Indian Journal of Science and Technology, Vol 9(25), DOI: 10.17485/ijst/2016/v9i25/96615, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Monocular Vision-based Signer-Independent Pakistani Sign Language Recognition System using Supervised Learning

Habib Ahmed*, Syed Omer Gilani, Mohsin Jamil, Yasar Ayaz and Syed Irtiza Ali Shah

School of Mechanical and Manufacturing Engineering (SMME), National University of Sciences and Technology (NUST), H-12 Main Campus, Islamabad, Pakistan;habib_ahmed@msn.com, mohsin@smme.nust.edu.pk, irtiza@smme.nust.edu.pk

Abstract

Background/Objectives: To construct a Pakistani sign language learning-based gesture recognition system with a reasonable rate of accuracy. **Methods/Statistical Analysis**: It should be <70 words. Include the method adapted to study the objectives/sampling details or simulation or statistical analysis of data; technique employed; mention unique/ important points of modification of methodology in the current study. Mention about test samples the control employed or approach used for comparing the test sample. **Findings**: The proposed system uses static images to extract local and global, region and boundary-based descriptors for acquiring gesture information, which is provided as input for supervised learning method known as Support Vector Machine (SVM). The purpose of this research is to formally introduce a practical learning-based PSL recognition system, which can lay the groundwork for future research pertaining to PSL. The proposed system was developed and the ten class supervised learning based system was able to achieve an accuracy of 83%. **Application/Improvements**: It is a preliminary work, which will be further improved to construct a real-time static and dynamic gesture based PSL system that is able to recognize words and sentences information.

Keywords: Fourier Descriptors, Gesture recognition, Hu Moments, Sign Language Recognition, Support Vector Machines

1. Introduction

Gesture recognition continues to play an important part within research, with widespread practical applications and potential future uses within the academia and society at large. Sign language recognition is a single research focus within a larger research area of gesture recognition, which also includes the following: (i) augmented reality[1,2], (ii) human-robot interaction (HRI)[3,4], (iii) human-computer interaction (HCI)[5,6], (iv) industrial and domestic applications[7,8]. Gestures can be classified as either static (one hand is used without movement) or dynamic (involves both hands, their motion as well as other body language and facial cues) in nature. The literature within sign language recognition systems employ many different types of sensors that include single camera[9], multiple cameras[10], Microsoft® Kinect[11], sensor gloves[12] and surface-mounted electromyography (sEMG) sensors[13]. Each sensor has varying properties and therefore, it is difficult to cross-examine all the systems built using different vision and non-vision-based sensors. Some of the widely-practiced regional variants of the sign languages such as American Sign Language (ASL), Chinese Sign Language (CSL) and British Sign Language (BSL) have received considerable attention within the sign language research area.

Some of the earliest researches on sign language recognition relied on color cameras to segment gesture information from the background[15]. However, accurate automatic real-time recognition of sign language remains a challenging task using traditional color cameras, owing to the lack of depth information, noise and difficulty in clean segmentation of the region-of-interest from varying backgrounds. However, the advent of low-cost sensors such as Microsoft® Kinect allowed development

*Author for correspondence

Figure 1. Ten Urdu language alphabet gestures for PSL.

of much more efficient sign language recognition systems at reduced costs. Unlike conventional color cameras, Kinect provides color data, depth map and skeletal image containing 3D position information of signers' different important joints[16]. Research by Sun et al.[17] proposed exemplar coding based approach for ASL recognition using Kinect sensor to provide maximum accuracy of 86.8% for around 2,000 different phrases. The combined use of accelerometers and sEMGs has been devised as another approach towards sign language recognition with promising results[13]. This system was able to achieve gesture recognition accuracy of more than 90% for 100+ different signs from CSL using shape, orientation and motion information from different sensors[13].

This study proposes and develops a framework for PSL static alphabet recognition system based on supervised learning technique. Other sign languages such as Chinese Sign Language, BSL and ASL, have pre-existing database containing words and sentence samples, there is no existing database for PSL alphabets and words. As a result, the data had to be collected and database had to be maintained for completing this research. The proceeding sections of the paper have been divided into four parts. Section 2 outlines the various salient features of the proposed framework for PSL alphabet recognition. While, section 3 highlights the salient features of the experiments conducted towards practical implementation of the system and within section 4, experimental results and the overall effectiveness of the system is discussed. Finally, section 5 concludes the findings of the research and possible future research with regards to developing a complete PSL recognition system.

2. System Framework

The overview of the developed system have been outlined in Figure 2 in which the different modules of the system have been given independently. Some of the core components of the system include image pre-processing module, feature extraction module and machine learning module. Within figure 2, the system shows input images on the left-hand side of the figure, being processed using different image processing techniques such as image resizing, de-noising, ROI segmentation using HSV color space, and morphological operations to reduce error and segmentation noise. The processed image from the first module is given to the second module as input for extracting relevant features. These features, once extracted are provided to the one-against-all SVM model (constructed with the help of learning using features from the database images) for classification into the appropriate gesture class. The details of the different modules and their internal workings have been elaborated in the proceeding subsections.

2.1 Pre-processing Module

Within this module, the primary functions being conducted include segmentation and morphological operations. Segmentation allows the separation of the background from the foreground or Region-Of-Interest (ROI). Segmentation in itself is very complex and context-specific in nature, while for an image segmentation method to be considered successful, the ROI should be homogenous and uniform in comparison with some fixed reference texture or color tone[18]. Skin segmentation can be considered as a region-based segmentation technique, as it differentiates between foreground and background based on the skin color. Particularly concerning applications focusing on human beings, skin segmentation has been extensively used within gesture and facial recognition. Many of the color spaces have been used in this regard (RGB, normalized-RGB, HSV, HSI, YCbCr,

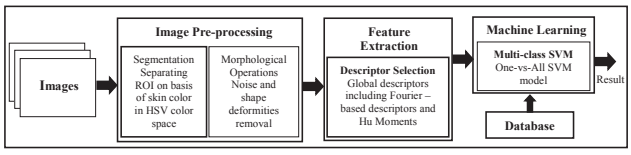

Figure 2. Proposed system framework for PSL recognition.

YUV, and YIQ) and their overall performance has been examined by many different studies[19–21]. Within this module, skin segmentation has been carried out by converting from RGB to HSV color space and using global thresholding to separate skin pixels from other regions not containing skin pixels. This technique provided the most reliable results, when compared to using YCbCr and RGB color space-based thresholding. The threshold used within the proposed system has been outlined below:

$$H < 0.25 \quad (1)$$

$$0.9 > S > 0.15 \quad (2)$$

Although, different sources have provided varying margins for HSV thresholds, but actual implementation of the HSV color segmentation required manually testing and setting the threshold, based on optimum performance under the given conditions. The morphological operations were used to ensure that the slight defects in obtaining ROI could be removed, as over-segmentation or under-segmentation can easily corrupt the desired results[22]. Practically, it is important to ensure that extracted region is homogenous and without any irregular breaks and distortions in the contour. The operations performed included filling holes within the selected region and using morphological functions for closing and removing redundant regions.

2.2 Feature Extraction Module

Feature extraction is another critical aspect of the classification problem, as it extracts relevant information from images, which can be used by the learning system for observing inter-class trends and intra-class variations to differentiate and classify information highlighted within the images. Within this research, a host of different global features were used, which include region- and boundary-based features. Some of the simple global features include length, area, rectangularity, eccentricity, convexity, solidity, circularity ratio and basic rectangle (minimum and maximum axis). Global features are computed based on the global similarities in patterns, regions and groups of neighboring points. Some of the primary global shape features used by this system include Fourier Descriptors and Hu's invariant moments. Region segmentation allow the utilization of region-based (such as Hu Moments) and boundary-based features (such as Fourier Descriptors) simultaneously. Fourier descriptors have been extensively employed within the field of object classification, retrieval and recognition, making it extremely useful and popular[23–25]. Mathematically, the discrete Fourier Transform can be represented in the following manner:

$$f_x = \frac{1}{N}\sum_{t=0}^{N-1} s(t)e^{\frac{-j2\pi xt}{N}}, x = 0, 1, \ldots N-1 \quad (3)$$

In the above equation, $s(t)$ is the equation of the shape signature; a one-dimensional representation of the boundary of the region-of-interest. Shape boundary signatures can be represented in different forms that include Complex Coordinates, Chord Length Signature, Curvature Signature, Centroid Distance Function and Area Function[23]. Fourier Descriptors can be reconfigured to ensure scale, rotation and starting point invariance, which allows the computed descriptors to be robust and efficient. Hu[26] proposed a set of seven moments from the normalized unscaled central moments that demonstrate invariance to translation, rotation and scaling. As a result of these properties, they can be used for shape classification and representation in conjunction with other shape feature descriptors. The seven moments outlined by Hu[26] are given in equation (4), where m_{AB} refers to unscaled central moments (while the order of the moment is $A^{+}B$):

$$\begin{aligned}
&Hu_1 = m_{20} + m_{02}, \qquad Hu_2 = [\![(m]\!]_{20} - m_{02})^2 + 4m_{11}{}^2 \\
&Hu_3 = [\![(m)]\!]_{30} - 3m_{12})^2 + (3m_{21} - m_{03})^2 \\
&Hu_4 = [\![(m]\!]_{30} + m_{12})^2 + (m_{21} + m_{03})^2 \\
&Hu_5 = [\![(m]\!]_{30} - 3m_{12})^2 (m_{30} + m_{12})[(m_{30} + m_{12})^2 - 3(m_{21} + m_{03})^2] \\
&\quad + (3m_{21} - m_{02})(m_{21} + m_{03})[3(m_{30} + m_{03})^2 - (m_{30} + m_{12})^2) \\
&Hu_6 = [\![(m]\!]_{20} - m_{02})[(\mathrm{m}_{30} + m_{12})^2 - (m_{21} + m_{03})^2] + 4m_{11}[\![(m]\!]_{30}\, m_{12})[\![(m]\!]_{21} + m_{03}) \\
&Hu_7 = [\![(3m]\!]_{21} - m_{03})[\![(m]\!]_{30} + m_{12})[(m_{30} + m_{12})^2 - 3(m_{21} + m_{03})^2 \\
&\quad - (m_{03} - 3m_{12})(m_{21} + m_{03})[3(m_{30} + m_{12})^2 - (m_{21} + m_{03})^2)
\end{aligned} \quad (4)$$

2.3 Machine Learning Module

Supervised learning was employed within this module, which required maintaining a database of images and their respective feature vectors in order to train the system to recognize the individual gestures effectively. Earlier studies have considerably relied on Hidden Markov Models (HMMs)[15,27], from the earliest stages of gesture recognition as an academic research subject. Relatively recently, different variations of neural networks[14,28] have also been used within some of the researches. This system uses a one-against-all multi-class Support Vector Machine (SVM) classifier for ten alphabet gesture classes of PSL. Originally, SVM has been implemented for binary classification, i.e. it can only differentiate between two classes[29]. For binary SVM with two classes (a_1, a_2) and training data

($\chi = \{x_1, x_2, \ldots x_L\} \subset \Re^p$), the SVM classifier will separate the input vectors into a high dimensional feature space F, which is able to separate the two classes into two separate hyperplanes (P1 and P2) by computing a linear function within that feature plane. The SVM classification rule for the binary case is given below:

$$y_k = \begin{cases} +1 & x_k \in a_1 \\ -1 & x_k \in a_2 \end{cases} \quad (5)$$

The transformation of the input vector into the feature space is accomplished using eq. (6), (where $\varphi: \Re^p \rightarrow \mathfrak{F}$ and $w \in \mathfrak{F}$, while the (.) in eq. (6) denotes dot product) and this equation is constrained by a condition, which is highlighted in eq. (7)[30]. Consequently, the two hyperplanes in F are given as P_1: $w \cdot \varphi(x_k) + b = +1$ and P_2: $w \cdot \varphi(x_k) + b = -1$[31].

$$f(x) = (w \cdot \varphi(x)) + b \quad (6)$$

$$y_k (w \cdot \varphi(x)) + b - 1 \geq 0 \quad (7)$$

The multi-class variants of binary SVM have been developed and being used successfully, such as: (i) one-against-one, (ii) one-against-all, and (iii) Directed-Acyclic Graph-based SVM (DAGSVM) techniques11. Within one-against-all SVM, all the input vector data is transformed into a single large optimized formulation having L separate SVM models (L is total the number of classes) and one-against-one SVM reconfigures a single multi-class problem into multiple binary problems (total number of SVM models are given as: $\frac{L(L-1)}{2}$) using data from two classes within each SVM model32. The SVM one-against-all model has been given below:

Figure 3, given above provides an internal model-based representation of one-against-all SVM recognition system. Within one-against-all SVM model having a total of L SVM models, the kth classifier model is trained using data from class k as positive instances and all the other classes as negative instances. However, wrong output from any single SVM model can cause incorrect classification resulting in erroneous outcomes. As a result, the Error-Correcting Output Code (ECOC) based SVM technique was employed, as it significantly reduces error generated from redundant coding information[33]. ECOC assigns a single unique code to each class having a total of C unique codes with a Hamming distance of d and each code is H bits long. Several methods have been proposed by studies, such as using BCH codes, exhaustive codes, random codes and Hadamard-Matrix codes[33]. The SVM model takes two inputs; one of the input is the complete database containing a large array of feature information pertaining to the gesture images for each of the ten classes. This information is used for the supervised learning to construct SVM model. The second input of the learning system takes data from the feature extraction module, which contains feature information from the input image. The SVM model analyzes feature information of the input image in order to classify it into one of the ten classes for which the model has been constructed.

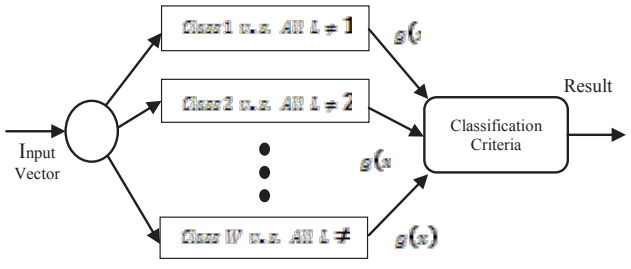

Figure 3. Internal working of one-against-all SVM-based model.

3. Experiments

This research focuses on classification of the ten different static gestures for Urdu language alphabets within PSL. The details of different gestures have been outlined within fig. 1. Hand gesture samples were taken from total of 60 different participants (35 male and 25 female) at varying lightning conditions in an indoor environment with uncluttered background using 5 megapixel camera from a standard mobile phone. As a result, the collected samples amounted to 600 images and the dimensions of individual image exceeding 1,700 × 2,600 pixels in size. Therefore, the size of the images were reduced to standard dimensions of 640 × 480 pixels, while ensuring that the aspect ratio remained unaffected in the process. The image data for each class ranges from 55 to 60. Data was processed and analyzed using MATLAB® R2015A on an Intel® Core i3 machine with 2.7 GHz maximum processing speed. For the purpose of cross-validation testing of the system, the collected data was randomly divided into training and testing portion based on the ratio of 3:2 for training and testing images respectively. The SVM model was constructed using 60% of the data for supervised learning purposes. While, the testing of the SVM model was conducted on the remaining 40% of the data in order to reveal the overall classification

accuracy of the proposed system. Figure given below provides a confusion matrix of the ten PSL gesture classes:

4. Discussion

The proposed system was successfully implemented and in order to analyze its effectiveness in terms of overall classification accuracy, the testing and training of the system was repeated for 1,000th iterations in order to satisfactorily validate the obtained results. Figure 4 provides the change in classification accuracy of the system up till the 1,000th iterations, which reveals that the overall system accuracy is equal to 83%. The overall accuracy and confusion matrix was computed after every iteration and the final result obtained after 1,000 iterations was the average of the results from successive iterations. Fig. 4 also shows that the highest error and misclassification rate from the ten classes is between Bay and Fay. One of the primary reasons for the error rate is visual similarity (refer to Figure 1 for visual information of the ten gesture classes) in the two gesture classes and the use of visual features, which ultimately culminate in the error rate within the two gesture classes. When defining gestures in terms of their visual properties, Bay can be defined in terms of all fingers facing upwards, thumb extending inwards (inside the palm) and the palm facing towards the camera. Similarly, Fay can be described in terms of palm facing the camera with all the fingers pointing upwards except index finger (bent downwards) and thumb (facing upwards) touching the index finger.

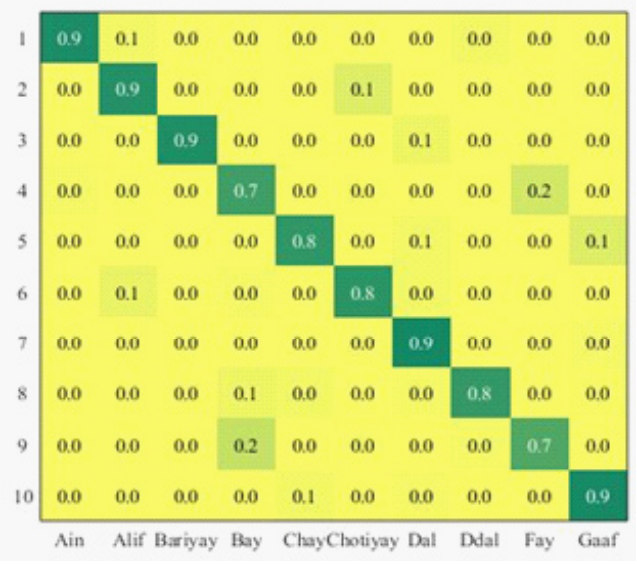

Figure 4. Confusion matrix for the ten PSL gesture classes.

5. Conclusion and Future Work

There has not been an adequate level of attention given to PSL within the realm of sign language recognition and this research attempts to bridge the widening gap between other sign languages (such as ASL and BSL) and PSL. The lack of adequate attention given to PSL within academia as a separate sign language within Pakistan is one of the reasons for developing this system. The proposed system has demonstrated promising classification accuracy of 83%, which is at par with many of the current researches related to other sign languages. The purpose of the study was to provide a benchmark for gesture recognition systems for PSL. Future studies will be conducted not only to improve the existing system, but also to provide enhancements by eradicating the practical issues within the existing system. Some of the other ways in which the current system can be improved is by reducing the error rate, increasing the overall amount of image data incorporated, number of gestures used within the system and using a mixture of static and dynamic word gestures from PSL. At the same time, gesture-based systems can be further used within the context of HRI and HCI in order to effectively contribute in terms of improvement within implementation and application from previous works.

6. References

1. Lahamy H. and Litchi, D. Real-time hand gesture recognition using range cameras. In: Proceedings of the 2010 Canadian Geomatics Conference and Symposium of Commission I, Calgary, Canada. 2010. p. 38–40.
2. Lambrecht J, Kruger J. Spatial Programming for Industrial Robots based on Gestures and Augmented Reality. Presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012. p. 466–72.
3. Correa M, Ruz-del-Solar J, Verchae Lee-Ferng J, Castillo N. Real-Time Hand Gesture Recognition for Human-Robot Interaction. Springer-Verlag: Berlin. 2009.
4. Hashimoto M, Takahashi K, Shimada M. Wheelchair control using an EOG- and EMG-based gesture interface. In: Proceedings of IEEE/ASME International Conference on Advanced Mechatronics. 2009. p. 1212–7.
5. Sohn K, Kim Y, Oh P. 2-Tier Control of a Humanoid Robot and Use of Sign Language Learned by Monte Carlo Method. Presented at The 22nd IEEE International Symposium on Robot and Human Interactive Communication. 2013; 547–52.
6. Ding J, Yu L, Wang Y, Pan Z. EasyHouse-I: A virtual house presentation system based on Internet. Presented at the

11th International Conference on Human-Computer Interaction. Las Vegas, Nevada, 2005.
7. Kela J, Korpipaa P, Mantyjarvi J, Kallio S, Savino G, Jozzo L, Marca D. Accelerometer-based gesture control for a design environment. Personal and Ubiquitous Computing. 2006; 10(5):285–99.
8. Abid MR, Melo LBS, Petriu EM. Dynamic Sign Language and Voice Recognition for Smart Home Interactive Application. Presented at the IEEE International Symposium on Medical Measurements and Applications (MeMeA). 2013; 139–44.
9. Lee S-H, Sohn M-K, Kim D-J, Kim H. Smart TV Interaction System using Face and Hand Gesture Recognition. Presented at the IEEE International Conference on Consumer Electronics. 2013; 173–4.
10. Pu Q, Gupta S, Gollakota S, Patel S. Whole-Home Gesture Recognition using Wireless Signals. Presented at the MobiCom'13. Miami, FL. 2013.
11. Hsu C-W, Lin C-J. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks. 2002; 13:415–25.
12. Padden C, Humphries T. Deaf in America: Voices from a culture. Cambridge, MA: Harvard University Press, 1988.
13. Li Y, Chen X, Zhang X, Wang K, Wang ZJ. A Sign-Component-Based Framework for Chinese Sign Language Recognition using Accelerometer and sEMG Data. IEEE Transactions on Biomedical Engineering. 2012; 59(10):2695–704.
14. Gweth Y, Plahl C, Ney H. Enhanced continuous sign language recognition using PCA and neural network features. CVPR Workshop on Gesture Recognition. 2012.
15. Starner T, Weaver J, Pentland A. Real-time American sign language recognition using desk- and wearable computer-based video. Proceedings of IEEE Transactions of Patterns Analysis and Machine Intelligence. 1998. p. 1371–5.
16. Oszust M, Wysocki M. Recognition of Signed Expression Observed by Kinect Sensor. In: Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance. 2013. p. 220–5.
17. Sun C, Zhang T, Bao B-K, Xu C, Mei T. Discriminative Exemplar Coding for Sign Language Recognition with Kinect. In: IEEE Transactions on Cybernetics. 2013; 43(5):1418–28.
18. Haralick RM, Shapiro LG. Image Segmentation Techniques. Computer Vision, Graphics, and Image Processing, 1985; 29(1):100–32.
19. Zarit BD, Super BJ, Quck FKH. Comparison of five color models in skin pixel classification. In: Proceedings of International Conference on Computer Vision, 1999; 58–63.
20. Yao H, Gao W. Face detection and location based on skin chrominance and lip chrominance transformation from color images. Pattern Recognition. 2001; 34:1555–64.
21. Jones MJ, Rehg JM. Statistical Color Models with Application to Skin Detection. International Journal of Computer Vision. 2002; 46(1):81–96.
22. Gonzales RC, Woods RE. Digital Image Processing, 3rd ed. New York, NY: Prentice-Hall, 2008.
23. Zhang D, Lu G. A Comparative Study of Fourier Descriptors for Shape Representation and Retrieval. Presented at The 5th Asian Conference on Computer Vision. 2002. p. 646–51.
24. Kunttu I, Lepisto L, Rauhamaa J, Visa A. Multiscale Fourier descriptor for shape-based image retrieval. Presented at the IEEE International Conference on Pattern Recognition. 2004. p. 765–8.
25. Nixon MS, Aguando A. Feature Extraction and Image Processing. Philadelphia, PA: Elsevier. 2007.
26. Hu M-K. Visual pattern recognition by moment invariants. IRE Transactions on Information Theory. 1962; 8(2):179–87.
27. Wilson AD, Bobick AF. Parametric hidden markov models for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1999; 21(9):884–900.
28. Araga Y, Shirabayashi M, Kaido K, Hikawa H. Real Time Gesture Recognition using Posture Classifier and Jordan Recurrent Neural Network. Presented at the IEEE World Congress on Computational Intelligence. Brisbane, Australia, 2012.
29. Cortes C, Vapnik V. The Nature of Statistical Learning Theory. New York, NY: Springer, 1995.
30. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2(2):955–74.
31. Wang X, Paliwal KK. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognition. 2003; 36:2429–39.
32. Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. Journal of Machine Learning Research. 2001; 2:265–93.
33. Dietterich TG, Bakiri G. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research. 1995; 2:263–86.