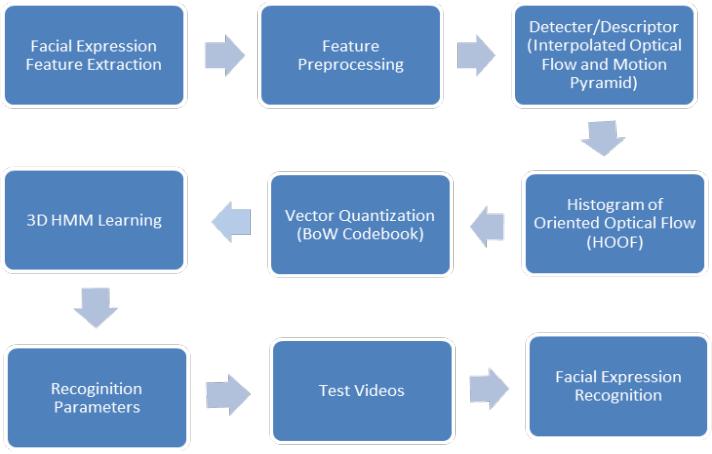


Figure 7. A system flow diagram.

4.1 Optical Flow

Optical flow is the motion estimate of image pixels or regions from one frame to the next. Between two consecutive frames, the following was assumed: 1) Brightness/color constancy between the same pixels, 2) Small displacement (However, image pyramids are to relax this assumption and track larger movements) and 3) Spatial coherence (neighboring points are on the same surface). To cope with the aperture problem, additional assumptions were made in the following Lucas-Kanade method [23] that all neighboring pixels in the patch move at the same speed and in the Horn-Schunck method [13], that a smoothing term was added in the optimization. However, motion discontinuities (e.g. occlusion at motion boundaries, pixels being visible in one frame only) and motion in texture-less regions would make optical flow estimation a challenging task.

4.1.1 Lucas–Kanade Method (Local)

First, Identify distinguished points (e.g., corners detected using Shi-Tomasi algorithm), then obtain image patches surrounding those pixel points, and use a least square optimum fit model to find the flow field between two frames. It is a local estimating method resulting in a sparse flow because it cannot provide flow information in the interior uniform regions of the image. Local method is more robust under noise. First picture below shows an optical flow tracking of the movement of highway vehicles. The second picture shows optical flow of facial expression.

Figure 8. An example of feature-points tracking.
(Source: http://docs.opencv.org/master/d7/d8b/tutorial_py_lucas_kanade.html#gsc.tab=0)

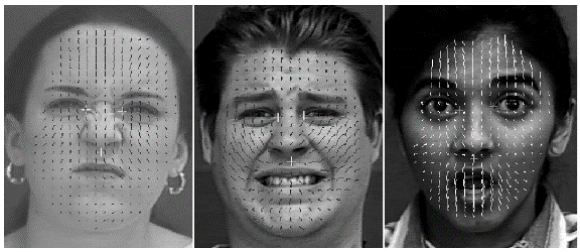


Figure 9. Examples of pixel-wise tracking of face. (Lien et al. [20,21])

4.1.2 Horn–Schunck Method (Global)

Horn-Schunck method is based on Lucas-Kanade method but add a smoothing constraint and integrate over the whole image. Because it is optimizing an integral function of the whole image based on residuals from the brightness constancy constraint, and a specific regularization term expressing the expected smoothness of the flow field, it is a global estimating method. It is resulting in a dense flow because it processes all the pixels. [13]

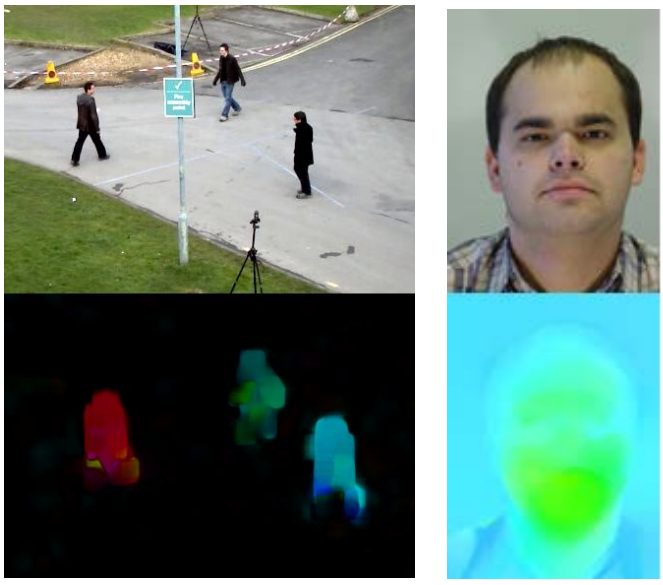


Figure 10. (Left) An example of dense optical flow, (Right) An example of facial optical flow from a pair of head movement frames (Source on the left is same as Figure 8.)

4.1.3 Coarse to Fine Lukas-Kanade Method

Horn-Schunck and Lucas-Kanade optical methods work only for small motion. If object moves faster, the brightness changes rapidly, derivative masks will fail to estimate spatiotemporal derivatives. Therefore, image pyramid is used to computer optical flows at coarser scale as the following diagram indicates. It is also called Multiresolution L-K method or L-K with Pyramid method.

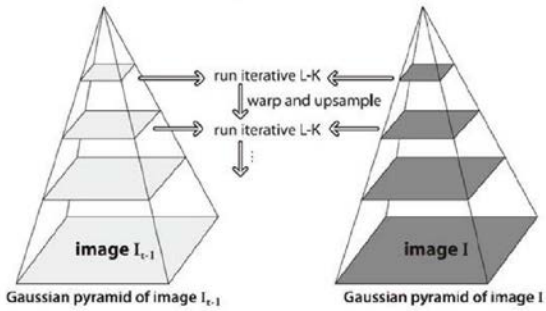


Figure 11. A pictorial coarse-to-fine optimal flow estimation.
(Source: Bradski et al. 2008. *Learning OpenCV: Computer vision with the OpenCV library,* O'Reilly)

4.1.4 Combined Local-Global (CLG) Methods

Bruhn et al. [4] proposed a method combining the robustness of local methods with the density of global approaches. Hence, it is called combined local and global method. It can be considered as a noise-robust generalization of the Horn and Schunck technique because it uses the same concept of integration concept of H-S method. Instead of integrating over the individual pixels, it integrates over the patch of pixels which L-K method users to avoid the aperture problem.

4.1.5 A Proposed Interpolated Lucas-Kanade Method

We proposed an interpolated Lucas-Kanade optical flow method to transform sparse Lucas-Kanade Coarse-to-Fine optical flows to dense flows by adopting a distance-based interpolation method [19]. As the following left figure shows, we use a search circle, whose radius need to be defined beforehand or to be varied, centered at the interpolation point at location x to select data points p_i enclosed in the circle. The weight assigned to each is based on the square of distance from x to p_i. There are other weights can be used. Bigger circle will make the resulting surface smoother but loose the local continuity. Therefore, a good knowledge of the data set will be required to select the radius of the circle.

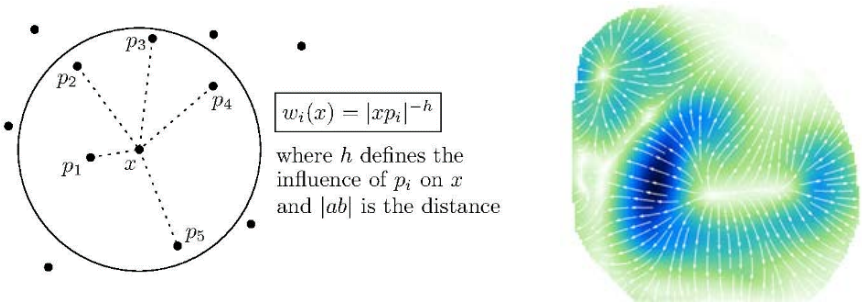


Figure 12. (Left) A pictorial distance-based interpolation circle (Ledoux et al. [19]). (Right) An example of Interpolated L-K optical flows

4.2 HMM

HMM is a tool to model a spatial, temporal or spatio-temporal process when the process is sequential and stochastic in nature. For example, Huang and Kennedy used HMM to uncover hidden spatial states and patterns underlying home sales prices. HMM is also widely used in speech and handwriting recognition, DNA sequencing, and music identification to characterize the spectral properties in time series and in dynamic scene understanding and behavior recognition in video footage to uncover spatio-temporal (space and time) patterns in events.

HMM is a generative graphical model, as well as a simple Dynamic Bayesian Network (DBN). It is based on three assumptions: (1) The first-order Markov chain assumption on the transition probability among hidden states - no influence on current state (t) beyond its predecessor in time ($t-1$), (2) The stationary assumption on transition probabilities – the transition probabilities are time-invariant, and (3) output independence assumption on the probabilities emitted from hidden states to observations - the emission probabilities are time-invariant. Researchers are developing variations of HMM in relaxing these assumptions like the defined state duration or varying the underlying structure or architecture of HMM to cope with the needs of various applications. Some common variants of HMM are continuous HMM with Gaussian or Mixture of Gaussian output, Input-output HMM, Coupled HMM, Factorial HMM, Layered HMM and Hierarchical HMM [3]. This paper is focused on 3D HMM.

HMM is using the observations to find the hidden states following the Bayesian theory of calculating the posterior probability from a prior probability and its likelihood. The model is also in line with the control theory view of the real-world system consisting of state space (unobservable) and output vector (observable). For example, in a stochastic, discrete-time, linear control system represented by

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}$$

$$z_k = Hx_k + v_k$$

where A corresponds to a state-transition matrix, H corresponds to a emission (observation) matrix and B, the input-output matrix, are coefficients, x is the state variable, u is the input variable, z is the output variable, w_k (process noise) and v_k (measurement error) are multivariate Gaussian distributions with covariance q_k, and is called Kalman Filter. The underlying model of Kalman Filter is a Bayesian model similar to HMM, but latent and observed variables are continuous with Gaussian distributions.

Due to the linearity and Gaussian distribution assumptions, the maximum likelihood estimate (MLE) used in Kalman Filter coincides with maximum a posteriori probability (MAP) estimate from expectation maximization (EM) algorithm of HMM.

Here is a summary of HMM algorithm:

Rabiner's* [30] *three basic problems

- What is the probability of an observed sequence, *O*, given a model? **(Evaluation)**
  - *joint probability of observations and state sequence – P(O,S|ϑ)*
  - *as useful as Markov Chain*
- What is the optimal sequence of states that "explains" the observed data? **(Decoding)**
  - *optimality criterion*
- how can one adjust the model parameters to maximize the probability of the observed data given the model **(Learning)**

Problem 1: Evaluation

For a model $\lambda(\theta)$ with parameters θ, what is $P(O|\lambda)$ or more simply $P(O|\theta)$?

- model parameters
  Q a set of states $Q = \{1,2,\ldots,N\}$
  A the state transition matrix $a_{ij} = P(q_{t+1} = j | q_t = i)$
  B the emission probabilities $b_j(k) = P(o_t = k | q_t = j) \quad 1 \le k \le M$
  ω an initial probability distribution $\omega_i = P(q_0 = i) \quad 1 \le i \le N$
  Observations $O = (o_0, o_1, \ldots o_T)$
- $P(O|\theta) = \sum_{\pi} P(O|\pi,\theta) P(\pi|\theta)$
- Assume the observations are independent

$$P(O|\pi,\theta) = \prod_{i=0}^{T} P(o_t | q_t,\theta)$$
$$= b_{q0}(o_0) b_{q2}(o_2) \cdots b_{qT}(o_T)$$

  The probability of a a particular state sequence (or path), π,
  $P(\pi|\theta) = \omega_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}$
- There are N^T state sequences and *O(T)* calculations so the brute force complexity is $O(TN^T)$
- *Forward algorithm (Viterbi)*
  - α_t *is the probability of observing the partial sequence* $(o_0, o_1, \ldots, o_t)$ *given that state* $q_t = i$
  - $\alpha_t(i) = P(o_0, o_1, \ldots, o_t, q_t = i|\theta)$
    $\alpha_{t+1} = (\sum_{i=1}^{N} \alpha_t(i) a_{ij}) b_j(o_{t+1})$ *with* $\alpha_0(i) = \omega_i b_i(o_0)$
  - $P(O|\theta) = \sum_{i=1}^{N} \alpha_T(i)$
  - *complexity* $O(N^2 T)$
- *Backward algorithm*
  - almost the same as forward algorithm

- $\beta_t(i)$ is the probability of observing the partial sequence $(o_{t+1}, o_{t+2}, \dots o_T)$ given that state $q_t = i$

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots o_T | q_t = i, \theta) \text{ with initial condition } \beta_T(i) = 1$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad t = T-1, \dots 0$$

$$\alpha_{t+1} = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right) b_j(o_{t+1})$$

$$\alpha_t(i) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\, b_j(o_t)$$

Problem 2: Decoding

What is the optimal sequence of states that "explains" the observed data?

- Optimality criteria
  - *the path π that maximizes the correct number of individual states, i.e., the path where the states are individually most likely*
  - *the most probable single path, maximize* $P(\pi | O, \theta)$ *or equivalently* $P(\pi, O|\theta)$ *— Viterbi algorithm*

- The optimal path is the path that maximizes $P(q_0, q_1, \dots, q_T | O, \theta)$
  let $\delta_t(i)$ be the highest probability path ending in state i

$$\delta_t(i) = \max_{q_0, q_1, \dots q_{t-1}} P(q_0, q_1, \dots, q_t = i, o_0, o_1, \dots, o_t | \theta)$$

$$\delta_t(j) = \max_i \left[\delta_t(i) a_{ij} \right] b_j(o_{t+1})$$

- Keep track of argument that maximizes $\delta_t(j)$ at each position *t*

Problem 3: Learning

- Find the parameters, $\hat{\theta}$, that maximizes $P(O|\theta)$

$$\theta = \arg\max_{\theta} \left(P(O|\theta) \right)$$

- No analytical solution requires iterative solution (Baum-Welch algorithm)
  1. *initial model* θ_0*, repeat*
  2. *compute parameters* θ_i *based on* θ_{i-1} *and observations O*
  3. *if* $\log P(O|\theta_i) - \log P(O|\theta_{i-1}) < \varepsilon$*, stop*
     else, accept θ_i*, goto 2.*
- With Baum-Welch algorithm likelihood is proven to be greater or equal at each step

Training

- Need to update the transition probabilities. The probability of being in state *i* at time *t* and state *j* at time *t+1* is $\xi(i, j)$

$$\xi(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\theta)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

- The probability of being in state *i* at *time t, given the observed sequence O is* $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$ or in terms of ξ, $\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j)$

- Derived quantities
  expected number of times state i is used $\sum_{t=0}^{T}\gamma_y(i)$
  expected number of transitions from state i to state *j* $\sum_{t=0}^{T-1}\xi_t(i,j)$

- Baum-Welch parameter updates
  - $\omega_i = \gamma_0(i)$ *probability of starting in state I*
  - $a_{ij} = \frac{\sum_{t=0}^{T}\xi_t(i,j)}{\sum_{t=0}^{T}\gamma_t(i)}$
  - $b_j(k) = \frac{\sum_{t=0}^{T}\delta_{o_t}\gamma_t(j)}{\sum_{t=0}^{T}\gamma_t(j)}$ *where* $\delta_{o_t} = \begin{cases}1, & \text{if } o_t = k\\ 0, & \text{if } o_t \neq k\end{cases}$

4.2.1 3D HMM

HMM is exploring a latent statistical model to describe the observed statistical phenomenon, assuming the observed variations of sequence of events were generated from an underlying inherent Markovian stochastic process. The model is used for recognition, classification and interpretation of future sequence of events.

1D HMM is applied to a one dimensional spatial or temporal chain, e.g., speech recognition, musical score analysis, and sequencing problems in bioinformatics. 3D HMM is applied to a two dimensional spaces, e.g., aerial image segmentation, automatic face recognition.

3D HMM is an extension of 3D HMM into a spatio-temporal volume in which consecutive frames of video images are stacked to form a third dimension - time.

Suppose there are M states $\{1,2,\ldots,M\}$ for each node (i,j,k) in a 3D structure, $i = \{1,2,\ldots,I\}$, $j = \{1,2,\ldots,J\}$, *and* $k = \{1,2,\ldots,K\}$ where I,J are the numbers of row and column of each frame and K is the number of frames in the original image, the feature vector (observation) is $o(i,j,k)$, the corresponding hidden state is $s(i,j,k)$, and the class of the node is $c(i,j,k)$. The transition probability of state $s(i,j,k)$ depends on its adjacent neighboring states in vertical, horizontal and across the frame directions following a predefined lexicographic order as

$$(i',j',k') < (i,j,k) \text{ if } k' < k \text{ or } k' = k, j' < j \text{ or } k' = k, j' = j, i' < i.$$

A graphical illustration for the 3D grid is depicted in Figure 13.

We can define the state transition probability as

$$P\{s_{i,j,k} = l \mid \omega\} = a_{p,m,n,l}$$

where $\omega = \{(s_{i'j'k'}, u_{i'j'k'}): (i', j', k') < (i, j, k)\}$ is the set of states and feature vectors of all points preceding (i, j, k) in the lexicographic order and their state values are p, m and n. where $p = s_{i,j,k-1}$, $m = s_{i-1,j,k}$ and $n = s_{i,j-1,k}$ [16].

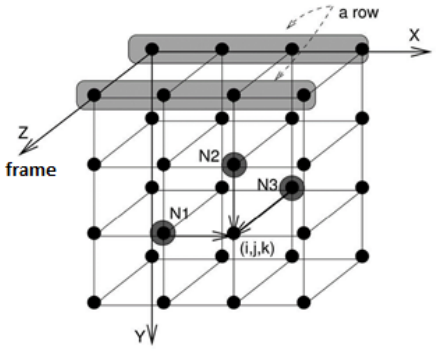


Figure 13. A pictorial 3D HMM structure (Joshi et al. [16]).

We can define the emission probability as normal density distribution. Given the state $s_{i,j,k}$ of a point (i, j, k), the feature vector $o_{i,j,k}$ follows a multivariate Gaussian distribution with a covariance matrix Σ_l and a mean vector μ_l determined by the state $l = \{1, \ldots, M\}$ as follows [16]:

$$b_l(o) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} e^{-(1/2)(o-\mu_l)' \Sigma_l^{-1}(o-\mu_l)}.$$

If the state $s_{i,j,k}$ of a point *(i, j, k)* is known, its observed vector $o_{i,j,k}$ is conditionally independent of the rest of the points in the 3D grid, we can define a Viterbi training algorithm for estimation of Parameters as follows.

In 1-D HMM a Baum-Welch algorithm, a special case of expectation-maximization algorithm is usually used to derive local maximum likelihood estimate of the parameters of HMM. An approximation to the Baum-Welch algorithm is the Viterbi training, in which each observation and its assigned parameters were used to get the most likely state sequence as shown in (1). Then the sequence was used to re-estimate the hidden parameters as shown in (2) – (4). Instead of maximizing the likelihood of observed data, the Viterbi training is to maximize the probability of the most likely sequence of states and ends up saving significant computational time by sacrificing some accuracy [16].

$$\{s^*_{i,j,k} : (i,j,k) \in C\} = \underset{s_{i,j,k}:(i,j,k)\in C}{\arg\max} P(s_{i,j,k} : (i,j,k) \in C | o_{i,j,k}(i,j,k) \in C; \lambda^{(t)})$$

$$= \underset{s_{i,j,k}:(i,j,k)\in C}{\arg\max} P(s_{i,j,k}, o_{i,j,k} : (i,j,k) \in C; \lambda^{(t)}) \quad (1)$$

$$\mu_\ell^{(t+1)} = \frac{\sum_{(i,j,k)\in C} o_{i,j,k} I(s^*_{i,j,k}=l)}{\sum_{(i,j,k)\in C} I(s^*_{i,j,k}=l)} \quad (2)$$

$$\sum_l^{(t+1)} = \frac{\sum_{(i,j,k)\varepsilon C}(o_{i,j,k}-o_l^{(t+1)})(o_{i,j,k}-o_l^{(t+1)})' I(s^*_{i,j,k}=l)}{\sum_{(i,j,k)\varepsilon C} I(s^*_{i,j,k}=l)} \quad (3)$$

$$a_{p,m,n,l}^{(t+1)} = \frac{\sum_{(i,j,k)\varepsilon C} I,(s^*_{i,j,k-1}=p) I(s^*_{i-1,j,k}=m) I(s^*_{i,j-1,k}=n) I(s^*_{i,j,k}=l)}{\sum_{(ij,k)\varepsilon C} I(s^*_{i,j,k-1}=p) I(s^*_{i-1,j,k}=m) I(s^*_{i,j-1,k}=n)} \quad (4)$$

Where $C = \{(i,j,k), 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K\}$, s* is the optimal state sequence, $I(\cdot)$ is the indicator function that equals 1 when the argument is true and 0 otherwise, λ is collection of parameters of 3D HMM [16].

In 1-D HMM Viterbi algorithm for finding the states $\{s_t^*\}$ requires $wM(M-1)$ comparisons and wM memory records (assuming the length of HMM is w and M possible states). Hence the computational and storage complexities are $O(wM^2)$ and $O(wM)$ respectively. For a 3D HMM there are M^{w^3} possible combinations of states for the entire 3D grid. When Viterbi algorithm is applied to frame-by-frame, it avoids exhaustive search along Z axis, but it still needs to consider all the possible combinations of states in each frame. Therefore, the computational complexity for the optimal set of states is at least $O(wM^{2w^2})$.

Joshi, et al. [16] proposed a locally optimal algorithm to search for the set of states with the MAP probability based on row sequence, then row by row adjustment. It improves the complexity of one iteration of the proposed algorithm to $O(w^3M^2)$ due to 3D grid total of w^2 rows.

Jiten and Merialdo [15] introduced a three dimensional dependency tree HMM (3D-DT HMM) which relaxes the dependencies between neighboring state nodes to a random uni-directional dependency. It allows us to estimate the model parameters along a 3D path linearly through a random decision tree, as the following figure shows. Hence, it maintains a linear computational complexity.

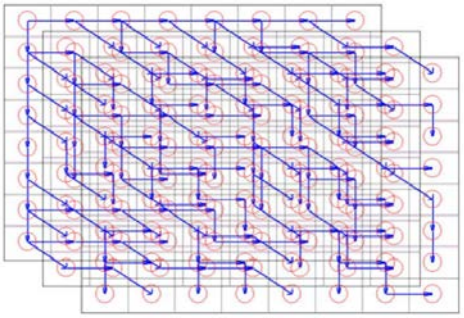


Figure 14. A pictorial 3D-DT (Dependency Tree) HMM structure (Jiten et al. [15]).

5 Experiments and Analysis

The experiments are performed as follows:

- First, select testing video dataset that has variations in the face profile such as sex, age and ethnicity with variety of partial occlusions on their face such as hair, glasses, headscarf etc. (see Figure 15)
- Split video into frames and run interpolated optical flow code to obtain optical flow vectors
- Select block size and run HOOF (Histogram of Oriented Optical Flow) code to obtain histogram for each block of optical flow vectors
- Run VQ (Vector Quantization) code to obtain BoW (bag of words) for all histograms
- Transfer all histogram data to data in terms of BoW and feed them to 3D HMM code for each facial expression to obtain parameters
- Run the parameters against the validation set.

5.1 Data Set

The most widely used database for research on facial expression recognition is the Cohn-Kanade facial expression database [17]. It contains image sequences of approximately 100 subjects, each posing a set of 23 facial displays. They were labeled with FACS codes in addition to basic emotions. There are two main limitations of this database [27]. First, the facial expression recording ends at the apex of the facial motion, instead of full cycle of onset, apex and offset in motion research. Secondly, many recordings have date/time stamp over the chin of the subject making the changes in the appearance of chin difficult to see and track.

MMI facial expression database was developed to overcome the limitations. It was conceived in 2002 by Maja Pantic, Michel Valstar and Ioannis Patras [27]. Initially it was focused on collections of high-quality still images of AUs and AU combinations. It video-recorded 1767 clips of 20 participants (Sessions 1 to 1767). Later data to distinguish the six basic emotions were added and formed the second part of the MMI database. It video-recorded 238 clips of 28 subjects displaying six basic emotions (Sessions 1767 to 2004). Part III is similar to part I, consisting of high quality still images with all AUs and six basic emotions (Sessions 2401-2884). Recently, spontaneous data were added to form Part IV (Sessions 2005 to 2388) and Part V (Sessions 2895 to 2903).

Part II of MMI database was used for this research. Six basic emotions of 20 randomly selected subjects were taken for the study.

Figure 15. Example of subjects in MMI facial expression database (size 720x576). (Source: http://mmifacedb.eu/)

5.2 Optical Flow Outputs

After extracting 20 videos of the six basic emotions (Happiness, Sadness, Anger, Fear, Disgust and Surprise), we split each video into 7 frames and feed the frames pairwise to optical flow program to obtain optical flow vectors. Figure 16 - 18 shows the results of the two emotions – Anger and happiness. The first row depict the extracted frames of videos, the second row exhibit the color-coded optical flow diagrams calculated from frame pairs. Figure 19 and 20 display the first six histograms of the optical flow magnitude and orientation from anger data set.

Figure 16. (Top) An example of 7 frames from Anger data set. (Bottom) The optical flows obtained.

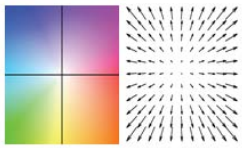


Figure 17. Optical flow color-code (hue for direction and saturation for magnitude).

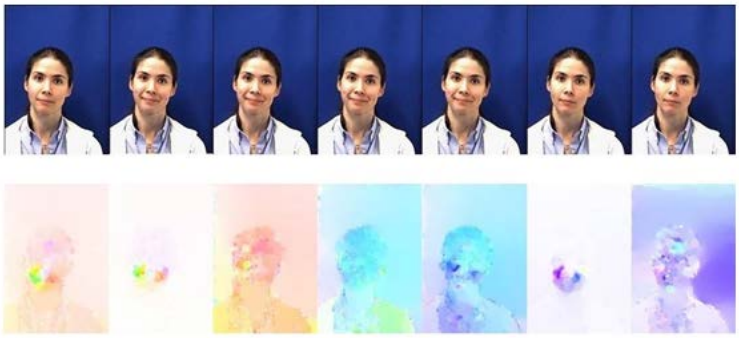


Figure 18. (Top) An example of 7 frames from Happiness data set. (Bottom) The optical flows.

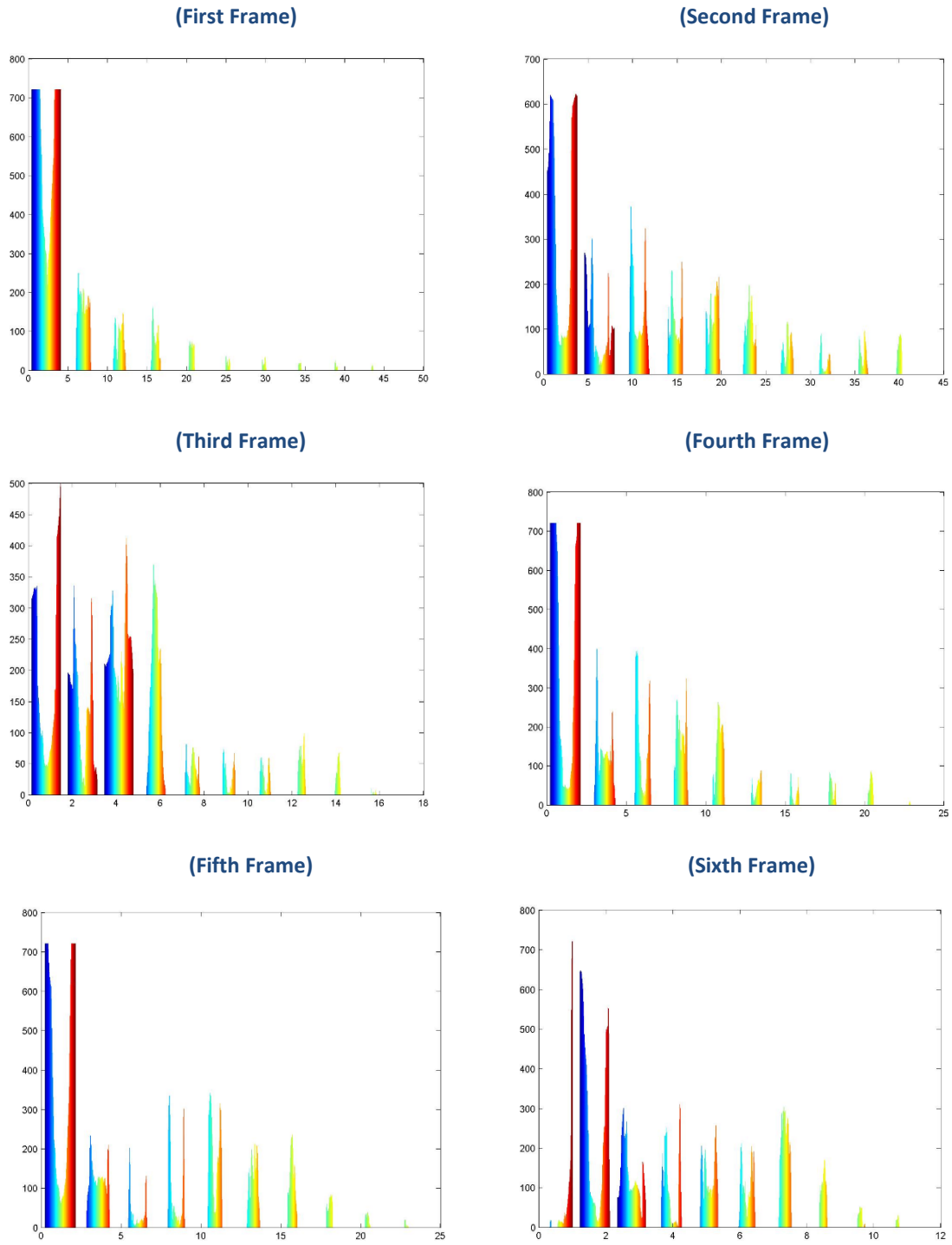


Figure 19. Histogram of magnitude of optical flow from Anger data set in Figure 16.

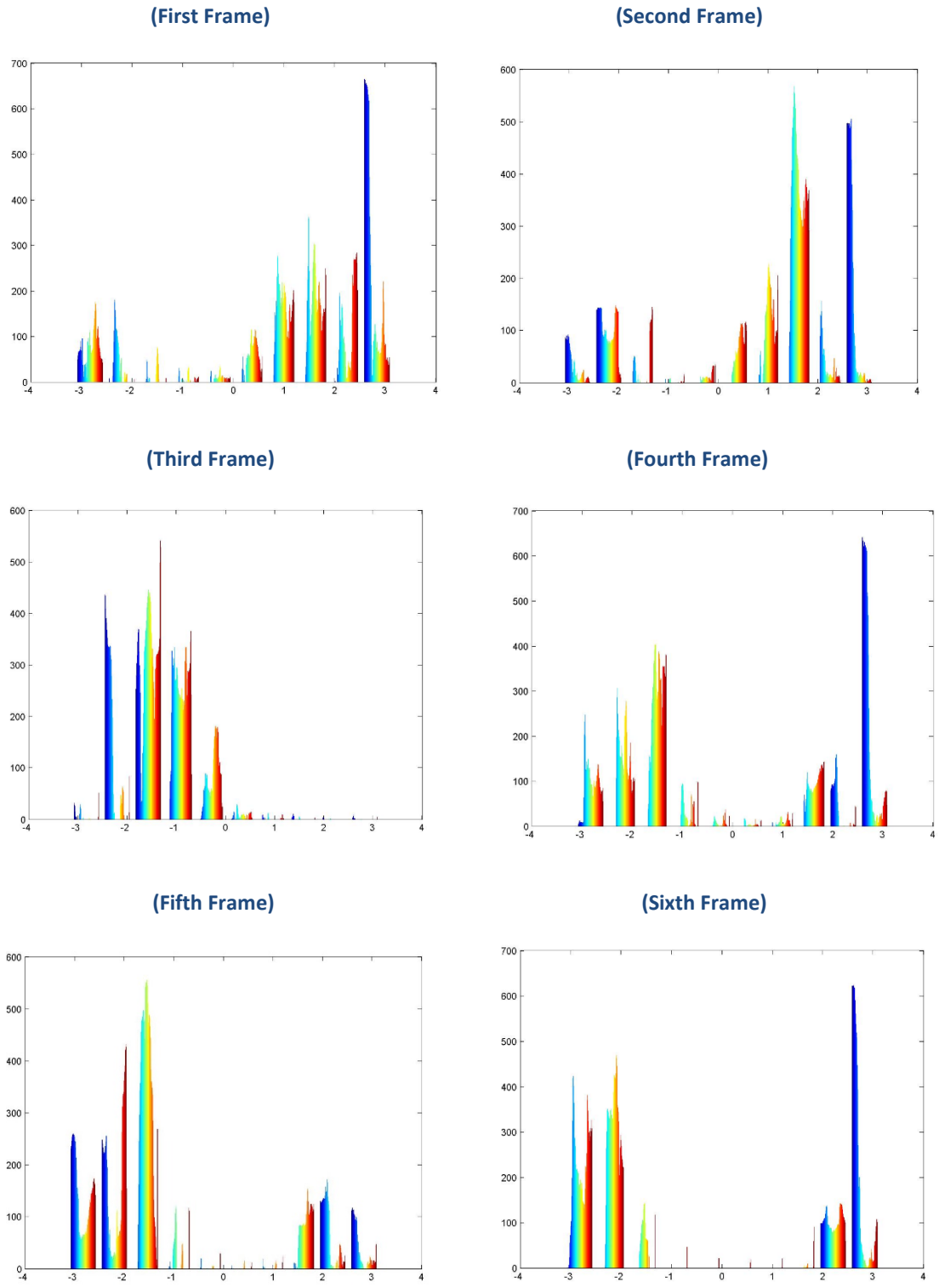


Figure 20. Histogram of orientation of optical flow from Anger data set in Figure 16.

5.3 Histogram of Oriented Optical Flow (HOOF) Outputs

We use HOOF algorithm developed by Chaudhry et al. [5]. As indicated earlier using histogram as the descriptor has advantage of being robust to small geometric distortions and small errors in the region detection. Raw optical flows may not be useful as a descriptor if the number of pixels is different in the face and the optical flow computation is prone to background noise, scale variation and direction of movement. No two faces are alike but faces are symmetric. The eyebrow raise on the right half of face is same as the eyebrow raise on the left half of face. The lip stretch to the right is same as to the left with only difference of direction. We need a descriptor from optical flow but is invariance to scale and direction of motion.

First, we computer optical flows at every frame of the video. Each flow vector is binned according to its primary angle from the horizontal axis and it is weighted according to its magnitude. Thus, all optical flow vectors, $f = [u, v]^T$ with direction of $\theta = \tan^{-1}\left(\frac{v}{u}\right)$ in the range

$$-\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{B}$$

will have $\sqrt{u^2 + v^2}$ added to the sum in bin b $(1 \leq b \leq B)$. There is a total of B bins. B can be varied (Figure 18 shows B = 4). Finally the histogram is normalized to a sum of 1.

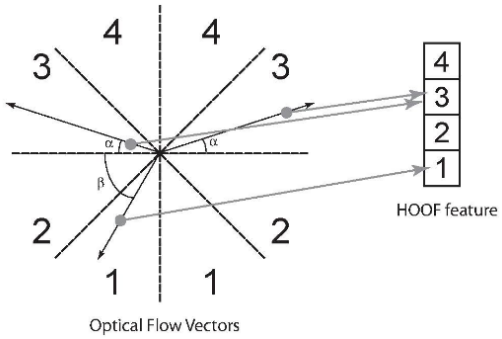
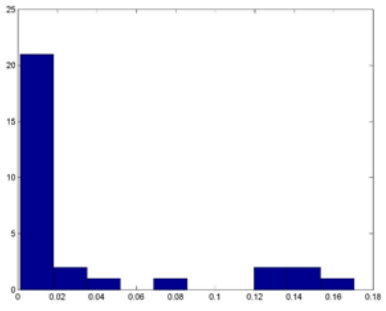
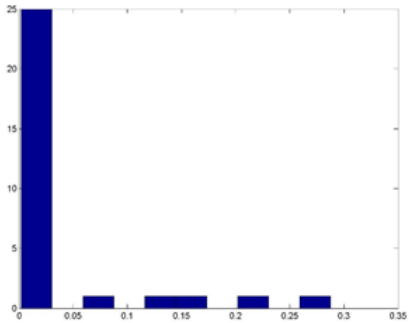


Figure 21. An example of histogram formation with four bins (Chaudhry et al. [5])

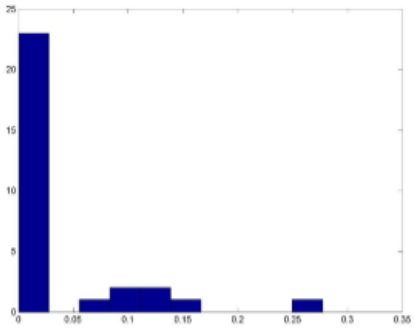
The following figure shows the histogram of the whole image.



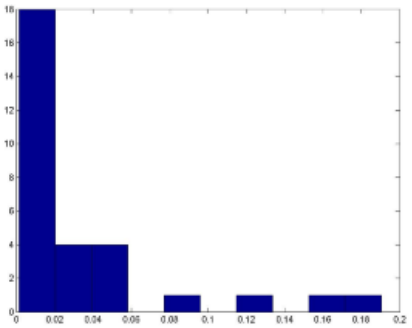
(First Frame)



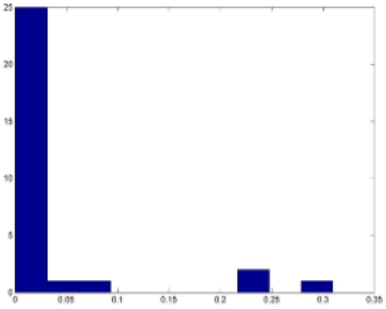
(Second Frame)



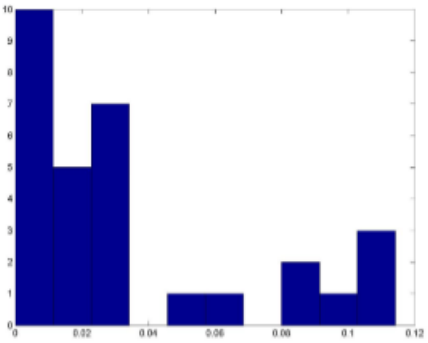
(Third Frame)



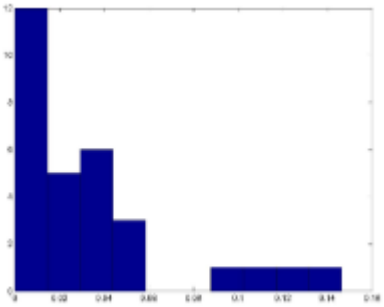
(Fourth Frame)



(Fifth Frame)



(Sixth Frame)



(Seventh Frame)

Figure 22. Histogram of Oriented Optical Flow from whole image of Anger Data set in Figure 16.

However, in our application we divided the whole image by blocks of size 24x24. There is 720 blocks (30x24) of images for each frame and therefore 720 HOOF histograms for each frame.

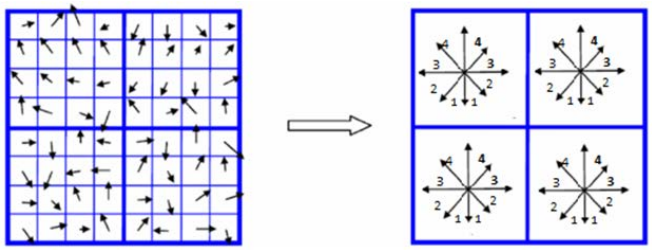


Figure 23. HOOF generated from a 4x4 block of image (our experiment uses blocks of 24x24).

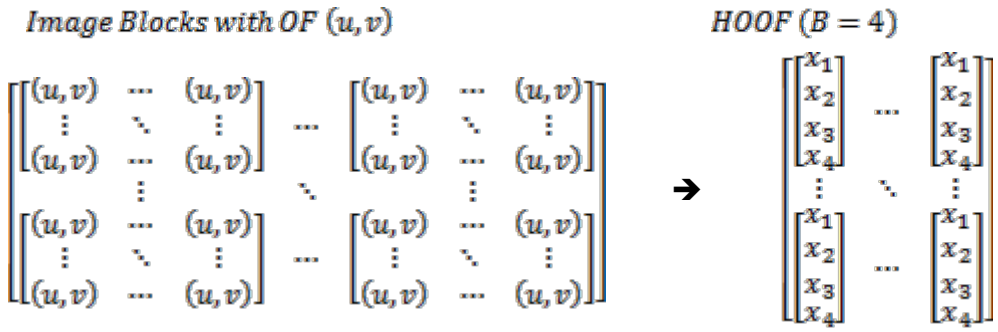


Figure 24. HOOF generation in matrix form.

5.4 Bag of Words (BoW) Outputs

After extracting 20 videos of the six basic emotions (Happiness, Sadness, Anger, Fear, Disgust and Surprise), we obtain observation symbol sequence via vector quantization. Because a discrete 3D HMM is employed to decode the temporal variations of the facial expression features and the discrete HMMs are usually trained and tested using sequence of symbols, the feature vectors need to be transformed to symbols by comparing with the codeword vectors of a codebook. To obtain the codebook, vector quantization algorithm is performed on the feature vectors from the training datasets. A clustering algorithm (e.g., K-means) is generally used for codebook generation. It selects the initial centroids and splits the whole dataset based on the centroids. Then, it recalculate the centroids from the split datasets and continues to split the dataset according to the codeword size. The optimization is obtained when the distortion between two consecutive centroids is minimized. As shown in figure 25 each centroid vector is a codeword of a code book (similar to word dictionary).

The main idea is to quantize the extracted feature vector of each patch/block into one of visual words, and then represent the observations of 3D HMM by sequences of the visual words. A codeword can be considered as a representative of several similar patches/blocks.

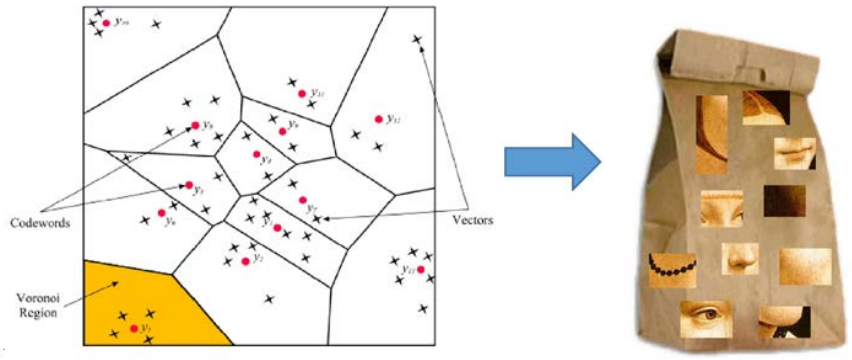


Figure 25. Left shows VQ (Vector Quantization) codewords in 2-dimensional space with input vectors (x), codewords (●) and Voronoi regions. Right shows a pictorial bag of words (BoW) from an image of a face. (Source: http://www.mqasem.net/vectorquantization/vq.html and http://people.csail.mit.edu/fergus/iccv2005/bagwords.html)

5.5 3D HMM Classification Result

Table 1. 3D HMM classification confusion matrix

	Happiness	Sadness	Anger	Fear	Disgust	Surprise
Happiness	92.7%	0.5%	1.6%	3.7%	1.0%	0.5%
Sadness	0.8%	87.8%	3.2%	2.4%	5.0%	0.8%
Anger	0.5%	5.3%	87.3%	3.2%	3.5%	1.2%
Fear	2.0%	3.1%	4.2%	89.2%	1.0%	0.5%
Disgust	0.8%	2.1%	4.2%	0.5%	91.8%	0.6%
Surprise	0.5%	0.5%	4.2%	0.4%	0.5%	93.9%

6 Conclusion

In this work we have shown the viability of doing facial expression recognition by characterizing facial actions in terms of optical flows, transforming magnitude and direction of optical flow to a histogram and subsequently into bag of words for classification of 3D HMM. A regular HMM has limited dimension of either time or space but not both. It was first applied to facial expression recognition by Yamato et al. [43]. We used a 3D spatio-temporal HMM to apply to facial expression recognition and showed how the extension of dimension is more intuitive to track the motion or behavior in a video. Future extensions include exploring other spatio-temporal features and extending the current 1D or 2D HMM applications to the spatio-temporal domain. Schmidt et al. [32] compared three different features (principal component analysis, orientation histograms and optical flow estimation) on emotional recognition with HMM. Using features detected at multiple scales and segmentation of video frames at shorter interval should improve performance at the cost of computer time. Another possible direction of future work is to incorporate a parallelism in running of 3D HMM using a massively parallel hardware architecture and high performance of floating point arithmetic and memory operations on GPU (Graphics Processing Unit).

ACKNOWLEDGEMENT

We thank Jasser Jasser of ECE Department of Oakland University for assistance with development and running of the experiment and sharing of his pearls of wisdom with us during the course of this research.

REFERENCES

[1]. Bashyal, Shishir, and Ganesh K. Venayagamoorthy. 2008. Recognition of facial expressions using gabor wavelets and learning vector quantization. *Engineering Applications of Artificial Intelligence* 21 (7): 1056-64.

[2]. Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110 (3): 346-59.

[3]. Bouchaffra, Djamel. 2010. Conformation-based hidden markov models: Application to human face identification. *Neural Networks, IEEE Transactions on* 21 (4): 595-608.

[4]. Bouchaffra, Djamel, and Abbes Amira. 2008. Structural hidden markov models for biometrics: Fusion of face and fingerprint. *Pattern Recognition* 41 (3): 852-67.

[5]. Bruhn, Andrés, Joachim Weickert, and Christoph Schnörr. 2005. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision* 61 (3): 211-31.

[6]. Chaudhry, Rizwan, Arunkumar Ravichandran, Georg Hager, and René Vidal. 2009. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. Paper presented at Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, .

[7]. Cohen, Ira, Nicu Sebe, Ashutosh Garg, Lawrence S. Chen, and Thomas S. Huang. 2003. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding* 91 (1): 160-87.

[8]. Donoser, Michael, and Horst Bischof. 2006. Efficient maximally stable extremal region (MSER) tracking. Paper presented at Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, .

[9]. Draper, Bruce A., Kyungim Baek, Marian Stewart Bartlett, and J. Ross Beveridge. 2003. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding* 91 (1): 115-37.

[10]. Ekman, Paul, and Wallace V. Friesen. 1977. Facial action coding system.

[11]. Elramsisi, AM, MA Zohdy, and NK Loh. 1991. A joint frequency-position domain structure identification of nonlinear discrete-time systems by neural networks. *Automatic Control, IEEE Transactions on* 36 (5): 629-32.

[12]. Hamm, Jihun, Christian G. Kohler, Ruben C. Gur, and Ragini Verma. 2011. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods* 200 (2): 237-56.

[13]. Horn, Berthold K., and Brian G. Schunck. 1981. Determining optical flow. Paper presented at 1981 Technical symposium east, .

[14]. Jafri, Rabia, and Hamid R. Arabnia. 2009. A survey of face recognition techniques. *Journal of Information Processing Systems* 5 (2): 41-68.

[15]. Jitén, Joakim, and Bernard Merialdo. 2007. Video modeling using 3D hidden markov model. Paper presented at VISAPP (1), .

[16]. Joshi, Dhiraj, Jia Li, and James Z. Wang. 2006. A computationally efficient approach to the estimation of two-and three-dimensional hidden markov models. *Image Processing, IEEE Transactions on* 15 (7): 1871-86.

[17]. Kanade, Takeo, Jeffrey F. Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. Paper presented at Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, .

[18]. Karray, Fakhreddine, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. 2008. Human-computer interaction: Overview on state of the art.

[19]. Khan, Aftab Ali, and MA Zohdy. 1997. A genetic algorithm for selection of noisy sensor data in multisensor data fusion. Paper presented at American Control Conference, 1997. Proceedings of the 1997, .

[20]. Ledoux, H., and CM Gold. 2005. Interpolation as a tool for the modelling of three-dimensional geoscientific datasets. Paper presented at Proceedings ISPRS

[21]. Lien, James J., Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. 1998. Automated facial expression recognition based on FACS action units. Paper presented at Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, .

[22]. Lien, Jenn-Jier James. 1998. *Automatic Recognition of Facial Expressions using Hidden Markov Models and Estimation of Exprssion Intensity*.

[23]. Lowe, David G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60 (2): 91-110.

[24]. Lucas, Bruce D., and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. Paper presented at IJCAI, .

[25]. Mahoor, Mohammad H., Steven Cadavid, Daniel S. Messinger, and Jeffrey F. Cohn. 2009. A framework for automated measurement of the intensity of non-posed facial action units. Paper presented at Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, .

[26]. Matas, Jiri, Ondrej Chum, Martin Urban, and Tomás Pajdla. 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing* 22 (10): 761-7.

[27]. Moreno, Plinio, Alexandre Bernardino, and José Santos-Victor. 2005. Gabor parameter selection for local feature detection. In *Pattern recognition and image analysis.*, 11-19Springer.

[28]. Pinto, Sílvia Cristina Dias, Jesús P. Mena-Chalco, Fabrício Martins Lopes, Luiz Velho, and Roberto Marcondes Cesar Jr. 2011. 3D facial expression analysis by using 2D and 3D wavelet transforms. Paper presented at Image Processing (ICIP), 2011 18th IEEE International Conference on, .

[29]. Praseeda Lekshmi, V., and M. Sasikumar. 2009. Analysis of facial expression using gabor and SVM. *International Journal of Recent Trends in Engineering* 1 (2): 1-43.

[30]. Rabiner, Lawrence R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77 (2): 257-86.

[31]. Sánchez, A., José V. Ruiz, Ana Belén Moreno, Antonio S. Montemayor, Javier Hernández, and Juan José Pantrigo. 2011. Differential optical flow applied to automatic facial expression recognition. *Neurocomputing* 74 (8): 1272-82.

[32]. Schmidt, Miriam, Martin Schels, and Friedhelm Schwenker. 2010. A hidden markov model based approach for facial expression recognition in image sequences. In *Artificial neural networks in pattern recognition.*, 149-160Springer.

[33]. Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing* 27 (6): 803-16.

[34]. Sirovich, Lawrence, and Michael Kirby. 1987. Low-dimensional procedure for the characterization of human faces. *Josa a* 4 (3): 519-24.

[35]. Soyel, Hamit, and Hasan Demirel. 2012. Localized discriminative scale invariant feature transform based facial expression recognition. *Computers & Electrical Engineering* 38 (5): 1299-309.

[36]. Tian, Ying-li, Takeo Kanade, and Jeffrey F. Cohn. 2002. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. Paper presented at Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, .

[37]. ———. 2001. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23 (2): 97-115.

[38]. Tsai, Ching-Chih, You-Zhu Chen, and Ching-Wen Liao. 2009. Interactive emotion recognition using support vector machine for human-robot interaction. Paper presented at Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, .

[39]. Turk, Matthew, and Alex P. Pentland. 1991. Face recognition using eigenfaces. Paper presented at Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on, .

[40]. Uddin, Md Zia, JJ Lee, and T-S Kim. 2009. An enhanced independent component-based human facial expression recognition from video. *Consumer Electronics, IEEE Transactions on* 55 (4): 2216-24.

[41]. Valstar, Michel, and Maja Pantic. 2006. Fully automatic facial action unit detection and temporal analysis. Paper presented at Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, .

[42]. Wu, Jian, Zhiming Cui, Victor S. Sheng, Pengpeng Zhao, Dongliang Su, and Shengrong Gong. 2013. A comparative study of SIFT and its variants. *Measurement Science Review* 13 (3): 122-31.

[43]. Yamato, Junji, Jun Ohya, and Kenichiro Ishii. 1992. Recognizing human action in time-sequential images using hidden markov model. Paper presented at Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on, .

[44]. Zohdy, M., D. Bouchaffra, and J. Quinlan. 2001. Optimal mapping from chromosome space to feature space for solving sequential pattern recognition problems. Paper presented at Circuits and Systems, 2001. MWSCAS 2001. Proceedings of the 44th IEEE 2001 Midwest Symposium on, .

SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 6
ISSN 2054-7390

Research on Linear Fractional Town Traffic Flow Model Tactic

Chunna Zhao
School of Information Science and Engineering, Yunnan University, Kunming, Yunnan, China
chunnazhao@163.com

ABSTRACT

Traffic flow is a worldwide problem. It has many influencing factors and it is the complex system. Fractional calculus is a powerful tool for dealing with complex systems. Fractional calculus is a direct way of extending traditional integer order calculus, which allows the order to be a fraction. Fractional order model achieves better results than the integer order model. A linear fractional order model based on Grunwald–Letnikov's definition for traffic flow is proposed in this paper. City road traffic flow system is composed of a large number of complex dynamic behaviors of traffic participant. It is a highly nonlinear and non-stationary complex system. Firstly, fractional order calculus is introduced. Then the linear fractional order traffic flow model is proposed based on fractional calculus. The fractional order parameters can be determined by a large number of data and mathematical statistics method. The proposed model was simulated and applied to actual Linghai town road traffic flow. The practicability and effectiveness of the method have been validated.

Keywords: Fractional Order Calculus; Grunwald–Letnikov's Definition; Linear; Traffic Flow; Model.

1 Introduction

Traffic flow theory is a scientific which uses physics and mathematical tools to describe the traffic characteristics. The earliest theory was formed in the 1930s. The probabilistic methods were adopted in the early theory of traffic flow basically. Now researchers are using a variety of methods to study the traffic flow. Reference [1] proposed a new stochastic model of traffic flow with state dependent headways. And the proposed model is consistent with the CTM in the mean dynamic sense. The flow of traffic was analyzed under the operation of variable speed limits in the reference [2]. A new traffic flow model called shockwave profile model was developed in reference [3]. The model is suitable for modeling traffic flow on congested arterials and is empirically validated using field data. The influence of vehicle-to-vehicle communication on traffic flow has been studied in reference [4]. A peak hour traffic scenario with open boundaries and an on-ramp was examined in that paper. Traffic flow of mixed modal share generates dissimilar road density with time in reference [5]. Scale of pollutant dispersion depends upon different traffic flow conditions and spatial distribution of concentrations changes with the traffic flow pattern. Optimal traffic flow management was achieved by delay control in reference [6]. Parallel computing platform accelerates the computation for traffic flow management. And decomposition methods are promising for a real-time traffic flow management platform. The slow-to-start effect in two-dimensional Biham–Middleton–Levine traffic flow model

DOI: 10.14738/tmlai.36.1688
Publication Date: 28th December, 2015
URL: http://dx.doi.org/10.14738/tmlai.36.1688

with traffic light periods was studied in reference [7]. And they have explained this via the evolution process from a designed regular initial configuration.

At present, non-equilibrium traffic flow is managed by high order continuous model. But some relevant parameters are difficult to accurately identify, especially the function of specific traffic behaviours and mechanism. These have a direct impact on the reliability of simulation results. Therefore, the establishment of a suitable traffic flow model has important theoretical significance and practical value. We should explore the traffic laws and establish a set of strict theoretical basis taking real-time, accuracy, reliability and mixed traffic flow into account. There is a wealth of nonlinear dynamic characteristics in the traffic flow system. And it is a highly nonlinear complex system involved human intervention. The natural world is essentially a fractional. Fractional model is suitable for processing with the human factors and fractional characteristics of complex systems. Fractional character is also suitable for dealing with nonlinear fluid mechanics characteristics.

It is well known that fractional order systems itself is an infinite dimensional filter due to the fractional order in the differentiator or integrator while the integer-order systems are with limited memory (finite dimensional). There has been a surge of interest in the possible engineering application of fractional order differentiation. Examples may be found in [8] and [9]. The significance of fractional order theory is that it is a generalization of classical integral order theory, which could lead to more adequate modeling and more robust control performance.

Fractional order systems could model various real materials more adequately than integer order ones and thus provide an excellent modeling tool in describing many actual dynamical processes[10]. Fractional model provided the scientific basis for prevention and treatment of satellite monitoring absorption rate [11]. The nematode movement can be simulated through fractional model [12]. It may be used for building “love” models using fractional-order system [13]. Reference [14] modeled iron meteorites crystallization by fractional theory. And there are some people pay close attention to unemployment rates by means of fractional calculus [15].

Linear fractional traffic flow model is proposed in this paper. Model parameters can be obtained by the corresponding actual data. The artificial and natural factors for the road traffic flow can be effectively dealt with fractional model. And it can improve the rationality and accuracy. Fractional traffic flow model can describe traffic flow more objective and accurate. Fractional traffic flow control strategy can be put forward on the basis of fractional traffic flow forecast.

The remaining part of this paper is organized as follows. In Section II, fractional order calculus is introduced. In Section III, the linear fractional model method is presented for traffic flow. In Section IV, simulation examples are presented to verify the feasibility. Finally, conclusions are drawn in the paper.

2 Fractional Order Calculus

Although the fractional order calculus is a 300-years old topic, the theory of fractional order derivative was developed mainly in the 19th century. Reference [16] provided a good source of references on fractional order calculus. Fractional order calculus is a generalization of integration and differentiation to a fractional, or non-integer order fundamental operator ${}_aD_t^{\alpha}$, where a and t are the lower/upper bounds of integration and α the order of the operation.

$$ {}_aD_t^{\alpha} = \begin{cases} \dfrac{d^{\alpha}}{dt^{\alpha}} & R(\alpha) > 0 \\ 1 & R(\alpha) = 0 \\ \int_a^t (d\tau)^{(-\alpha)} & R(\alpha) < 0 \end{cases} \tag{1} $$

which $R(\alpha)$ is the real part of α. Moreover, the fractional order can be a complex number as discussed in [17]. In this paper, we focus on the case where the fractional order is a real number.

Due to its importance in applications, the Grunwald–Letnikov's definition of the fractional order calculus is considered in this paper based on the generalization of the backward difference.

Definition 1: Introducing the positive integer *m*, $[m]$ is the integer part of *m*. Then this definition has the form

$$ \begin{aligned} {}_aD_t^{\alpha} f(t) &= \lim_{h \to 0} h^{-\alpha} \sum_{j=0}^{[(t-\alpha)/h]} (-1)^j \binom{\alpha}{j} f(t - jh) \\ &= \lim_{h \to 0} \frac{1}{\Gamma(\alpha) h^{\alpha}} \sum_{j=0}^{[(t-\alpha)/h]} \frac{\Gamma(\alpha + j)}{\Gamma(j+1)} f(t - jh) \end{aligned} \tag{2} $$

where

$$ \binom{\alpha}{j} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-j+1)}{j!} = \frac{\alpha!}{j!(\alpha-j)!} \tag{3} $$

Firstly, the n-order derivative of continuous function $f(t)$ is

$$ f^{(n)}(t) = \lim_{h \to 0} h^{-n} \sum_{j=0}^{[(t-\alpha)/h]} (-1)^j \binom{n}{j} f(t - jh) \tag{4} $$

where

$$ \binom{n}{j} = \frac{n(n-1)(n-2)\cdots(n-j+1)}{j!} = \frac{n!}{j!(n-j)!} \tag{5} $$

When n is negative integer, there is

$$ \binom{-n}{j} = (-1)^n \binom{n}{j} \tag{6} $$

When n is any real number, there is the Grunwald–Letnikov's definition.

Compared with conventional calculus, fractional order calculus has the following qualities:

- If $f(t)$ is an analytic function of the variable t, the derivative $D_t^{\alpha} f(t)$ is an analytic function of t and α.
- The operation D_t^{α} has the same result with the usual derivative when $\alpha = n$ and $n \in Z^+$.
- The operation D_t^{α} has the same result with the usual n-fold integral when $\alpha = n$ and $n \in Z^-$.
- The operator of order $\alpha = 0$ is the identity operator.

- The operator should be linear:

$$D_t^{\alpha}[af(t)+bg(t)]=aD_t^{\alpha}f(t)+bD_t^{\alpha}g(t) \tag{7}$$

- For the fractional-order integrals of arbitrary order, it holds the additive law of exponents (semi-group property):

$$D_t^{\alpha}[D_t^{\beta}f(t)]=D_t^{\alpha+\beta}f(t) \tag{8}$$

- It can be seen that the fractional-order integral can be expressed as a causal convolution of the form：

$${}_aI_t^{\alpha}f(t)=\Phi_{\alpha}(t)*f(t),\qquad \alpha\in R^{+} \tag{9}$$

where $\Phi_{\alpha}(t)=\dfrac{t_{+}^{\alpha-1}}{\Gamma(\alpha)}$, $\alpha\in R^{+}$.

As in the case for conventional linear systems, for linear fractional-order systems, linear fractional-order differential equations are the fundamental governing equations. The linear fractional-order differential equation is defined as

$$\begin{aligned}&a_nD_t^{\beta_n}y(t)+a_{n-1}D_t^{\beta_{n-1}}y(t)+\cdots+a_1D_t^{\beta_1}y(t)+a_0D_t^{\beta_0}y(t)\\&=b_mD_t^{\alpha_m}x(t)+b_{m-1}D_t^{\alpha_{m-1}}x(t)+\cdots+b_1D_t^{\alpha_1}x(t)+b_0D_t^{\alpha_0}x(t)\end{aligned} \tag{10}$$

Denoting the left hand side of the equation by

$$u(t)=b_mD_t^{\alpha_m}x(t)+b_{m-1}D_t^{\alpha_{m-1}}x(t)+\cdots+b_0D_t^{\alpha_0}x(t) \tag{11}$$

The original fractional-order differential equation can be rewritten in the form

$$a_nD_t^{\beta_n}y(t)+a_{n-1}D_t^{\beta_{n-1}}y(t)+\cdots+a_0D_t^{\beta_0}y(t)=u(t) \tag{12}$$

Substituting Grunwald–Letnikov's definition into the above equation, one may find that

$$\frac{\alpha_0}{h^{\beta_0}}\sum_{j=0}^{[(t-a)/h]}\omega_j^{(\beta_i)}y_{t-jh}+\cdots+\frac{\alpha_m}{h^{\beta_n}}\sum_{j=0}^{[(t-a)/h]}\omega_j^{(\beta_n)}y_{t-jh}=u(t) \tag{13}$$

where the binomial coefficients $\omega_j^{(\beta_i)}$ can still be evaluated recursively with

$$\omega_0^{(\beta_i)}=1,\quad \omega_j^{(\beta_i)}=\left(1-\frac{\beta_i+1}{j}\right)\omega_{j-1}^{(\beta_i)},\quad j=1,2,\cdots \tag{14}$$

By slight rearrangement of the terms, the closed-form solution of the fractional order differential equation can be obtained as

$$y_t=\frac{1}{\sum_{i=0}^{n}\dfrac{\alpha_i}{h^{\beta_i}}}\left[u_t-\sum_{i=0}^{n}\frac{\alpha_i}{h^{\beta_i}}\sum_{j=1}^{[(t-a)/h]}\omega_j^{(\beta_i)}y_{t-jh}\right] \tag{15}$$

The linear fractional traffic flow model is considered in this paper. Model parameters can be obtained by the corresponding actual data. It aims to describe the traffic flow of the actual situation more accurately, so that fractional control systems can be used on traffic flow in the future. It can reduce the complexity while improving the scientific validity of the system model based on the linear fractional order calculus.

3 Linear Fractional Model

Accurate mathematical and computer simulation model is a very effective and vital tool for supporting a new road and transport infrastructure planning. The appropriate traffic flow model can be used to evaluate design alternatives, which can greatly reduce the costs. Reference [18] introduced a new hybrid approach which combines the complementary features and capabilities of both continuum mathematical models e.g. and knowledge-based models e.g. in order to describe effectively traffic flow in road networks. Convective instability is relevant for describing congested traffic flow in reference [19]. Only two out of five stability classes agree with observations. And synchronized traffic and jams can be described by "two-phase" models. A short-term traffic forecasting model which combines the support vector regression model with continuous ant colony optimization algorithms (SVRCACO) was studied in reference [20]. And the SVRCACO model is a promising alternative for forecasting traffic flow. A new lattice model of traffic flow with the consideration of the traffic interruption probability has been presented in reference [21]. The traffic interruption probability on the stability of traffic flow was explored. Quantitative criteria for validating models with respect to traffic instabilities were applied to a database of hundreds of congestions in reference [22]. The railway traffic flow was simulated when a train suddenly stops with malfunction in reference [23]. Two ways are proposed to alleviate the harmful influence caused by train failure. Increase of train operation interval can decrease trains' delay. And decrease of maximum speed can reduce the go-and-stop wave. Mainstream traffic flow control (MTFC) is proposed as a novel and efficient motorway traffic management tool, and its possible implementation and principal impact on traffic flow efficiency was analyzed [24]. Variable speed limit was considered as one (out of several possible) way(s) for MTFC realization, either as a stand-alone measure or in combination with ramp metering. A novel non-linear modeling formalism was suggested for freeway traffic flow in reference [25]. A second-order macroscopic model has been transformed into a Linear Parameter Varying (LPV) model. A new driver's forecast lattice model of traffic flow has been presented in reference [26]. The analytical and numerical results show that the driver's forecast effect can improve the stability of traffic flow. A linear fractional model is introduced in this paper. Traffic flow of a location on urban highway is represented by the fractional model.

In this system the input is hours and the output is traffic flow. The order of fractional has the important meaning. They show man-made and natural factors. Combined other correlation coefficients fractional model is shown as the following:

$$\mathrm{a}_1 D^{\alpha_1} y(t) + \mathrm{a}_2 D^{\alpha_2} y(t) + \mathrm{a}_3 D^{\alpha_3} y(t) + \mathrm{a}_4 D^{\alpha_4} y(t) = bx(t) \tag{16}$$

where $x(t)$ expresses the hours, and $y(t)$ is the traffic. Fractional coefficient α_1 shows emergency situations. In general, there are many emergency situations. The value indicates the emergency of possibility. The larger the value of fractional coefficient shows more likely emergencies. And it is converse if the value is smaller. This fractional coefficient is decided by location and position. Fractional coefficient α_2 shows driver's state of mind. In Linghai traffic condition, psychological state of driver is good. So the value of this coefficient is relatively large. Fractional coefficient α_3 shows holidays. A position vehicle flow is related to data collection time. Whether it is holiday? Holidays are approaching? Fractional coefficient α_4 shows rain and snow status. When weather is bad, traffic will

be reduced, which is well-recognized fact. Common coefficients are determined by some vehicle factors. Coefficient a_1 shows overtaking and changing lanes phenomenon. This coefficient represents the possibility of overtaking and diversion. The greater value shows that this possibility is greater. Coefficient a_2 expresses the interaction of individual vehicles. Coefficient a_3 expresses randomness of determination. Coefficient a_4 expresses vehicle inertia. Coefficient b expresses speed and distance between vehicles. The MATLAB function is written to implement the step response of the system.

4 Simulation Examples

Here Linghai town road traffic flow should be considered in this paper. This place actual traffic data has been gotten. Parameters of the above model can be determined by the data through other methods. So the linear fractional traffic flow mode is established, which can provide the basis for traffic management.

The corresponding data can be obtained by the perception of space equipment. The data should be dealt with through the quadratic approximation, the weighted minimum and fractional approximation methods. And then corresponding coefficients can be received. This fractional coefficient α_1 is decided by location and position. There are a lot of unexpected events at this location. From the point of statistical data, this is moderate attention orientation. So there is $\alpha_1 = 3.1$. In Linghai traffic condition, the psychological state of driver is not good. So the value of this coefficient is relatively small. It is $\alpha_2 = 0.9$. A position vehicle flow is related whether it is holiday. There are more festivals in September. In addition to the weekend, there is Mid-Autumn Festival, Teacher's Day and so on. So there is $\alpha_3 = 2.8$. The fractional coefficient α_4 shows Weather conditions. In general the weather is better in September. The inclement weather is less. So it is $\alpha_4 = 4.5$. Common coefficients show vehicle factors. The location is in the main ring road. It is not close to the export or import. In general, there is not much overtaking and lane changing. The value is not lager. $a_1 = 8$. The coefficient a_2 represents the interaction between vehicles. The Impact between vehicles is decided by many factors. Based on the existing data and through fractional analysis method, we get $a_2 = 7$. In all cases the determination randomness is quite lager. Which also shows that traffic flow model is not unique. So establishment uniform traffic flow model is not realistic. The actual location is considered in this part. There is the certain basis on the existing data. Here the coefficient is $a_3 = 4$. Coefficient a_4 expresses vehicle inertia. Vehicle inertia is great. So it is $a_4 = 6$. On the urban expressway, the relation of speed and distance between vehicles is very lager. The more value is taken, which shows that the relationship is very close. There is $b = 8$. Based on the above data and related methods, the fractional program evaluation model can be obtained:

$$8D^{3.1}y(t) + 7D^{0.9}y(t) + 4D^{2.8}y(t) + 6D^{4.5}y(t) = 8x(t) \tag{17}$$

For the fractional model, simulation model can be built by using the above algorithm. It is taken as the basis for traffic management that the curve can be gotten from the simulation model. According to the algorithm proposed in this paper, the corresponding traffic flow curve can be obtained with MATLAB from 15:00 to 22:00, as shown in Figure 1.

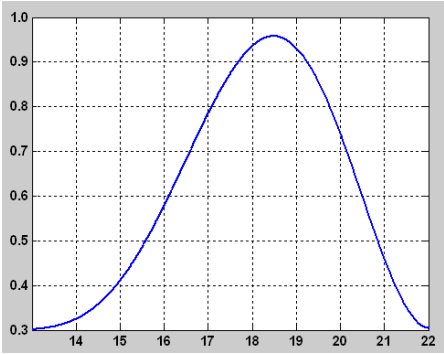


Figure 1. Output traffic flow curve

The result shows that peak traffic flow is gotten when it is six-half at the afternoon. Then the traffic flow falls. The traffic flow tends to the usual level of traffic more than nine-point in the evening. This result is consistent with the actual traffic flow. It is validated that the proposed method is very accurate and effective.

During the holidays, there have some changes in the corresponding data. Then the corresponding coefficient will change. According to the holiday period traffic flow data, the corresponding coefficients can be received. There is $\alpha_1 = 2.5$. The coefficient is $\alpha_2 = 0.7$. Because it is the New Year holiday period, there is $\alpha_3 = 3.2$. The fractional coefficient α_4 shows Weather conditions. Weather is changeable. So it is $\alpha_4 = 3.9$. Common coefficients show vehicle factors. There is $a_1 = 5$. The coefficient has $a_2 = 9$. Here the coefficient is $a_3 = 8$. Coefficient a_4 expresses vehicle inertia. So it is $a_4 = 4$. The relation of speed and distance between vehicles is very lager. The more value is taken, which shows that the relationship is very close. There is $b = 9$. Based on the above data and related methods, the fractional program evaluation model can be obtained:

$$5D^{2.5}y(t) + 9D^{0.7}y(t) + 8D^{3.2}y(t) + 4D^{3.9}y(t) = 9x(t) \tag{18}$$

For the fractional model, simulation model can be built by using the above algorithm. It is taken as the basis for traffic management that the curve can be gotten from the simulation model. According to the algorithm proposed in this paper, the corresponding traffic flow curve can be obtained with MATLAB from 7:00 to 20:00, as shown in Figure 2.

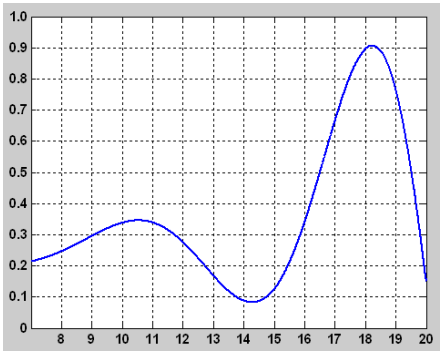


Figure 2. Output traffic flow curve

In the New Year holiday period, there are many friends eating and drinking at noon. So there is more traffic flow. At two and three point in the afternoon, most people are at rest. The peak traffic flow is gotten when it is six-point at afternoon. This result is consistent with the actual traffic flow. It is validated that the proposed method is very accurate and effective.

5 Conclusion

The traffic phenomenon is researched in traffic flow theory. General characteristics of transportation are described by measuring and modeling method. The mathematics and physics theory are applied in the various parameters of traffic flow and their qualitative and quantitative analysis. Traffic flow theory provides a theoretical basis and basic method for traffic engineering application. And the basic law of traffic flow should be interpreted. There are two functions of traffic flow model in traffic engineering application. The first function is to better understand the dynamics of traffic, especially to explain the formation and spread of congestion. Therefore, transportation scholars use identify the possible traffic bottleneck by using the traffic model. The second role is that the traffic flow model can be used as a simulation platform. Through this platform, various control strategies can be developed and assessed in order to improve traffic conditions. Traffic flow is a complex multi-factorial process, and it is not modeled by integer order model accurately. Fractional model is an effective method for analysis and processing of complex system.

Linear fractional modeling method of traffic flow is proposed in this paper. The coefficients of the model are obtained by a large number of related data. The method was applied to actual town transportation. And result indicates that this method is highly efficient for solving real-world problems. The validity of proposed method is validated by actual situation.

ACKNOWLEDGMENTS

This work was supported by the International S&T Cooperation Program of China (2010DFB10930, 2011DFG13000); the National Natural Science Foundation of China (60873006, 61070049, 61170304, 61104035, 61174145, 61201378); the Beijing Natural Science Foundation, Beijing Outstanding Talents Project, and the S&R Key Program of BMEC (4122017,KZ201210028036, KM201010028021, 2012D005016000011).

REFERENCES

[1]. Saif Eddin Jabari, Henry X. Liu. *A stochastic model of traffic flow: Theoretical foundations.* Transportation Research Part B: Methodological, 2012, 46(1): 156-174.

[2]. B.G. Heydecker, J.D. Addison. *Analysis and modelling of traffic flow under variable speed limits.* Transportation Research Part C: Emerging Technologies, 2011, 19(2): 206-217.

[3]. Xinkai Wu, Henry X. Liu. *A shockwave profile model for traffic flow on congested urban arterials.* Transportation Research Part B: Methodological, 2011, 45(10): 1768-1786.

[4]. Florian Knorr, Michael Schreckenberg. *Influence of inter-vehicle communication on peak hour traffic flow.* Physica A: Statistical Mechanics and its Applications, 2012, 391(6): 2225-2231.

[5]. Sharad Gokhale. *Traffic flow pattern and meteorology at two distinct urban junctions with impacts on air quality.* Atmospheric Environment, 2011, 45(10): 1830-1840.

[6]. D. Sun, A. Clinet, A.M. Bayen. *A dual decomposition method for sector capacity constrained traffic flow optimization.* Transportation Research Part B: Methodological, 2011, 45(6): 880-902.

[7]. Qiao-Hong Sui, Zhong-Jun Ding, Rui Jiang, Wei Huang, Duo Sun, Bing-Hong Wang. *Slow-to-start effect in two-dimensional traffic flow.* Computer Physics Communications, 2012, 183(3):547-551

[8]. chunna zhao, liming luo, and yu zhao. *Fractional Modeling Approach with Mittag-Leffler Functions for Linear Fractional-order System*. International Conference on Intelligent Computation Technology and Automation. 2012.1. 386-389.

[9]. chunna zhao, liming luo, and yingshun li. *Ecological environment evaluation method based on ideal correlation degree.* Advances in Biomedical Engineering, 2012, Vol. 7 : 142-148.

[10]. C. N. Zhao and X. D. Zhang, *The application of fractional order PID controller to position servomechanism*, IEEE WCICA, 2008:3380-3383.

[11]. O. Pontus and E. Lars, *Estimation of absorbed PAR across Scandinavia from satellite measurements.* Part II: Modeling and evaluating the fractional absorption, Remote Sensing of Environment, 2007, 110: 240-251.

[12]. S. Hapca, J. W. Crawford, K. MacMillan, et al, *Modelling nematode movement using time-fractional dynamics,* Journal of Theoretical Biology, 2007, 248: 212-224.

[13]. W. M. Ahmad and R. El-Khazali, *Fractional-order dynamical models of love*, Chaos, Solitons & Fractals, 2007, 33: 1367-1375.

[14]. J. W. Richard, F. M. William, H. Jenise, et al, *Modeling fractional crystallization of group IVB iron meteorites,* Geochimica et Cosmochimica Acta, 2008, 72: 2198-2216.

[15]. M. C. Guglielmo and A. G. Luis, *Modelling the US, UK and Japanese unemployment rates: Fractional integration and structural breaks,* Computational Statistics & Data Analysis, 2008, 52: 4998-5013.

[16]. Chunna Zhao, Yingshun Li, and Tao Lu. *Fractional System Analysis and Design.* National Defence Industry Press, 2011.

[17]. A. Oustaloup, F. Levron, B. Mathieu, and F.M. Nanot, *Frequencyband complex noninteger differentiator: characterization and synthesis*,IEEE Trans. Circuits Syst. I, 2000, 47: 25-39.

[18]. Jennifer McCrea, Salissou Moutari. *A hybrid macroscopic-based model for traffic flow in road networks.* European Journal of Operational Research, 2010, 207(2): 676-684.

[19]. Martin Treiber, Arne Kesting. *Evidence of convective instability in congested traffic flow: A systematic empirical and theoretical investigation.* Transportation Research Part B: Methodological, 2011, 45(9): 1362-1377.

[20]. Wei-Chiang Hong, Yucheng Dong, Feifeng Zheng, Chien-Yuan Lai. *Forecasting urban traffic flow by SVR with continuous ACO.* Applied Mathematical Modelling, 2011, 35(3): 1282-1291.

[21]. G.H. Peng, X.H. Cai, B.F. Cao, C.Q. Liu. *A new lattice model of traffic flow with the consideration of the traffic interruption probability.* Physica A: Statistical Mechanics and its Applications, 2012, 391(3): 656-663.

[22]. Martin Treiber, Arne Kesting. *Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns.* Transportation Research Part C: Emerging Technologies, 2012, 21(1): 31-41.

[23]. San-Tong Zhang, Yi-Chuan Chen. *Simulation for influence of train failure on railway traffic flow and research on train operation adjusting strategies using cellular automata.* Physica A: Statistical Mechanics and its Applications, 2011, 390(21): 3710-3718.

[24]. Rodrigo C. Carlson, Ioannis Papamichail, Markos Papageorgiou, Albert Messmer. *Optimal mainstream traffic flow control of large-scale motorway networks.* Transportation Research Part C: Emerging Technologies, 2010, 18(2): 193-212.

[25]. Tamás Luspay, Balázs Kulcsár, István Varga, József Bokor. *Parameter-dependent modeling of freeway traffic flow.* Transportation Research Part C: Emerging Technologies, 2010, 18(4): 471-488.

[26]. G.H. Peng, X.H. Cai, C.Q. Liu, B.F. Cao. *A new lattice model of traffic flow with the consideration of the driver's forecast effects.* Physics Letters A, 2011, 375(22):2153-2157.