Hand Gesture Recognition using SIFT

S. Pandita[1], S. P. Narote[2]
Department of E&TC, Sinhgad College of Engineering,
University of Pune, Pune, India.

Abstract

Gesture refers to expressive movement of human body parts having a particular message to be communicated to a receiver. Gesture recognition pertains to understanding meaning of human body part movement,which involves the movement of hand, face, head, arms or body .Today's world is witnessing enormous improvements in processing speeds and visualization displays, but still input devices for the most part have lagged behind, presenting a bottleneck in applications. Gesture recognition is very important in designing an efficient HCI. This paper presents a method for recognizing hand gestures by extracting distinctive invariant features fromimages that can be used to perform efficient matching between different views ofa hand gesture. The features are invariant to image scale and rotation, and provide robust matching across a considerable range of affine distortion,change in 3D viewpoint, addition of noise, and change in illumination [1].

Keywords: SIFT algorithm, Human Computer Interaction (HCI), Hand Gestures, Hand GestureRecognition.

1. Introduction

Gestures have got deep roots in our communication. The remarkable ability of the human is the gesture recognition, it is commonly noticedwhile hearing impaired people communicate with each other as well as with hearing people via sign language. Hand Gesture Recognition is becoming an increasingly important for many applications including human machine interfaces, multimedia, security, communication, visually mediated interaction and provides a separate complementary modality to speech for expression of ideas. In this paper we consider the American Sign Language (ASL) gestures as the images to be worked upon by decoding them with English alphabet. Our implementation focuses on deriving SIFT features from an image and trying using these features to perform gesture recognition [2]. The approach of SIFT feature detection taken in our implementation is similar with the one taken by Lowe et.al. [3], which is used for object recognition. According to[3] the invariant features extracted from images can be used to perform reliable matching between different views of an object or scene. The features have been shown to be invariant to image rotation and scale and robust across a substantial range of affine distortion, addition of noise, and change in illumination. The approach is efficient on feature extraction and has the ability to identify large numbers of features[2].The organization of the paper is as follows. Section 2 presents the SIFT algorithm. The gesture recognition details are presented in Section3. Section 4presents the results, section 5 discusses the conclusion.

2. SIFT Algorithm

The scale invariant feature transform (SIFT) algorithm, developed by Lowe [1,3,4], is an algorithm for image features generation which are invariant to image translation, scaling, rotation and partially invariant to illumination changes and affine projection. SIFT algorithm can be used to detect distinct features in an image. Once features have been detected for two different images, one can use these features to answer questions like "are the two images taken of the same object?" and "given an object in the first image, is it present in the second image?"[5].Computation of SIFTimage features is performed through the four consecutive phases which are briefly described inthe following:

2.1.Scale-Space Local Extrema Detection

This stage of the filtering attempts to identify those locations and scales that are identifiable from different views of the same object. This can be efficiently achieved using a "scale space" function. Further it has been shown under reasonable assumptions it must be based on the Gaussian function. The scale space is defined by the function:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad (1)$$

Where * is the convolution operator, $G(x, y, \sigma)$ is a variable-scale Gaussian and $I(x, y)$ is the input image.Various techniques can then be used to detect stable keypoint locations in the scale-space. Difference of Gaussians is one such technique, locating scale-space extrema, $D(x, y, \sigma)$ by computing the difference between two images, one with scale k times the other. $D(x, y, \sigma)$ is then given by:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2)$$

To detect the local maxima and minima of $D(x, y, \sigma)$ each point is compared with its 8 neighbours at the same scale, and its 9 neighbours up and down one scale as presented in figure 1.

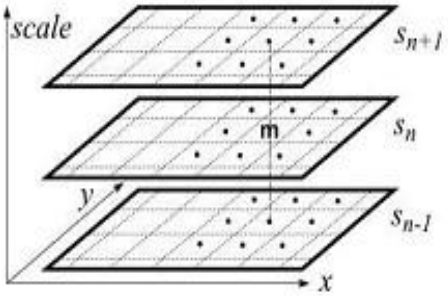


Figure1. An extrema is defined as any value in the DoG greater than all its neighbours in scale-space.[2]

If this value is the minimum or maximum of all these points then this point is an extrema.The search for extrema excludes the first and thelast image in each octave because they do not have a scale above and a scale belowrespectively. To increase the number of extracted features the input image is doubledbefore it is treated by SIFT algorithm, which however increases the computational time significantly.

2.2.Keypoint Localization

The detected local extrema are good candidates for keypoints.However, they need to be exactly localized by fitting a 3D quadratic function to thescale-space local sample point. The quadratic function is computed using a secondorder Taylor expansion having the origin at the sample point. Then, local extrema withlow contrast and such that correspond to edges are discarded because they aresensitive to noise.

2.3.Orientation Assignment

Once the SIFT-feature location is determined, a mainorientation is assigned to each feature based on local image gradients as shown in figure 2. For each pixelof the region around the feature location the gradient magnitude and orientation arecomputed respectively as:

$$m(x, y) = \surd(L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2 \qquad (3)$$

$$\theta(x, y) = \tan^{-1}((L(x, y + 1) - L(x, y - 1))/(L(x + 1, y) - L(x - 1, y))) \qquad (4)$$

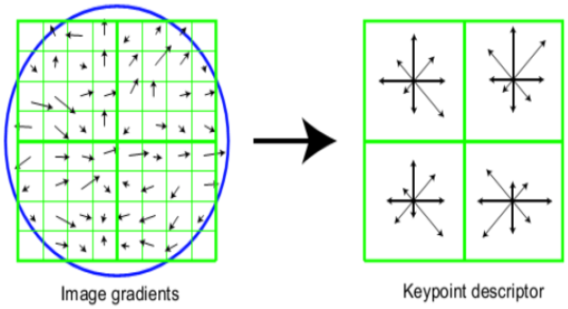


Figure 2. Orientation Assignment[2]

The gradient magnitudes are weighted by a Gaussian window whose size depends onthe feature octave. To detect the local maxima and minima of D(x, y, σ) each point is compared with its 8 neighbours at the same scale, and its 9 neighbours up and down one scale. If this value is the minimum or maximum of all these points then this point is an extrema.

2.4.Keypoint Descriptor

The local image gradients are measured at the selected scale in the region around each keypoint[6]. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination[1].

3. Gesture Recognition

In order to get a reliable recognition, it is quite important that the features extracted from the training image are detectable even under changes in image scale, noise and illumination[7]. Such points generally lie on high-contrast regions of the image, for example object edges. Gesture recognition is initially performed by matching each keypoint independently to the database of keypoints extracted from training images[1]. Many of these initial matches will be incorrect due to ambiguous features or features that arise from background clutter. Therefore, clusters of some features are first identified that agree on an object and its pose, as these clusters have a much higher probability of being correct than individual feature matches. Then, each cluster is checked by performing a detailed geometric fit to the model, and the result is used to accept or reject the interpretation.

4.Results

Results from our implementation are shown in figures 3 &4. Noise adjustment is a very essential part for our approach which could result in inefficientor false matching. However, we have used parameterswhich should help the keep the featurematching robust to noise in this implementation.

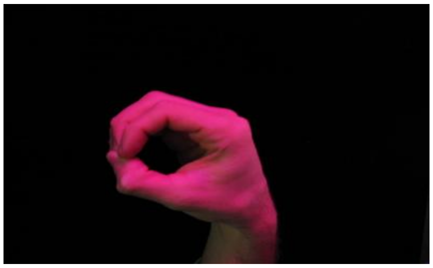
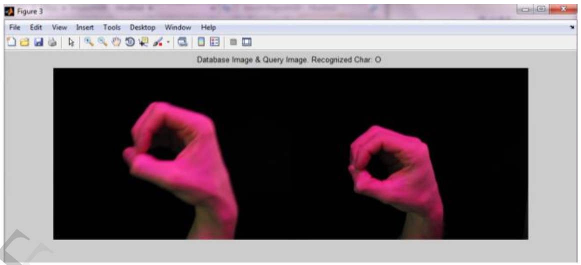


Figure 3(a) .Input query image

This input image representing character O is a colour image. When this is applied as an input query image , it matches with the scaled image of character O present in the database .



Database Image & Query Image. Recognized Char: O

Figure3 (b).Matched Database image Vs Input query image for character O

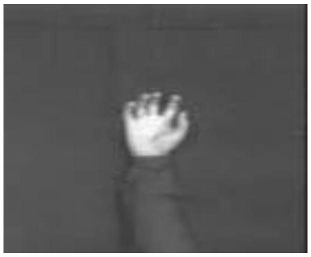
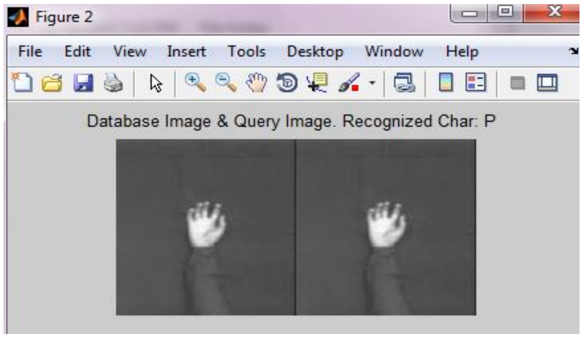


Figure 4(a) .Input query image



Database Image & Query Image. Recognized Char: P

Figure4 (b).Matched Database imageVs. Input query image for character P.

This input image shown in figure 4(a) is a grey scale image representing character P. We have also seen that SIFT issuccessfully able to detect similarities between images, eventhough the image has went through transformation. Some of the transformations thatwetested in our implementation were:

Saturation- Even though colour removal was done from the input query image, the database image,and both, the results were correct.

Scale- We scaled down the source image as well as the destination image and still the recognition was found to be correct.

Rotation- We took pictures at a skewed angle and then did the matching and results came out to be correct.

The summary of the results is presented in table1 below

Table1. Result Table

Sr.No.	Input Gesture	Recognized Gesture
1	B	B
2	B	B
3	B	B
4	L	L
5	H	H
6	I	I
7	C	C
8	O	O
9	A	No matches found
10	Y	S
11	AE	P

5. Conclusion

With the help of our algorithm we were able to decode gestures successfully. The feature extraction was done efficiently using SIFT. 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 keypoints between different hand gestures. The proposed approach was tested on real images . Also computation time was found to be lesser for grey scale images than with color images.

References

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

[2] Yu Meng and Dr. Bernard Tiddeman, “Implementing the Scale Invariant Feature Transform (SIFT) Method,” University of St. Andrews.

[3] David G. Lowe, “Object recognition from local scale-invariant features”, *International Conference on Computer Vision, Corfu, Greece*, pp. 1150-1157, September 1999

[4] David G. Lowe, “Local feature view clustering for 3D object recognition”, *IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, pp. 682-688, December2001

[5] Thomas Bakken, “R&I Research Note” Telenor ASA, 2007.

[6] Faraj Alhwarin, Chao Wang et.al., “Improved SIFT-Features Matching for Object Recognition”, *BCS International Academic Conference Visions of Computer Science,2008.*

[7] Pallavi Gurjal et. al., “ Real Time Hand Gesture Recognition Using SIFT”,*International Journal of Electronics and Electrical Engineering,* Volume 2, pp.19-33, Issue 3 ,March 2012.