Enhanced Feature Extraction Method for Hand Gesture Recognition using Support Vector Machine

E.M.P.S.Edirisinghe, P.W.G.D.Shaminda, I.D.T.Prabash, N.S.Hettiarachchige , L.Seneviratne and U.A.A.Niroshika

Abstract- **In this paper, a method is proposed to maximize the accuracy during the feature extraction stage in a real time system for hand gesture recognition by escalating the number of parameters of the feature set for support vector machine. Numerous former researches utilized hu moments but they didn't correspond to the complete description of an image, and was suitable only for giving very rough estimation of possible match. Thus matching performance was not acceptable for image retrieval. On the other hand, the accuracy of the support vector machine (SVM) depends on the number of support vectors. Hence adding features that significantly improve the splitting probability of training images decrease the number of support vectors and improves the performance of the SVM. Therefore to enhance the harmonizing of images, together with hu moments, edge histogram descriptor and circularity shape parameter is used to compose the feature vector. Experiments on series of test images show that the proposed method yields better matching performance. Integrated feature based approach to hand gesture recognition has been tested over 23 gestures and it gave promising results.**

Index Terms—feature extraction, gesture recognition, support vector machines, hu moments, circularity, edge histogram descriptor

I. INTRODUCTION

Gesture recognition is extensively employed in applications including human machine interaction, sign language and immersive game technology etc. Since late decades areas of Image Processing and Computer Vision have been predominantly developed aligning with the advancement of gesture recognition in order to achieve an ample acknowledgment in translating sign language into text/voice. Similar kinds of systems reveal numerous efforts to develop good techniques and features using image processing.

One such method is matching input gestures with a known gesture in the database with the aid of Least Mean Squares method [1] in order to reduce ambiguity.

Kaushik [2] and his fellows have used template matching with normalized cross-correlation along with two wrist bands to identify and extract the gesture. Kinect sensor which is a skeletal tracker and support vector machine technique was also used to predict the American Sign Language signs which were shown by the user, using the color and depth information in order to identify the shape of the hand but it is capable of detecting only a single hand [3]. Finger-Earth Mover's Distance (FEMD) which is a novel shape distance metric was used in its gesture recognition process. Of the many techniques so far devised, all most every system is extremely cumbersome and complex and has brought forth motivation to develop simple and manageable vision-based approaches to identify gestures.

This paper addresses the usage of gesture recognition in a sign language translating system developed by the research team which significantly facilitates non verbal communication via hand gestures. The sign language performs a major role in conveying meaning using a composite way of communication via expressions and body language instead of acoustically put across sounds. The proposed methodology is tested in recognizing British Sign Language alphabet (BSL) which is primarily interpreted by the system. The sequence of input images captured by the image capturing device is gone under pre-processing steps for background noise removal (Background Subtraction) which utilizes a novel hybrid method. Subsequent to noise removal the hand object is segmented and ultimately used for recognizing gestures. Gesture recognition can be considered in two phases namely feature extraction and classification. All features necessary to match and differentiate between gestures has to be preserved during the feature extraction phase. If values are not strong enough to grant accurate measures, performance of the support vector machine is intrinsically crippled during the classification phase. Therefore the proposed framework uses an improved feature set against the former research techniques that have only used less number of features.

II. METHODOLOGY

Main components of the proposed framework are shown in the Fig. 1. The stages can be declared as background subtraction and extracting of the hand region, shot segmentation, feature extraction and classification.

E.M.P.S.Edirisinghe, P.W.G.D.Shaminda, I.D.T.Prabash, N.S.Hettiarachchige , L.Seneviratne and U.A.A.Niroshika are with the Department of Information Technology, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka (e-mail: puuj1989@gmail.com, dlnshaminda@gmail.com, tharindutpd@gmail.com, thrani1989@gmail.com, lasantha.s@sliit.lk and aruni.n@sliit.lk).

978-1-4799-0910-0/13/$31.00 ©2013 IEEE

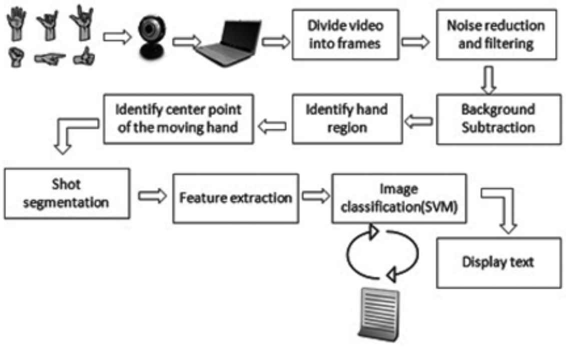

Fig. 1. Main stages of the proposed methodology

A hybrid approach is used in background subtraction and the descriptors obtained in the feature extraction stage are used as input for the SVM classifier.

A. Background subtraction and extracting the hand region

The initial approach to extract hand region is based on an enhanced background subtraction technique. An image capturing device is used to capture the hand movements of the user where all the captured frames are in RGB color format. Other than the interested area which is the hand region, captured images furthermore consists of foreground and background objects. An ideal background subtraction technique capable of eliminating unwanted regions and extracting the hand region precisely is essential for further processing.

Therefore a novel method is developed combining skin color pixel detection and motion tracking. Skin color detection algorithms are used to detect human body and the primitive algorithm which assures this requirement is the HSV lookup, where H is for hue , S is for saturation and V is for value. First classification of a pixel, as a skin component or not, is done according to following definition and the pixel is converted to its corresponding HSV value as given below [4],

$$\left.\begin{array}{l} H_{min} = 0 \\ H_{max}=50 \end{array}\right\} \longrightarrow \quad H_{min} \leq H \leq H_{max}$$

$$\left.\begin{array}{l} S_{min} = 0.23 \\ S_{max}=0.68 \end{array}\right\} \longrightarrow \quad S_{min} \leq S \leq S_{max}$$

If a pixel is not identified as a skin component then it is coloured white. The resultant output is later used as the input to the OR operation, which is then integrated and processed with motion tracking output. Temporal difference of frames, which is the difference of two consecutive frames, is taken initially for motion tracking and it provides a thin contour of the moving hand. At this point all background objects are eliminated along with most of the noise points. The resulting frame which appears with the moving hand region is then eroded to remove the rest of the noise points. Since the resulting contour is very thin the frame is subsequently dilated in order to get a clearer moving region.

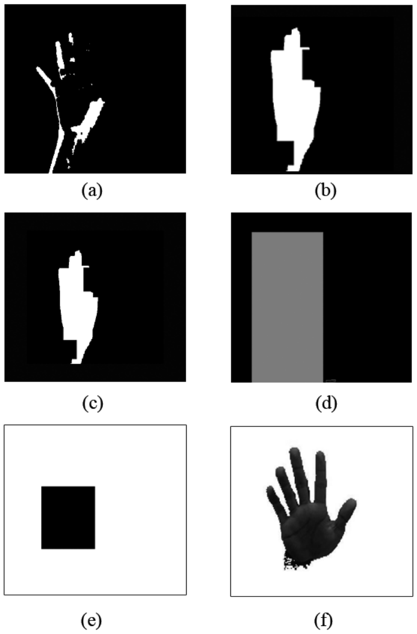

Fig.2. Motion tracking process where (a) Difference of two consecutive frames, (b) Frame after dilation and Erosion, (c) Convex hull drawing for filling and connecting small components, (d) Bounding rectangle drawing , (e) Mask for OR operation, and (f) OR operation with skin color detection frame

Fig. 2, briefly explains the motion tracking process and Fig. 3, shows the results of image extraction. Contour detection and convex hull techniques are used to fill and merge the dilated contours and the bounding rectangle is drawn around the merged contour. As the concerned region is the palm area of the hand, a threshold value is given to disregard diminutive bounding rectangles. Center points of the two biggest contours are assumed as corresponding to the center points of two hands of the user and a mask is created for OR operation in conjunction with the previously detected skin color area frame. The end result grants the skin color components within the motion area, which is the interested area or more precisely the hand region.

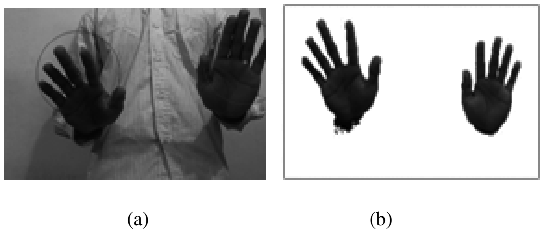

Fig. 3.Extracted hand region where (a) Original Image, (b) Skin colour pixels

B. Shot Segmentation

Since the signs are quickly and continuously changing from one to another it is extremely hard to distinguish between the start and end points of the sign in a real-time system. Therefore shot segmentation refers to the process of breaking down the sentence to the words within the context of the system and through this process system is able to identify the start and the end point of a particular sign. The shot segmentation process moreover includes the excluding of the hand movement while changing from one sign to another sign.

In order to carry out the shot segmentation following rule set is considered. Gestures are contained in movements that start with a slow initial move from the rest position, continue with a phase with substantially increased speed (the stroke), and end by returning to the rest position. The hand assumes a particular configuration during the stroke. Slow motions between resting positions are not gestures. Static hand gestures require a finite period of time to be recognized. Repetitive movements can be gestures [5].

Discontinuities are detected to determine end points in a sequence of gesture inputs for segmentation. The discontinuity detection is done by time-varying parameter (TVP) detection [6]. As for the system speed variation of the hand movement is taken as the TVP and the Fig. 4, illustrates the possible speed variation for the static gestures. However considering only the speed variation is not adequate to realize the specific functionality. Specific signs which change only the postures but don't change the position can be identified within the context. If merely the speed variation is considered in such a scenario where the user enters two signs continuously the system recognizes these two signs as one sign since within the particular time slot speed variation is zero. In order to address this issue, the shot segmentation is extended as a hybrid of speed variation and fingertip detection.

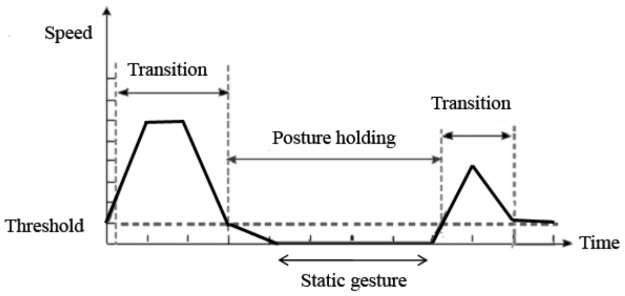

Fig. 4. Speed variations for static gestures

C. Feature Extraction

During the feature extraction stage a feature vector is composed integrating hu moments, edge histogram descriptor and circularity shape parameter values. Moment invariants were firstly introduced by Hu [7]. Hu derived six absolute orthogonal invariants and one skew orthogonal invariant based upon algebraic invariants, which are not only independent of position, size and orientation but also independent of parallel projection. The moment invariants have been proved to be the adequate measures for tracing image patterns regarding the image translation, scaling and rotation under the assumption that the images are with continuous functions and also noise-free. Therefore after the background and noise removal seven rotation and scale-invariant hu moments for each training image is computed. This seven-number code serves as a brief description of an image. However, this approach of considering only the hu moment values definitely suffers from obvious drawbacks. Seven hu moments by no means give a complete description of an image, hence are suitable only for giving very rough estimation of possible match [8]. When taking above mentioned particulars into consideration hu moments itself, is not sufficient for a very much accurate classification using support vector machine.

Accuracy of the support vector machine can be referred in terms of the computational complexity of the model and it is linear to the number of support vectors. Therefore fewer support vectors implies faster classification of test points. Adding a feature that significantly improves separating prospects of training images, decrease the number of support vectors which on the other hand improves the performance. Therefore texture based feature extraction is integrated with the hu moment values. Texture is identified as a dominant low-level descriptor and edge histogram (EHD) is one of the three texture descriptors defined by MPEG-7 Visual. The MPEG-7 Visual Standard allows user to measure similarity in images or videos based on visual decisive factors. It indicates the spatial distribution of four directional edges and one non-directional edge which construct five types of edges in local image regions [9].

The edge histogram is regarded as extremely valuable for indexing and retrieving images since it exploits the characteristics, such as translation and rotation invariance. The normative edge histogram for MPEG-7 is designed to contain only local edge distribution with 80 bins. These 80 histogram bins are the only standardized semantics for the MPEG-7 edge histogram descriptor [10]. MPEG-7 low level color and texture descriptors were extracted from images using the MPEG-7 Low Level Feature Extraction Library. It uses OpenCV2.0 and includes DLLs in the archive for convenience. The command line arguments and the procedure required in extracting the edge histogram values can be comprehended through referring to the archive [11].

In addition to above features, a geometric feature called circularity is used to enhance the accuracy of the SVM furthermore. Circularity is the measure of the shape that how much the object's shape is closer to the circle. In the ideal case, circle gives the circularity as one. The range of circularity varies from 1 to infinity [12]. In unison, feature set is composed with 88 features and used by the SVM in the classification stage.

D. Classification

LIBSVM is deemed as an integrated software for support vector classification and the package includes the library in C++. Therefore its 3.17 version is used in classification phase [13]. LIBSVM, which is an open source machine learning library, is used according to a defined procedure [14].

Step 1: Data Preparation for SVM
Step 2: Convert data into SVM format
Step 3: Conduct simple scaling on the data
Step 4: Consider the RBF kernel $K(x; y) = e$
Step 5: Use cross-validation to find the best parameter C and gamma
Step 6: Use the best parameter C and gamma to train the whole training set

Initially data is arranged for SVM via capturing binary images from three different people. Since the data set should contain both positive and negative data both positive and negative images are utilized in calculating feature values for each image and is written into a text file according to the LIBSVM format which accepts only numerical values. Then scaling is conducted on the text file with the feature values as recommend in linearly scaling, each attribute to the range [0; 1]. The original figures maybe too vast or undersized in range, thus rescaling accelerate training and predicting speed. Avoiding attributes in greater numeric ranges which dominates smaller numeric ranges is the main advantage of scaling and meanwhile it also avoids numerical complexities during the computation.

The output of scaling is used for creating the model and the scaling parameters are saved to the range file. Then it is used for scaling the test data since the same range must be applied in scaling both training and testing data and much better accuracy can be obtained. After scaling the data set, linear kernel function which was chosen for creating the model is utilized in order to attain a higher accuracy as the training set contains large number of instances and testing set contains only one instance. Cross validation is performed to achieve the parameters, C and gamma for the kernel so as to get the best cross validation accuracy. Finally for the testing, the training data set is trained using C value as the linear kernel accepts only one parameter.

III. RESULTS AND DISCUSSION

The Gesture recognition database consists of 2300 images representing 23 signs of BSL alphabet and images are of 640×480 pixels dimension in the form of binary. Since the proposed model works with static signs, H and J are not trained. H and J are dynamic as they enclose motions. Fig. 5, shows a number of instances of Gesture recognition database.

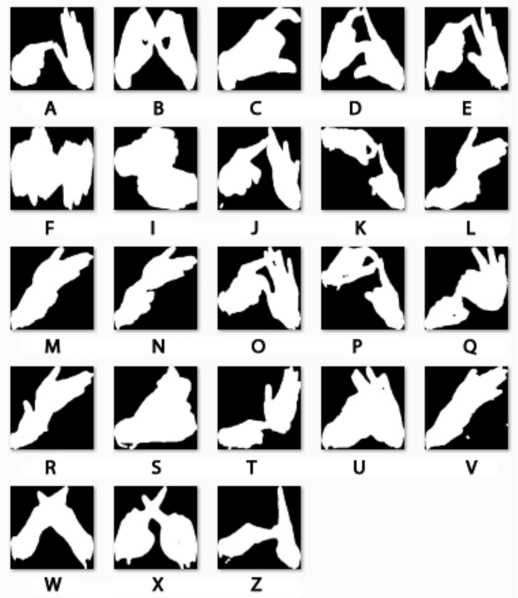

Fig. 5. Instances of gesture database

The theoretical values of the accuracy were calculated for each sign using sensitivity and specificity statistical classification. It statistically computes the performance of the experiment and is usually used in Binary classification Tests. In a binary classification, the given data set is divided into two categories on the basis of whether they have common properties or not by identifying their significance [16].

Table I below is the evidence for test results gained using above stated statistical method. The sensitivity indicates how likely the test will grant a positive response for a specimen with the characteristic. According to this data set which is listed in Table I, average of 98.55% sensitivity can be observed. As an example, if the user inputs sign "A", then there is an average of 98.55% possibility to identify it as sign "A". The specificity indicates how likely the test will grant a negative response for a specimen without the certain characteristic. Table I indicates an average of 71.49% specificity. As an example, if the user inputs a sign which is not "B", then there is an average of 71.49% possibility to result that it is not sign "B". According to the above estimates, average accuracy for the entire system is 75.17%.

True positive indicate correctly identified number, false positive indicates incorrectly identified number and true negative indicates correctly rejected number while false negative signifies incorrectly rejected number. Some complex gestures produce weak accuracy results as in "C" and "G" characters. Table I confirms that they result low accuracy values (52%-54%) compared to the average accuracy level. As the trained SVM machine has very high false positive (FP) values and low true negative (TN) values it tend to ignore the actual sign. Figure 6 below shows Accuracy measures in graphical form.

TABLE I

SPECIFICITY AND SENSITIVITYSTATISTICAL RESULTS FOR MEASURING ACCURACY

Letter	TP	FN	FP	TN	Sensitivity	Specificity	PPV	NPV	Accuracy
A	12/12	0/12	14/88	74/88	100%	84.09%	46.15%	100%	86%
B	12/12	0/12	11/88	77/88	100%	87.5%	52.17%	100%	89%
C	12/12	0/12	46/88	42/88	100%	47.73%	20.68%	100%	54%
D	12/12	0/12	19/88	69/88	100%	78.41%	38.71%	100%	81%
E	12/12	0/12	8/88	80/88	100%	90.91%	60%	100%	92%
F	12/12	0/12	2/88	86/88	100%	97.73%	85.74%	100%	98%
G	12/12	0/12	48/88	40/88	100%	45.45%	20%	100%	52%
I	8/12	4/12	8/88	80/88	66.67%	90.91%	50%	95.2%	88%
K	12/12	0/12	13/88	75/88	100%	85.23%	48%	100%	87%
L	12/12	0/12	15/88	73/88	100%	82.95%	44.44%	100%	85%
M	12/12	0/12	50/88	30/88	100%	34.09%	19.35%	100%	52%
N	12/12	0/12	17/88	71/88	100%	80.68%	41.38%	100%	83%
O	12/12	0/12	20/88	68/88	100%	77.27%	37.5%	100%	80%
P	12/12	0/12	13/88	75/88	100%	85.23%	48%	100%	87%
Q	12/12	0/12	18/88	70/88	100%	79.55%	40%	100%	82%
R	12/12	0/12	42/88	46/88	100%	52.27%	22.22%	100%	58%
S	12/12	0/12	57/88	31/88	100%	35.23%	17.39%	100%	43%
T	12/12	0/12	12/88	76/88	100%	86.36%	50%	100%	88%
U	12/12	0/12	45/88	43/88	100%	48.86%	21.05%	100%	55%
V	12/12	0/12	6/88	82/88	100%	93.18%	66.66%	100%	94%
W	12/12	0/12	30/88	58/88	100%	65.91%	28.57%	100%	70%
X	12/12	0/12	60/88	28/88	100%	31.82%	16.67%	100%	40%
Z	12/12	0/12	15/88	73/88	100%	82.95%	44.44%	100%	85%
					98.55%	71.49%	39.96%	99.7%	75.17%

Existing literature which elaborates on standard techniques of feature extraction such as artificial neural networks, Hidden Markov Model has a less recognition rate when compared with the proposed framework. Such experiment for the recognition of Brazilian Sign Language using distance and local alignment based approach gives the average accuracy of 75% only [15].

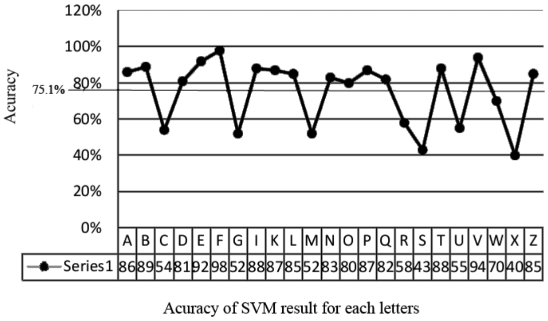

Fig.6. Accuracy measures in graphical form

IV. CONCLUSION

As BSL exercises both left and right hands in showing gestures, it is more complex when weigh against other sign languages. In this paper, a model is proposed to segment the hand region from a noisy background and to compute an integrated feature vector consisting of hu moments, EHD values and circularity parameter in order to maximize the accuracy of support vector machine in the classification phase. Additionally an innovative hybrid method is developed combining skin color pixel detection and motion tracking for noise removal and segmentation of the hand region.

A gesture shown by the user is identified correctly with an average of 98.55% possibility while there is an average of 71.49% possibility to result that it is not any other sign. As a number of complex signs blend with each other and provide low accuracy, the average accuracy for the entire system has dropped to 75.17% which is still a promising result.

As future work, the generic approach proposed can be used to train dynamic gestures effortlessly. Moreover, consideration is given to replace binary images with gray scale in order to improve the performance.

REFERENCES

[1] A. K. Alvi et al.,“Pakistan Sign Language Recognition Using Statistical Template Matching.” presented at the *International journal of information technology* 2004, Jan.

[2] K. Deb, M.I. Khan, H.P. Mony, and S. Chowdhury. “Two-Handed Sign Language Recognition for Bangla Character Using Normalized Cross Correlation,” *Global Journal of Computer Science and Technology*. vol. 12, no. 3, Feb.. 2012. [On-line]. Available: http://computerresearch.org/stpr/index.php/gjcst/article/viewArticle/1005 [Accessed Feb. 10, 2013].

[3] Z. Ren, J. Meng, J. Yuan and Z. Zhang. “Hand Gesture Recognition with Kinect Sensor,” Nanyang Technological University, USA, 2011,Jan.

[4] C. Chouse. “Skin Color Detection with HSV Lookup” 2009, [on-line]. Available: http://www.chasanc.com/content/view/65/98/ [Accessed Aug. 12, 2013].

[5] BogdanIonescu et al,. “Dynamic Hand Gesture Recognition Using the Skeleton of the Hand,”.*EURASIP Journal on Applied Signal Processing*. 2005, pp. 2101-2109.

[6] R. Liang and M. Ouhyoung. “A Real-time Continuous Gesture Recognition System for Sign Language,” *presented at the IEEE International Conference on Automatic Face And Gesture Recognition japan. 1998, pp 558-565*

[7] Z. Huang and J. Leng, “Analysis of Hu’s Moment Invariants on Image Scaling and Rotation,” *in Proc. 2nd International Conference on Computer Engineering and Technology (ICCET)*, Chengdu China, 2010, pp. 476-480

[8] O. Busaryev and J. Doolittle, “Gesture Recognition with Applications,” Dept.Computer Science and Eng., Ohio State Univ.,Columbus,Class Project Rep., 2010

[9] M. Bober and S. Paschalakis. “MPEG-7 Visual.”, 2011 [Online]. Available: http://mpeg.chiariglione.org/standards/mpeg-7/visual [Accessed Mar. 12, 2013].

[10] S. J. Park, D. K. Park and C. S. Won, “Efficient Use of Local Edge Histogram Descriptor,” MPEG7 document M5984, Korea, May, 2009

[11] M. Baştan. “MPEG-7 Feature Extraction Library” [Online]. Available: http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo-7/Software.html [Accessed Apr. 12, 2013].

[12] M. Elmezain et al.,"Posture and Gesture Recognition for Human-Computer Interaction," 2009 August [online] Available: http://www.intechopen.com/books/advanced-technologies/posture-and-gesture-recognition-for-human-computer-Interaction [Accessed Jun. 12, 2013].

[13] Lekshmi D. LIBSVM tutorial, 2012, February. [Online]. Available: http://lekshmideepu.blogspot.com/2012/02/libsvm-tutorial.html9 [Accessed Jun. 10, 2013].

[14] C. Chang and C. Lin. A Library for Support Vector Machines, 2011, April 01 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[15] Teodoro, Beatriz, and Luciano Antonio Digiampietri. "A Local Alignment Based Sign Language Recognition System."

[16] Aswathi B.L. “Sensitivity, Specificity, Accuracy and the relationship between them,” 2009 March. (1st edition)[Online]. Available: http://www.lifenscience.com/bioinformatics/sensitivity-specificity-accuracy-and [Accessed Aug. 19, 2013].