FEATURE EXTRACTION USING MFCC

Shikha Gupta[1], Jafreezal Jaafar[2], Wan Fatimah wan Ahmad[3] and Arpit Bansal[4]

Universiti Tecknologi PETRONAS, CIS Dept, Perak, Malaysia
Shikha.cs88@gmail.com[1], jafreez@petronas.com.my[2]
fatimhd@petronas.com.my[3]
[4]Indian institute of Information and Technology, Allahabad, India
Arpit06bansal@gmail.com

ABSTRACT

Mel Frequency Ceptral Coefficient is a very common and efficient technique for signal processing. This paper presents a new purpose of working with MFCC by using it for Hand gesture recognition. The objective of using MFCC for hand gesture recognition is to explore the utility of the MFCC for image processing. Till now it has been used in speech recognition, for speaker identification. The present system is based on converting the hand gesture into one dimensional (1-D) signal and then extracting first 13 MFCCs from the converted 1-D signal. Classification is performed by using Support Vector Machine. Experimental results represents that proposed application of using MFCC for gesture recognition have very good accuracy and hence can be used for recognition of sign language or for other household application with the combination for other techniques such as Gabor filter, DWT to increase the accuracy rate and to make it more efficient.

KEYWORDS

Hand gesture, 1D signal, MFCC (Mel Frequency Cepstral Coefficient), SVM (Support Vector Machine).

1. INTRODUCTION

Currently, there is a great focus on developing easy, comfortable interfaces by which human can communicate with computer by using natural and manipulation communication skills of the human. In HCI the input domain requires capturing and then interpretation of the face, facial expression, arms, hands, sometimes whole body motion as well. Among all the inputs, gesture is a powerful means for the communication purpose among human beings. Even the use of gesture is very common while talking on the telephone. The Gesture recognition system has two phases: first one is the feature extraction phase where by using some specific methods few values are assigned for each gesture by using training dataset. It involves extracting important information associated with the given gesture and removing all the remaining useless information. And another phase is the classification processes were based on the training and the testing database the intended gesture get analyzed. Basically Mel frequency Capstral coefficients (MFCC) are very common and one of the best method for feature extraction when talking about the 1D signals. So this paper presents an application of MFCC for hand gesture recognition. Features are extracted by converting input image into 1D signal. For classification purposes SVM is used. SVM it is a supervised Learning method. The benefit of SVM is that it can also use kernels for non-linear data transformation. The Law behind

using SVM is to divide the given data into two dissimilar category and then to get hyper-plane to partition the given classes.

The organization of the rest of the paper is as follow: Section II is all about the recognition system for hand gestures. Section III highlights details about the Mel Frequency Cepstral Coefficients. Section IV describes the experiment results and discussion. Section VI concludes the paper along with small description about the possible future work.

2. RECOGNITION SYSTEM FOR HAND GESTURE

The proposed gesture recognition system is divided into three important stages as shown in figure1: Image conversion from 2D to 1D signal, feature extraction and feature matching also known as classification process. The 2D converted image is given as input to MFCC for coefficients extraction. By doing feature extraction from the given training data the unnecessary data is stripped way leaving behind the important information for classification. The output after applying MFCC is a matrix having feature vectors extracted from all the frames. In this output matrix the rows represent the corresponding frame numbers and columns represent corresponding feature vector coefficients [1-4].

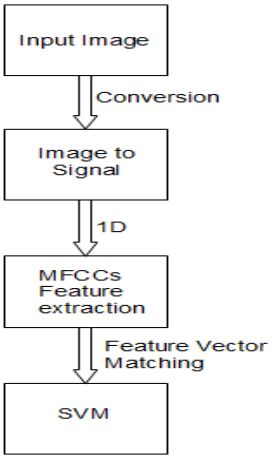

Figure 1: Step by step processing of recognition system

Finally this output matrix is used for classification process. The classification process is divided into two stages: training phase and testing phase. Feature extraction plays a very important in the recognition process. This is basically a process of dimension reduction or feature reduction as this process eliminates the irrelevant data present in the given input while maintaining important information. Several feature extraction techniques [5-14] are there for gesture recognition but in this paper MFCC have been used for feature extraction which is mainly used for speech recognition system. The purpose for using MFCC for image processing is to enhance the effectiveness of MFCC in the field of image processing as well. As per the study MFCC already have application for identification of satellite images [15], face recognition [16] and palm print recognition [17].

Steps for calculating MFCC for hand gestures are the same as for 1D signal [18-21]. Since MFCC works for 1D signal and the input image is a 2D image, so the input image is converted from 2D to 1D signal. Remaining calculation for features extraction is same as for speech signals as shown in figure 3.

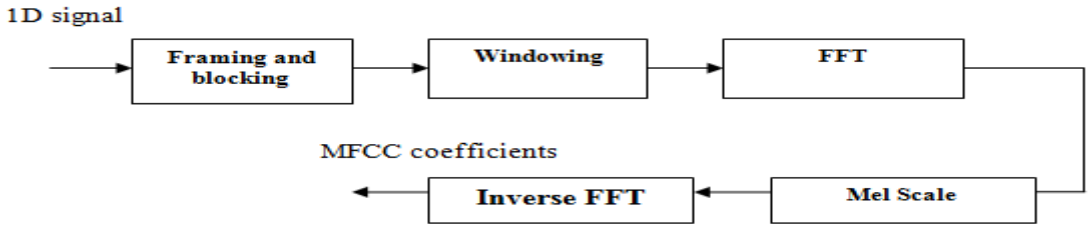

Figure 3: MFCC Processing

2.1. Framing and blocking

In this step the continuous 1D signal are blocked into small frames of N samples, with next frames separated by M samples (M<N) with this the adjacent frames are overlapped by N-M samples. As per many researches the standard value taken for N= 256 and M= 100 with a reason of dividing the given 1D signal into small frames having sufficient samples to get enough information. Because, if the frame size smaller than this size is taken then the number of samples in the frames will not be enough to get the reliable information and with large size frames it can cause frequent change in the information inside the frame. So, while working with MFCC these parameters are very common in practice. This process of breaking up the signals into frames will continue until the whole 1D signal is broken down into small frames.

2.2. Windowing

Windowing is done for minimizing the disruptions at the starting and at the end of the frame, the frame and window function is being multiplied. If the window being defined is W_n (m), $0 \leq m \leq N_m$-1 where N_m stands for the quantity of samples within every frame, the output after windowing the signal will be presented as Y (m) = X (m) W_n (m), $0 \leq m \leq N_m$ -1 where Y (m) represents the output signal after multiplying the input signal represented as (Xm) and Hamming window represented by W_n(m). Basically, many window functions exist such as rectangular window, flat top window and hamming window but, mainly hamming window is applied for carrying out windowing which usually represented as:

$$W_n(m) = 0.54 - 0.46 \cos(2\pi m/(N_m - 1)),\ 0 \leq m \leq N_m - 1 \qquad (1)$$

2.3. *FFT (*Fast Fourier Transform)

FFT is used for doing conversion from the spatial domain to the frequency domain. Each frame having N_m samples are converted into frequency domain. Fourier transformation is a fast algorithm to apply Discrete Fourier Transform (DFT), on the given set of N_m samples shown below:

$$D_k = \sum_{m=0}^{N_m-1} D_m\, e^{\frac{-j2\pi km}{Nm}} \qquad (2)$$

Where k= 0, 1, 2 ….. N_m-1

Basically the definition for FFT and DFT is same, which means that the output for the transformation will be the same; however they differ in their computational complexity. In case of DFT, each frame with N-M samples directly will be used as a sequence for Fourier transformation. On another, in case of FFT this frame will be divided into small DFT's and then computation will be done on this divided

small DFT's as individual sequence thus the computation will be more fast and easy. Thus it is in digital processing or other area instead of directly using DFT, FFT is used for applying DFT.

Commonly, D_k are the combination of real and imaginary numbers thus it represents the complex numbers but, merely absolute values (frequency magnitudes) are considered to carry out further process. The obtained sequence can be interpreted as positive frequencies $0 \leq f < F_s/2$ correspond to values $0 \leq m \leq N_m/2 - 1$, while negative frequencies $-F_s/2 < f < 0$ correspond to values $N_m/2 + 1 \leq m \leq N_m - 1$, F_s is the sampling frequency. By calculating DFT we can obtain the magnitude spectrum.

2.4. *Mel scale*

In this step, the above calculated spectrums are mapped on Mel scale to know the approximation about the existing energy at each spot with the help of Triangular overlapping window also known as triangular filter bank. These filter bank is a set of band pass filters having spacing along with bandwidth decided by steady Mel frequency time. Thus, Mel scale helps how to space the given filter and to calculate how much wider it should be because, as the frequency gets higher these filters are also get wider. For Mel- scaling mapping is need to done among the given real frequency scales (Hz) and the perceived frequency scale (Mels). During the mapping, when a given frequency value is up to 1000Hz the Mel-frequency scaling is linear frequency spacing, but after 1000Hz the spacing is logarithmic as shown in Figure 3. The formula to convert frequency f hertz into Mel m_f is given by Eq.

$$m_f = 2595 log_{10} \left(\frac{f}{700} + 1\right) \quad (3)$$

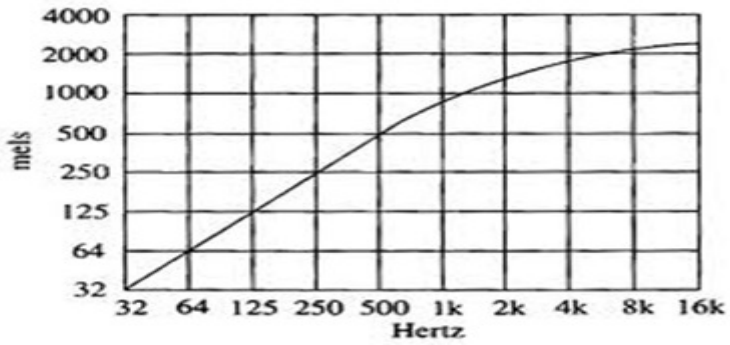

Figure **Error! No text of specified style in document.**: Mel Scale

Thus, with the help of Filter bank with proper spacing done by Mel scaling it becomes easy to get the estimation about the energies at each spot and once this energies are estimated then the log of these energies also known as Mel spectrum can be used for calculating first 13 coefficients using DCT. Since, the increasing numbers of coefficients represent faster change in the estimated energies and thus have less information to be used for classifying the given images. Hence, first 13 coefficients are calculated using DCT and higher are discarded.

2.5. Discrete cosine Transform (DCT)

This process of carrying out DCT is done in order to convert the log Mel spectrum back into the spatial domain. For this transformation either DFT or DCT both can be used for calculating Coefficients from the given log Mel spectrum as they divide a given sequence of finite length data into

discrete vector. However, DFT is generally used for spectral analysis where as DCT used for data compression as DCT signals have more information concentrated in a small number of coefficients and hence, it is easy and requires less storage to represent Mel spectrum in a relative small number of coefficients. This instead of using DFT DCT is desirable for the coefficients calculation as DCT outputs can contain important amounts of energy. The output after applying DCT is known as MFCC (Mel Frequency Cepstrum Coefficient

$$C_n = \sum_{k-1}^{k} (\log D_k) \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right] \quad (4)$$

where m = 0, 1… k- 1

where C_n represents the MFCC and m is the number of the coefficients here m=13 so, total number of coefficients extracted from each frame is 13.

In general, all steps carried out for image processing using MFCC can be summarized as follows: the input image is converted from 2D signal to 1D signal. Then, these 1D matrices are broken up into frames of N_m samples. To prevent loss of information the neighboring frames are separated by M where ($M<N_m$). Hence, the initial frame will constitute first N_m samples then the succeeding frame will begin with M samples after the previous frame to get overlap with the previous frame by N_m-M and so. This is to avoid a loss of the information. As per the study 60% of the overlapping is sufficient. Mostly the values for N_m= 256 and for M=100. After windowing process, each frame consists of N samples and is transformed into the frequency domain by using FFT, which is a high-speed algorithm to implement this conversion from time domain to frequency domain. After DFT calculation the magnitude spectrum is obtained, which is further transformed into the Mel frequency. In the last step, the log of this spectrum is processed in order to get the cepstral coefficients by doing DCT (discrete Cosine Transform). For implementation, in this paper only first 13 coefficients are taken.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this experiment benchmark database with static background is used for input images as shown in Figure 4. The images inside the database are segmented gray scale images having variation in size and orientations. Total numbers of images are 480 (10 different gestures with static background (dark and bright color) done by 24 different person).

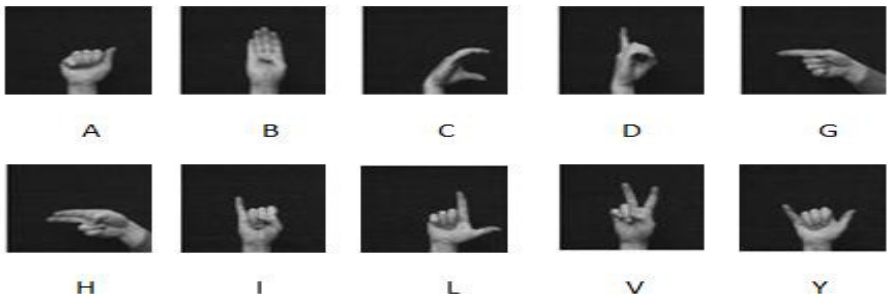

Figure 4: Sample hand posture images from Jochen Triesch dataset

After following the same procedure for the 2D images as used for 1D signal, when first 13 MFCC's are extracted for each given input, the plotting of these extracted features shows that

gestures within the same class even done by different user still look similar as compare to the MFCC's of different classes as shown Figure 5 which represents the raw images taken by different users with a variation in scale and angle and its respective MFCC representation. In the given graphs the X axis represents the number of frames or can say the number of MFCCs extracted from the given input signal. And Y axis represents the feature vector values for each frame.

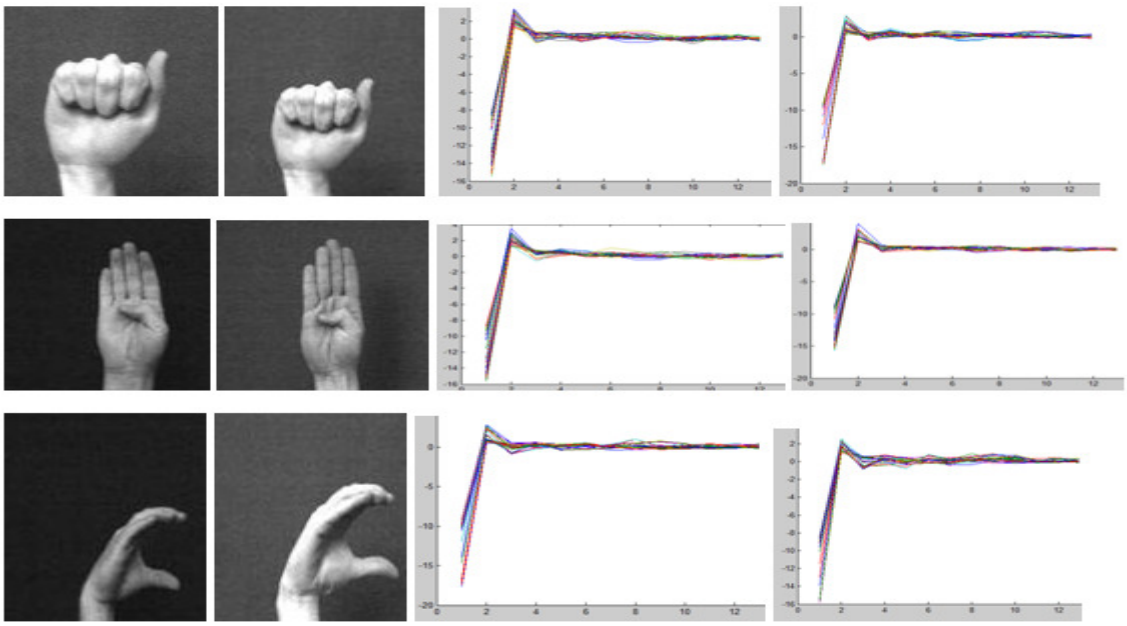

Figure 5: first 13 MFCCs extracted from the gray scaled images

After extracting MFCCs coefficients the feature vectors classification was done using SVM classify. While using SVM two choices are there either classification can be done using One against One SVM or One against all. One against one approach is a pair- wise method; we required (m (m-1) /2) SVM classifier to be trained where: m is the number of classes. The confusion matrix after classification is shown below in Table1.

The matrix displaying the classification results after using SVM for classification. In this work the classification is done between ten different classes. In this confusion matrix the accuracy and the false alarm rate are being shown. It depends individually how to show the accuracy using confusion matrix. In the above shown Matrix the diagonal represents the accuracy for the respective classes and the columns representing the gestures or can also instances in the determined class and in the rows the gesture recognized in the actual class. The errors or misclassified rate is shown outside the diagonal with non-zero values.

From the matrix it is clear that when class1 was classified against the remaining classes the accuracy for the given class is 100%. This accuracy rate represents that all the gesture from the class 1 is identified correctly as positive class gesture during the classification process. In case of class 2 the total accuracy for this class is 92% with error rate of 8%. The misclassification rate of 8% for class 2 shows that while doing the classification 8% of the gestures from class 2 was wrongly identified as class 1 gesture and due to this the false negative rate is 8% and true positive rate is 92% for class 2 with an overall accuracy as 92% for class 2.

For class 3 again the accuracy is just 83.33% which is very less as compare to class 1 and class 2 accuracy rate. When class 3 was classified against the given classes for all the classes the misclassification was zero but with class 4 erroneously 16.67% of the images from class 3 was identified as class 4 images and due to which all other columns have zero value while for class 3 4 the value is 16.67%.Similarly for the class 4 the overall accuracy is 90% with 10% misclassification rate of 10% with class 7.

Table 1 : Confusion matrix

	1	2	3	4	5	6	7	8	9	10
1	100	0	0	0	0	0	0	0	0	0
2	8	92	0	0	0	0	0	0	0	0
3	0	0	83.33	16.67	0	0	0	0	0	0
4	0	0	0	90	0	0	10	0	0	0
5	0	0	0	0	93	7	0	0	0	0
6	0	0	0	0	8	92	0	0	0	0
7	0	0	0	0	0	0	100	0	0	0
8	0	0	0	0	0	0	0	100	0	0
9	16.67	0	0	0	0	0	0	0	83.33	0
10	0	0	0	0	0	0	0	0	0	100

For Class 5 and Class 6 again the accuracy is 93% and 92% respectively. From the misclassification rate it is visible that when Class 5 was classified with other classes the misclassification rate for all the classes is 0% except for Class 6 having false alarm as 7%. For Class 6 also the misclassification rate is 0% with all other given classes except class 5 having an error rate of 8%. Finally for the remaining classes the accuracy is 100% such for class 7, class 8 and class 10 except for class 9 having an identification rate as 83.33% which is again very less as compare to other given classes accuracy.

In the last for overall accuracy all the diagonal percentage are total together and the sum is divided by the total number of classes to get the overall performance of the algorithms in terms of accuracy rate. In this research ten classes were used and after doing the calculation for the system accuracy the accuracy achieved was 93.6%. From the attain accuracy rate it is believe that the proposed application is feasible for hand gesture recognition.

4. CONCLUSION AND FUTURE WORK

This paper has represented a feasible method for hand gesture recognition using MFCC. In this work the given input are converted from 2D Images to 1D signal to be given as input to Mel frequency ceptral coefficients. After getting the first 13 MFCCs the extracted feature vectors are classified against SVM. From the resultant confusion matrix it is visible that the MFCC can be used as a feature extraction technique while working with images just like other available techniques and also shown a new application for MFCC which is always used for the voice based processing such as speaker identification, voice recognition, gender identification using the voice and recently in bio medical too to diagnosis the baby through its voice while crying.

This time the experiment was done only on 10 gestures may be in future the experiment can done on ASL database. But on the same side the misclassification rate is high between many classes. If the MFCC can be combine with other technologies may be this misclassification rate can be reduced and MFCC can be used in image processing like other techniques. Already from the previous study it

is clear that before this already MFCC was tried for palm recognition, face recognition and for satellite image recognition with very good accuracy. So if more emphasis will be given on this may be MFCC can be one of the best known algorithm in image processing as well just like the way it is famous in speech recognition, speaker identification. So, in future MFCC can be used with a combination of other techniques.

REFERENCES

[1] A. Khan, et al., "Speech Recognition: Increasing Efficiency of Support Vector Machines," International Journal of Computer Applications vol. 35, dec 2011.

[2] A. S. Mehendale and M. R. Dixit, "SPEAKER IDENTIFICATION," Signal & Image Processing: An International Journal (SIPIJ), vol. 2, june 2011.

[3] L. Muda, et al., "Voice recognition algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of computing vol. 2, 2010.

[4] A. Zulfiqar, et al., "A Speaker Identification System using MFCC Features with VQ Technique " Third International Symposium on Intelligent Informartion Technology Application, 2009.T. M. Talal and A. E.-. Sayad, "Identification of Satellite Images Based on Mel Frequency Cepstral Coefficients" 2009.

[5] D. C. Gope, "Hand Gesture Interaction with Human-Computer," Global Journal of Computer Science and Technology, vol. 11, dec 2011.

[6] T. Messer, "Static hand gesture recognition," University of Fribourg.

[7] S. K. Kang, et al., "Color Based Hand and Finger Detection Technology for user interaction," presented at the International Conference on Convergence and Hybrid Information Technology, 2008.

[8] M. A. amin and H. Yan, "Sign Language Finger Alphabet Recognition from Gabor -PCA Representation of hand gestures," presented at the Proceeding of the sixth Internaional Conference on Machine Learning and Cybernetics, Hong Kong, 2007.

[9] Chen, et al., "Hand gesture recognition using Haar-like features and a stochastic context-free grammer," IEEE Transactions on Instrumentation and Measurement vol. 57, p. 9, 2008.

[10] D. C. Gope, "Hand Gesture Interaction with Human-Computer," Global Journal of Computer Science and Technology, vol. 11, dec 2011.

[11] D.-Y. Huang, et al., "Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination " Expert Systems With Applications, vol. 38, p. 12, 2011.

[12] S. Padam and K.Prabin.Bora, "A Study on Static Hand Gesture Recognition using Moments," presented at the International Conference on Signal Processing and Communications (SPCOM), 2010

[13] J. J. Stephan and S. a. Khudayer, "Gesture recognition for Human Computer Interaction" International Journal of Advancements in computing Technology, vol. 2, 4 November 2010.

[14] K. Symeonidis, "Hand Gesture Recognition Using Neural Networks," Centre for Vision, Speech and Signal ProcessingAugust 23, 2000.

[15] T. M. Talal and A. E.-. Sayad, "Identification of Satellite Images Based on Mel Frequency Cepstral Coefficients " 2009.

[16] Sangeeta Biswas" MFCC based Face Identification" Titech Japan.

[17] M. M. M. Fahmy, "Palmprint recognition based on Mel frequency Cepstral coefficients feature extraction," Ain Shams Engineering Journal, p. 9, 2010.

[18] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman,"Speaker identification using Mel Frequency Cepstral coefficients".

[19] V. Tiwari, "MFCC and its applications in speaker recognition," International Journal on Emerging Technologies, 2010.

[20] S. Khan, Mohd Rafibullslam, M. Faizul, D. Doll, "Speaker recognition using MFCC" IJCSES (International Journal of Computer Science and Engineering System) 2(1): 2008.

[21] Mohd Rasheedur Hassan, Mustafa Zamil, Mohd Bolam Khabsani, Mohd Saifur Rehman " Speaker identification using MFCC coefficients " 3rd international conference on electrical and computer engineering (ICECE), (2004).