Sign Language Recognition System Using SIFT Based Approach

Ashwin S. Pol, S. L. Nalbalwar & N. S. Jadhav
Dept. of E&TC, Dr. BATU Lonere, MH, India
E-mail: ashwin.pol9@gmail.com, nalbalwar_sanjayan@yahoo.com, nsjadhav@dbatu.ac.in

Abstract – Sign language is one of the best methods used to communicate with handicapped people and robots. In this paper, we are used American Sign Language (ASL) as an example of a gestural language that uses a new class of local image features. 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. The objective of the paper is to decode a ASL Image into the appropriate alphabets.

Keywords – ASL, Feature Descriptor, SIFT.

I. INTRODUCTION

Hand sign signal language interpretation is important as it is one of the methods for handicapped people because verbal communication is impossible to exchange their ideas and to communicate with other people. Sign language recognition is important aspect for Human-Computer Interaction (HCI) and robotics design has drawn a lot of attentions from research fellow around the world in recent year. However, today there are lots of different issues faced by the research fellow in the effort to recognize the sign signal. Some of them are the variation of the hand gesture appearance, scaled, rotated version of image and the image processing speed as it involves many mathematical calculations. Traditionally, there are different approaches introduced to recognize hand gesture but most commonly used are glove-based and vision-based approach. Glove-based approach [5] usually signer requires to wear gloves in hand, also some sensors and so on that are used as sign signal to model the hand posture or gesture, whereas vision-based approach [6] signer does not require to wear anything natural hand used as sign signal to model hand posture or gesture. We are interested in vision-based techniques to recognize hand posture, which are a more natural way of communication with handicapped people and robots. Also, gestures can be divided into static gestures (hand postures) and dynamic gestures. We are currently working on static hand posture (image without motion) as example of sign signal.

II. 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 [4]. The block diagram shown in Figure.2 depicts the hand gesture recognition steps that are commonly applied by the researchers. Hand gesture recognition is a complex problem that has been dealt with in many different ways. Grobel and Assan (1996)[12] 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) [10] 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)[3] 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)[11] 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 order to extract several hand features to define a deterministic process of gesture recognition. Imagawa, Matsuo, Taniguchi, Arita, and Igi.(2000) [13] 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. Vogler and Metaxas (1997) used computer vision methods and HMMs to recognize continuous American Sign Language sentences with a vocabulary of 53 signs. They modeled context-dependent HMMs to alleviate the effects of movement epenthesis. An accuracy of 89.9% was observed. Triesch and Von der Malsburg (2001) [9] 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) [2] they suggested the method for texture classification using Local Binary Patterns. Rotation Invariant method is widely used for texture classification and recognition. Although it achieved a high accuracy of 96%, their system was limited only to 10 distinct signs. H. Desa, and W. Majid (2009) [5] developed an ASL finger spelling system using a Cyber glove, with the use of neural networks for data segmentation, feature classifier, and sign recognition. Using a tree-structured neural classifying vector quantize, a large neural network with 51 nodes was developed for the recognition of ASL alphabets. They claimed a recognition accuracy of 98.9% for the system.

III. SYSTEM FUNCTIONALITY

The system is designed to visually recognize static signs of the American Sign Language (ASL), all signs of ASL alphabets using bare hands. The user or signers are not required to wear any gloves or to use any devices to interact with the system. But, since different signers vary their hand shape size, and operation habit and so on, which bring more difficulties in recognition. The entire method consists of the following three main stages: Stage A: Image Preprocessing Stage B: Feature extraction using SIFT. Stage C: Gesture Recognition.

A. Image Preprocessing

In this Stage, the input gestures database is creating which contains the different gestures poses with many samples per gesture. Images of signs were resized to 640 by 480, by default MATLAB function “imresize” uses nearest neighbor interpolation to determine the values of pixels in the output image but other interpolation methods can be specified. Here ‘bicubic’ method is used because if the specified output size is smaller than the size of the input image, “imresize” applies a low pass filter before interpolation to reduce aliasing. Therefore we get default filter size 11 by 11.

B. Scale Invariant Feature Transform

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

1) Detection of scale space extrema

The first stage of key point detection is to identify locations and scales that can be repeatable assigned under differing views of the same object.

Fig.1. Basic Block Diagram of Recognition System

Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space[7] (Witkin,1983). Under a variety of Reasonable assumptions the only possible scale-space kernel is the Gaussian function. Therefore, the scale space of an image is defined as a function, L(X,Y,σ) that is produced from the convolution of a variable-scale Gaussian, G(X,Y, σ) with an input image, I(X,Y):

\[ L(X,Y,\sigma) = G(X,Y, \sigma) * I(X,Y) \]  \hspace{1cm} (1)

To efficiently detect stable key point locations in scale space, [8] proposed Using scale-space extrema in the difference-of-Gaussian function convolved with the image, D(X,Y, σ) which can be computed from the difference of two nearby scales separated by a Constant multiplicative factor k:

\[ D(x, \ y, \ \sigma) = (G(x, y, k \sigma) - G(x, y, \sigma)) * I(x, y) \]

\[ = L(x, \ y, k \sigma) - L(x, \ y, \sigma). \hspace{1cm} (2) \]

2) Locate maxima or minima in DOG image

In this step we coarsely locate the maxima or minima. We go through each pixel and check all its
neighbors. The check is done within the current image, and also the one above and below it. In Fig.4 X marks the current pixel and green circles mark the neighbors. This way, a total of 26 checks are made. X is marked as a “key point” if it is the greatest or least of all 26 neighbors. To increase chances of matching and stability of the algorithm, we also find sub pixel maxima or minima using Taylor series approximation.

4) Key point Descriptor

We want to generate a very unique fingerprint for the key point. We also want it to be relatively lenient when it is being compared against other key points. To do this, as shown in Fig.5 we take 16x16 window around the key point. This 16x16 window is broken into sixteen 4x4 windows. Within each 4x4 window, gradient magnitudes and orientations are calculated. These orientations are put into an 8 bin histogram. The amount added to the histogram bin depends on the magnitude of the gradient. This is done using a “Gaussian weighting function”. Doing this for all 16 pixels, you would’ve “compiled” 16 totally random orientations into 8 predetermined bins. You do this for all sixteen 4x4 regions. So you end up with 4x4x8 = 128 numbers. These 128 numbers form the “feature vector”. The key point is uniquely identified by this feature vector.

IV. GESTURE RECOGNITION

For Sign image matching, we saved the feature vectors for the training image set. When a New image is applied to the algorithm, preprocessing steps discussed in Section III are first performed. Then, we use SIFT algorithm to calculate the feature vectors for this input image. The minimum Euclidean distance between each feature vector of the query image and all the feature vectors of the database is found. The Gesture image having a feature vector with the minimum Euclidean distance to a feature vector of the query image is given a vote to be the right sign of alphabet. After we go over all the feature vectors of the query image giving votes to alphabet sign in the database, we observed that we always have the right sign of alphabet to be the one with the highest number of votes.

We decided to compare the highest vote (corresponding to the right ASL alphabet) and the second highest vote (corresponding to the most conflicting alphabet). If the difference between them is larger than a threshold, then there is a match and this match corresponds to the highest vote. If the difference is smaller than a threshold, then we declare a ‘No Match’.
Fig. 6 Sign A and V. Green circles of different sizes represent scale and we can detect interest point using Scale space of LOG.

V. EXPERIMENTAL RESULT

The hand gesture recognition system, presented in this paper is tested by using hand images from different people with varying shape, scale, rotation and size. The conclusions drawn based on the robustness of the features and Sign recognition. In our Experiment we take total seven gestures A, V, W, 1 etc. of hand of different people. In Fig.6 as we see circle of different sizes and each size correspond to scale we use σ value. This tells us we can detect interest point using scale space of LOG of image.

As shown Fig. 7 input image representing character V. When we calculate the feature descriptor of input image and it matches key points with the scaled image of character V present in the database. As shown in Fig. 8 input image representing character A. When we calculate the feature descriptor of input image and it matches key points with the scaled image of character A present in the database.

Fig.7. Matched Database image Vs. Input image of ASL character V

Fig.8. Matched Database image Vs. Input image of ASL character A

We have also seen that SIFT is successfully able to detect similarities between images, even though the image has went through transformation. We see for different images which has scale change, rotation version of image, also possible to recognize successfully.

VI. CONCLUSION AND FUTURE WORK

We use SIFT algorithm for feature vector composition. The SIFT features described in our implementation have been computed at the edges which are invariant to scaling, rotation, addition of noise. These features are useful due to their distinctiveness, which enables the correct match for key points between different hand gestures. This makes the recognition system more practical since signers do not have to wear gloves that make the signing process natural.

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. Also if we use Bag of Word approach for classification purpose then matching rate of signs possible to increase.

VII. REFERENCES


