Sign Language Recognition Using Scale Invariant Feature Transform and SVM

Ashwin S.Pol, Dr. S.L.Nalbalwar, Prof. N.S. Jadhav

Abstract— Sign language is one of the best methods used to communicate with handicapped people and robots. This paper presents system of sign language recognition using bag-of-words and multiclass support vector machine (SVM) approach that uses Scale invariant feature of image. The SIFT (Scale Invariant Feature transform) algorithm [1] takes an image and transforms it into a collection of local feature vectors. SIFT was firstly used to detect key points and describe them because the SIFT features were invariant to image scale and rotation and were robust to changes in the viewpoint and illumination. SIFT method is used to extract all the features of certain types of signs, then formed theirs code book from all the features by using K means clustering and finally classified by using the multiclass support vector machine. We are used American Sign Language (ASL) as an example of a gestural language. The objective of the paper is to decode ASL Image into the appropriate alphabets.

Index Terms—ASL (American Sign Language), SIFT (Scale Invariant Feature transform), SVM (Support Vector Machine), Bag-of-word (BOW).

1 Introduction

Vision based hand sign signal recognition is one of the emerging recent research is mainly on the human computer machine and robotics interaction applications. However, there are lots of issue are there to recognize the hand gesture, out of them are the variation of the hand gesture appearance, Scaled, rotated version of image and the image processing speed. The main aim of this paper is to develop sign language recognition system that is able to detect and translate the hand gesture (sign signal) from captured images to English alphabets.

There are different methods introduced to recognize sign signal out of them mostly used are glove based approach and vision based approach. Glove-based approach usually requires the signer to put on gloves, sensors that are used as measuring devices to model the hand postures whereas vision-based approach uses video camera to capture images and apply a particular image processing algorithm to recognize hand posture. We are interested in vision-based techniques, which are a more appropriate way of communication with people and machine. Also, gestures can be divided into static gestures (hand postures) and dynamic gestures. The hand motion conveys as much meaning as their posture does. A static sign is determined by a certain configuration of the hand, while a dynamic gesture is a moving gesture determined by a sequence of hand movements and configurations. The initial steps of this project are: (1) Extracting the feature of each ASL image, (2) Construction of codebook using K-mean clustering (3) Classification of ASL image features using multiclass SVM.

2 Literature Review

In last two decades, human hand gesture recognition provides a natural way to interact and communicate with machines has grabs much attention of many researchers around the globe. Various algorithms and techniques for recognizing hand gesture had been introduced by the researchers. Conventionally, for hand gesture recognition, the system should be consisting of four stages which are image acquisition, hand features extraction, processing extracted features and hand gesture recognition [9], [10]. The block diagram shown in Figure.1 depicts the hand gesture recognition steps that are commonly applied by the researchers.

L. Bretzner, I. Laptev, and T. Lindeberg (2002) [7] proposed scale-space color features are used to recognize hand gestures, which are based on feature detection and user independence. But, the system performance depends upon when there is no other skin color object present in the image. A. Argyros and M. Lourakis (2006) [8] suggest vision based hand gesture recognition based on a clear-cut and integrated hand contour then computed the curvature of each point on the contour. Grobel and Assan (1996) [11] used HMMs to recognize isolated signs with 91.3% accuracy out of 262 sign vocabulary. They extracted the features from video recording of signers wearing colored gloves. Kjeldsen and Kender (1996) [12] suggest an algorithm of skin color segmentation in the HSV color space and use a back propagation neural network to recognize gestures from the segmented hand images. Hongo et al. (2000)[13] use a skin color segmentation technique in order to segment the region of interest and then recognize the gestures by extracting directional features and using linear discriminant analysis. Manresa et al. (2000) [14] propose a method of three main steps: (i) hand segmentation based on skin color information, (ii) tracking of the position and the orientation of the hand by using a pixel based tracking for the temporal update of the hand state and (iii) estimation of the hand state in

- *Ashwin S. Pol is currently pursuing master's degree program in Electronics & Telecommunication Engineering in Dr. Babasaheb Ambedkar Technological University, India.*
  E-mail: ashwin.pol9@gmail.com
- *Dr.S.L.Nalbalwar is Associate Professor in Dr. Babasaheb Ambedkar Technological University, India.*
  E-mail:nalbalwar_sanjayan@yahoo.com
- *Prof. Narendra S. Jadhav is Assistant Professor in Dr. Babasaheb Ambedkar Technological University, India.*
  E-mail:nsjadhav@dbatu.ac.in

order to extract several hand features to define a deterministic process of gesture recognition. Imagawa, Matsuo, Taniguchi, Arita, and Igi.(2000) [15] present "A local feature extraction technique is employed to detect hand shapes in sign language recognition". They used appearance based Eigen method to detect hand shapes. Using a clustering technique, they generate clusters of hand shapes on an Eigen space. They have achieved accuracy of around 93% recognition of 160 words. Triesch and Von der Malsburg (2001)[16] propose a computer vision system that is based on Elastic Graph Matching, which is extended in order to allow combinations of different feature types at the graph nodes. Timi Ojala et. al. (2002)[17] They suggested the method for texture classification using Local Binary Patterns. Rotation Invariant method is widely used for texture classification and recognition. Murakami and Taguchi (1991)[10] investigated the use of recurrent neural nets for Japanese Sign Language recognition. Although it achieved a high accuracy of 96%, their system was limited only to 10 distinct signs.

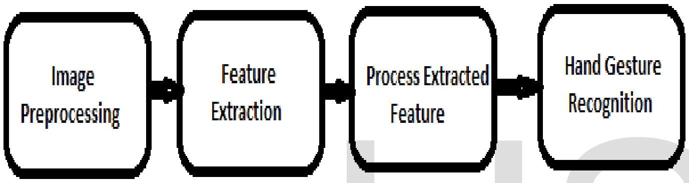


Fig.1 Basic Black Diagram of Recognition system

3 Algoritham Applied

3.1 Scale Invariant Feature Transform

The features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection [1]. These features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. SIFT [1] is divided into two stages, key point detection and key point description. Each stage consists of two sub-stages respectively.

1. Constructing a scale space: The first stage of computation searches over all scales and image locations. It is implemented efficiently by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation.

2. Key point localization: At each candidate location, a detailed model is used to determine location and scale. Key points are selected based on measures of their stability.

3. Assigning an orientation to the key points: One or more orientations are assigned to each key point location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.

4. Key point descriptor: The local image gradients are measured at the selected scale in the region around each key point. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

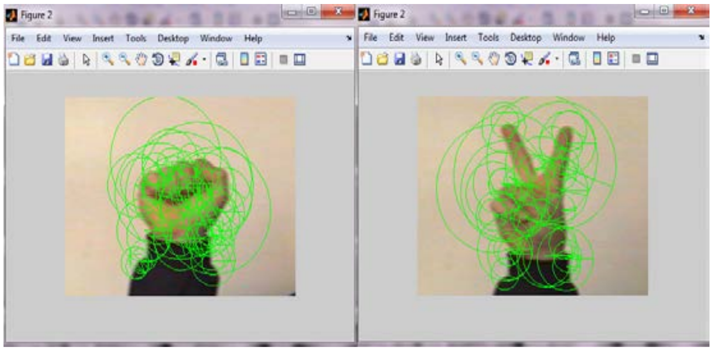


Figure 2
File Edit View Insert Tools Desktop Window Help
Figure 2
File Edit View Insert Tools Desktop Window Help

Fig. 2 Key point locazation of Sign A and V, Green circles of different sizes represent scale and we can detect interest point using Scale space of LOG.

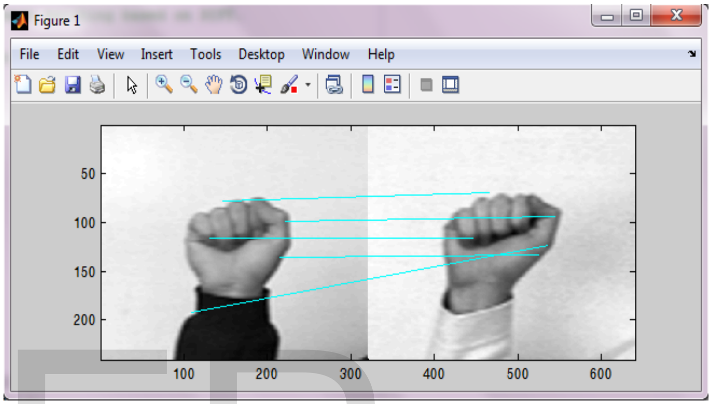


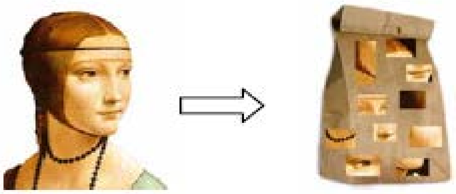
Fig.3 SIFT feature on sample image and corresponding matching point.

SIFT feature is a local feature. Compared with traditional overall feature, it improves the efficiency of the method. Fig.3 shows the SIFT features extracted on sample ASL images and some corresponding matching points in two sign images.

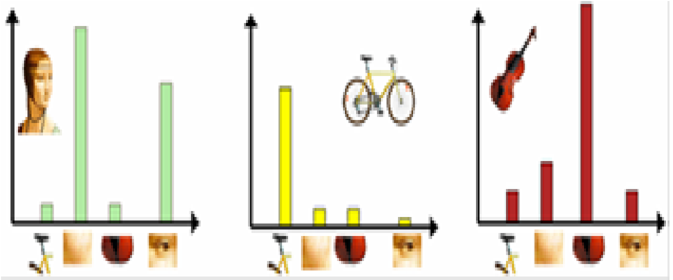
3.2 Bag-of-Word

In recent years," Bag-of-Words" model has been paid more and more attention in the field of object recognition. The image representation method based on "Bag-of-Words"[2] is shown in Fig.4.The basic idea of BoW is to categorize objects in terms of some particular features which called "code-words" and "codebook" consists of all the "code-words".

Bag-of-words is a valid classification from text analysis, two dimensional image data mapped to a collection of visual keywords, which is not only maintained the local features of the image but also compressed effectively the description of the image. Thus natural language processing techniques and methods can be very effectively applied to the field of object recognition.



(a) Local Feature Extracted



(b) After quantify the local features and construct codebook, an image is converted into a histogram of the number of occurrences of features descriptor in a given image

Fig. 4 Bag-of-words model [3].

3.3 Code Book Construction

Codebook construction process is shown in Fig.5 and as follows:

1. The SIFT features extracted from certain kind of training ASL images are clustered by using K-means method [6] and initialize k-means cluster centers with random values. The final cluster center is selected as candidate intrinsic feature.
2. After calculating the occurrence frequency of cluste center according to the Euclidean distance between each SIFT feature and cluster center. Top n of the occurrence frequency of cluster centers are selected as “code-words” and the size of “codebook” is n.

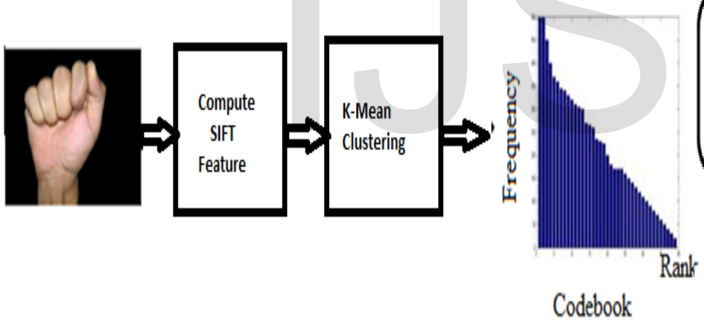
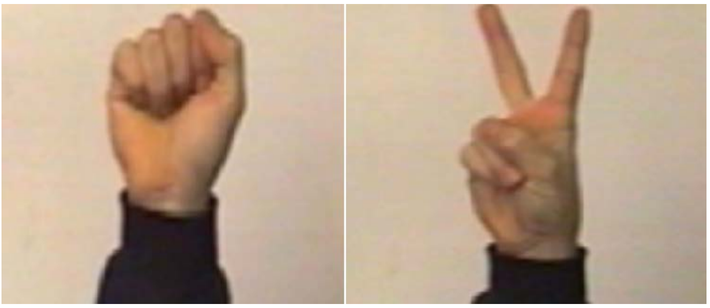
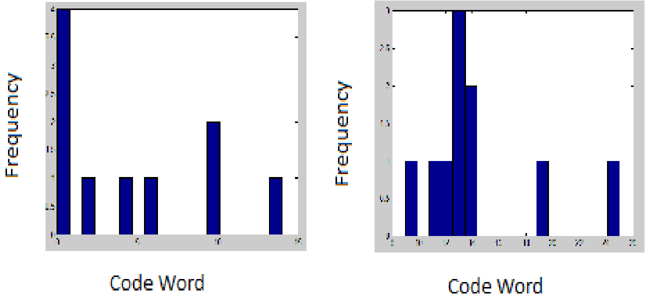


Fig. 5 Process of Codebook Constrction

Examples of converted code words by this method are shown in Fig.6. The difference of histograms among images is clearly found in the examples.



(a) Sign of English alphabet “A (b) Sign of English alphabet “V”



(c) Histogram of letter “A” (d) Histogram of letter “V”

Fig.6 Converted example of ASL sign

3.4 SVM Classifiers

We apply bag-of-words vector (histogram vector) with its related class number into a multiclass SVM classifier to create the multiclass SVM training classifier model. SVM is a group of related supervised learning methods used for classification. In our implementation, multiclass SVM training and testing are performed using the library for SVM [19]. This library supports multiclass classification and uses a one-against-one (OAO) method for multiclass classification in SVM. After extracting code-words from all kinds of ASL image, all training images are converted into histogram and four groups of SVMs are trained with these. Process of training is shown in Fig.7. And process of testing has shown in Fig.8.

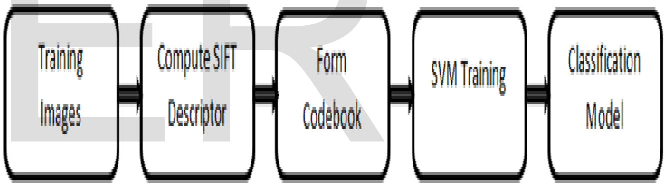


Fig.7 Process of training

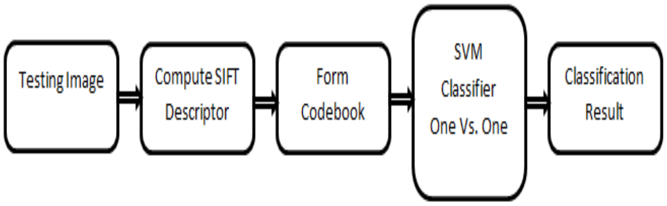


Fig.8 Process of testing

4 RESULT

The purpose of this study is to achieve higher recognition rate and lower false alarm rate for ASL hand gesture in most real scene. This database contains color images of four hand postures in ASL, which are A, B, V and Five, performed by different people against uniform backgrounds. In training stage we apply 60 image of each hand posture to build codebook then we train multiclass SVM model. In the testing stage, we apply new four hand postures images and create codebook of them classified using multiclass SVM classifier models that were built in the training stage to classify the four hand postures. The overall recognition accuracy is shown table.

TABLE 1
PERFORMANCE OF MULTICLASS SVM CLASIFFIER

Gesture Name	No. of images	Correct	Incorrect	Recognition Rate (%)
A	50	41	9	82
B	50	45	5	90
Five	50	42	8	84
V	50	43	7	87

5 Conclusion And Feature Work

In this paper we use SIFT algorithm for feature vector composition. The SIFT features described in our implementation which are invariant to scaling, rotation, addition of noise. Experiments considering the difficulties of ASL recognition such as lighting condition, angle of shooting and distance variation and were carried out on training image sets classified according to viewpoint and distance. As discussed in above section, SIFT method is used to extract all the features of certain types of signs, then formed theirs code book from all the features by using K means and finally classified by using the SVM. That means our method recognizes ASL by feature distribution. Result shows that the method is robust to different lighting condition, distance and angle of shooting. The system shows that the first stage can be useful for deaf persons or with speech disability for communicating with the rest of the people who do not know the language. As future work, it is planned to add to the system a learning process for dynamic signs.

References

[1] David G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, Vol.60, No. 2, pp. 91-110, 2004.

[2] Fei-Fei, L. and P. Perona, “A Bayesian hierarchical model for learning natural scene categories”, Computer Vision and Pattern Recognition, CVPR 2005.

[3] Fei-Fei,L.,2007.CVPR2007_tutorial_bag_of_words.From http://vision.cs.princeton.edu/documents/CVPR2007_tutorial_bag_of_words.ppt.

[4] Xiaoguang HU, Xinyan ZHU, “Traffic sign recognition using Scale invariant feature transform and SVM”, A special joint symposium of ISPRS Technical Commission IV & AutoCarto in conjunction with ASPRS/CaGIS Fall Specialty Conference November 15-19, 2010 Orlando, Florida

[5] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2006, pp. 2169–2178.

[6] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[7] L. Bretzner, I. Laptev, and T. Lindeberg, “Hand gesture recognition using multiscale color features, hieracrchichal models and particle filtering,” in Proc. Int. Conf. Autom. Face Gesture Recog., Washington, DC, May 2002.

[8] A. Argyros and M. Lourakis, “Vision-based interpretation of hand gestures for remote control of a computer mouse,” in Proc. Workshop Comput. Human Interact. 2006, pp. 40–51.

[9] W. K. Chung, W. Xinyu, and Y. Xu. 2008 “A realtime hand gesture recognition based on Haar wavelet representation”, in Proceedings of the IEEE International Conference on Robotics and Biomimetics, Washington, DC, USA, 2008, pp.336-341.

[10] K. Murakami and H. Taguchi, “Gesture recognition using recurrent neural networks,” in Proceedings of the Conference on Human Factors and Computing Systems, 1991,pp.237-242.

[11] K. Grobel, M. Assan. “Isolated sign language recognition using hidden markov models.” International Conference on Systems – ICONS 1996

[12] Kjeldsen, R., Kender, Finding skin in colour images. In: IEEE Second International Conference on Automated Face and Gesture Recognition, Killington, VT, USA, pp.184-188 1996.

[13] Hongo, H., Ohya, M., Yasumoto, M., Yamamoto, K, Face and hand gesture recognition for human–computer interaction. In: ICPR00: Fifteenth Interna- tional Conference on Pattern Recognition, Barcelona, Spain, 2000, pp.2921-2924.

[14] Manresa, C., Varona, J., Mas, R., Perales, Real-time hand tracking and gesture recognition for human–computer interaction. Electronic Letters on Computer Vision and Image Analysis, 2000, pp.1-7.

[15] Imagawa, I., Matsuo, H., Taniguchi, R., Arita, D., Lu, S., Igi, S. 2000 : Recognition of local features for camera-based sign language recognition system. In: Proc. 15th International Conference on Pattern Recognition. Volume 4. Pp.849-853.

[16] Triesch, J., Von der Malsburg, C., 2001. A system for person-independent hand posture recognition against complex backgrounds. IEEE Transanctions on Pattern Analysis and Machine Intelligence 23 (12), 1449-1453

[17] Timi Ojala, Matti Pietikainen and Topi Maenpaa, 2002 “Multi-resolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, pp.971-987.

[18] Nasser H. Dardas and Nicolas D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques" IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 60, NO. 11, NOVEMBER 2011.

[19] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multi-class support vector machines,” IEEE Trans. Neural Netw., vol. 13, no. 2, pp. 415– 425, Mar. 2002.