J. Inst. Eng. India Ser. B (February 2017) 98(1):19–26
DOI 10.1007/s40031-016-0250-8

ORIGINAL CONTRIBUTION

Distinctive Feature Extraction for Indian Sign Language (ISL) Gesture using Scale Invariant Feature Transform (SIFT)

Sandeep Baburao Patil[1] · G. R. Sinha[1]

Received: 12 August 2014 / Accepted: 4 July 2016 / Published online: 21 July 2016

Abstract India, having less awareness towards the deaf and dumb peoples leads to increase the communication gap between deaf and hard hearing community. Sign language is commonly developed for deaf and hard hearing peoples to convey their message by generating the different sign pattern. The scale invariant feature transform was introduced by David Lowe to perform reliable matching between different images of the same object. This paper implements the various phases of scale invariant feature transform to extract the distinctive features from Indian sign language gestures. The experimental result shows the time constraint for each phase and the number of features extracted for 26 ISL gestures.

Keywords Indian sign language · Gesture recognition · Scale invariant feature transform · Distinctive features · Key points · Difference of Gaussian (DOG)

Introduction

In recent years, there have been numerous research contributions in the field of Indian sign language (ISL) biometrics which helps in identification of persons based on their traits or characteristics. This area is not very easy because no standard databases are available for the ISL biometrics. The significance of the problem can be easily illustrated by using natural gestures applied together with verbal and nonverbal communication. The use of hand gestures in support of verbal communication is very useful for the people having no visual contact [1–6]. Different approaches for recognizing hand gestures are available in literatures; Few of them require wearing marked gloves or attaching extra hardware to the body of the subject [7–9]. These approaches are less likely to apply for real world applications whereas vision-based approaches are considered as non-intrusive and hence more likely to be used for real world applications.

ISL is an attempt in this direction that deals with recognition of various static and symbolic characters generated by hand gestures. The biometrics uses the ISL gestures as human traits and recognizes the characters accordingly, which is very useful for deaf and dumb people [10–15]. Many professional estimate that the deaf population in India is approximately 3 million and hard hearing people is 10 millions. Around 100 million people are associated and involved with these 13 million people caring them and helping them in communication. The involved peoples are family members, social workers, audiologist, professional, teachers etc. If any solution is developed for helping the deaf and dumb people to recognize the language then it would be a significant contribution towards society and mankind [11].

Literature Review

A numerous research works on ISL biometrics is being carried out and some of the researchers have already contributed in this area. Study of sign language biometrics and also ISL biometrics was made and below is a report of literature survey.

Bhuyan, et al. [4] proposed a novel approach for hand pose recognition for analyzing the textures and key

✉ Sandeep Baburao Patil
patilsandeepb1212@gmail.com

[1] Shri Shankaracharya Technical Campus, Shri Shankaracharya Group of Institutions, Bhilai, India

geometrical features of the hand. A skeletal hand model was constructed to analyze the abduction/adduction movements of the fingers and subsequently, texture analysis was performed to consider some inflexive finger movements. The feature extraction technique is more complex because the abduction and adduction angle of the fingers and internal variations for particular gestures [4]. Ghotkar, et al. [16] introduced a hand gesture recognition system to recognize the alphabets of ISL. The system consists of four modules: real time hand tracking, hand segmentation, feature extraction and gesture recognition. Genetic algorithm was used for gesture recognition. Computational time required for the system is more, since it has to process through four different modules [16]. Chakraborty, et al. [6]. suggested a novel approach towards recognizing of ISL gesture for humanoid robot interaction (HRI). The JAVA based software was developed to deal with the entire HRI process. This system found its limitations in real time applications where different speech and hearing would explore several issues [6]. Balakrishnan and Rajam [4] developed a Sign Language biometrics system for one of the south Indian languages. The system describes a set of 32 signs, each representing the binary 'UP' and 'DOWN' positions of the five fingers of right hand palm. This system was applied to single user both in training and testing phase. The earlier mentioned system is applicable to Tamil text only and find limited to five fingers of single hand with same background [5]. Quan [17] introduced a novel recognition method of sign language using the vision-based multi-features classifier. This can exclude some skin-like object and tracked the moving recognizes hand more precisely from the sign language video sequence. The system developed dynamic sign language appearance modeling first, and then classification method of SVMs for recognition has been used. The experiment was carried out over 30 groups of the Chinese manual alphabet images and the results proved that this appearance modeling method is simple, efficient, and effective for characterizing hand gestures. The overall result shows that using linear kernel function the best recognition rate of 99.7 % of letter 'F' has been achieved. The system suffers with limitations in real time with dynamic hand gesture [17]. Arulkarthick, et al. [2] presented a real-time system for sign language recognition using hand gestures. After detecting hands postures, hand gestures are recognized based on features of hand that are extracted using the Haar transform. K-means clustering algorithm was used to reduce the number of extracted features which reduced the computational complexity. This method was limited to small database, and for large database the performance of AdaBoost algorithm is not satisfactory [2]. Alon and Athitsos [3] introduced a unified framework that simultaneously performed spatial segmentation, temporal segmentation, and recognition. The framework carried both the information's that is bottom-up and top-down. A gesture has been recognized even when the hand location is highly ambiguous and gesture's begins and ends information are unavailable. The performance of this approach was evaluated on two challenging applications: recognition of hand-signed digits gestured by users wearing short-sleeved shirts, and retrieval of occurrences of signs of interest from a video database containing continuous, unsegmented signing of American Sign Language (ASL). This system is limited to five signs of ASL. The algorithm has to undergo through three major components that is feature extraction, spatio-temporal matching and intermodal competition that requires large processing time [3]. Many authors developed a novel algorithm for the hand gesture images detection and recognition. The detection process rotates an askew gesture to right position, and to delete the elbow and forearm parts from the captured pictures. The recognition process includes two phases (a) the model construction and (b) sign language identification. This algorithm gives 94 % accuracy for hand gesture recognition and the accuracy will be adversely affected if the orientation of hand gesture lies in skew or if the image part from the wrist to the arm is incorrect.

Jothilakshmi, et al. [7] used acoustic features to develop a two level language identification system for Indian languages. Firstly, the system identifies the family of the spoken language, and then it is fed to the second level to identify the particular language in the corresponding family. The system uses hidden Markov model (HMM), Gaussian mixture model (GMM), artificial neural networks (ANN). The system could not achieve good accuracy [7]. Kadam, et al. [18] studied an effective use of glove for implementing an interactive sign language teaching programmed. Sign language glove is very useful to aid in communication with the deaf. A translator was suggested to those who want to learn sign language. The flex sensors are required to mount on the finger of the glove whose resistance changes according to the finger position that is difficult to understand for the deaf.

Nguyen, et al. [11] described the facial expressions whose features are tracked to effectively capture temporal visual cues on the signers face during signing using probabilistic principal component analysis (PPCA). A test was conducted to recognize six isolated facial expressions representing grammatical markers in American sign language (ASL). The recognition accuracy reported for ASL facial expressions was 91.76 % in person dependent tests and 87.71 % in person independent tests. The proposed system is not fully automatic and some subject to tilt/rotate their head while thrusting the head forward, contributing to the lower accuracy for these two classes [14]. Paulraj, et al. [15] proposed a sign language recognition system that will help the hearing impaired to communicate more fluently with the normal people. Sign language recognition system employs skin color segmentation and neural network method for training. A segmentation process is carried out to separate the right and left hand regions from the image frame and a simple vertical interleaving method has been used in the preprocessing stage to reduce the size of the image. The system requires about 3218 average mean epochs to train the network model that exceeds the training time. Moreover there was confusion in first, fourth, eighth and ninth sign that greatly reduced the accuracy [15]. The first automatic Arabic sign language (ArSL) recognition system was introduced based on HMMs. The system operates in different modes including offline, online, signer-dependent, and signer-independent modes. Experimental results demonstrated that the given system has high recognition rate for all modes. The system did not rely on the use of data gloves or other means as input devices, and it allows the deaf signers to perform gestures freely and naturally. This system is limited to only ArSL [16]. Magdy and Samir [11] developed a post processing module based on natural language processing rules that is fast, easy and computationally simple implementation. The system was more accurate especially for domain-specific sign language recognition.

Problem Formulation

There have been research contributions in sign language biometrics but very limited work has been done in ISL biometrics. ISL gestures used in literatures were not common and no specific database available for ISL.

Methods developed for other languages do not perform well for ISL because the methods are biased for the language which is written differently than ISL.

Gloves have been used in many methods that lead to poor identification of fingers and their separation.

Database Collection

The image of ISL is shown in Fig. 1. Based on this image the hand gesture databases from 10 different persons were collected. The database consists of 26 images for each class. Figure 2 shows the database of one of the class. Each image was taken having a resolution of 284 × 215 pixels. Resolution of 284 × 215 produces a 0.06 mega pixel image.

Proposed Methodology

Scale invariant feature transform (SIFT) algorithm was first introduced by David Lowe in 1999, to perform reliable matching between different images of the same object. This algorithm is widely used for feature extraction because of the image stability over translation, rotation and scaling and to some extent invariant to change in the illumination and camera viewpoint. Localization is very well achieved in both spatial and frequency domains that helps in reducing the probability of disruption by occlusion, clutter, or noise. The proposed methodology is used to extract the feature of ISL gesture. The flow chart of the proposed methodology is shown in Fig. 3.

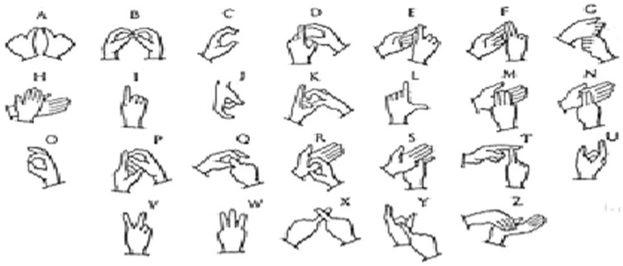


Fig. 1 Indian sign language

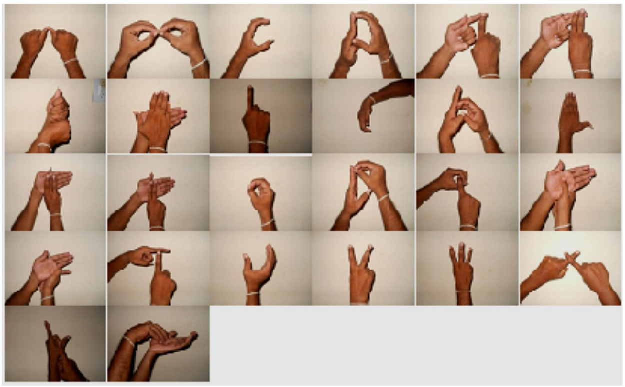


Fig. 2 26 ISL gesture from class 1

Scale Space Extrema Detection

This phase identify the stable points from different views of the same image over image rotation, translation and scaling. Figure 4 shows the internal stage of the extreme detection phase. The algorithm computes 'scale', 'difference of Gaussian' and 'extrema' over several 'octaves'.

The difference of Gaussian image $D\ (x, y, \sigma)$ is given by

$$D(x,y,\sigma) = L(x,y,k_i\sigma) - L(x,y,k_j\sigma) \tag{1}$$

where $L(x,y,k_j\sigma)$ is the convolution of the original image $I(x,\ y)$ with the Gaussian blur $G(x,\ y,\ K_\sigma)$ at scale K_σ, that is,

$$L(x,y,k_\sigma) = G(x,y,K_\sigma) * I(x,y) \tag{2}$$

where * is the convolution operator, $G(x,\ y,\ \sigma)$ is a variable-scale Gaussian and $I(x,\ y)$ is the input image.

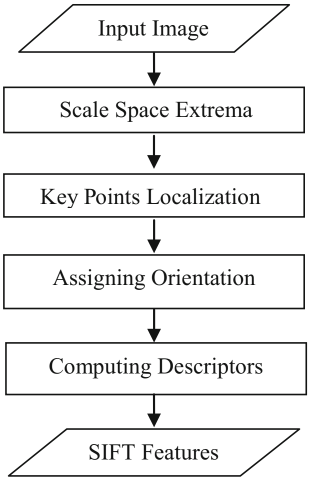


Fig. 3 Various phases of SIFT algorithm

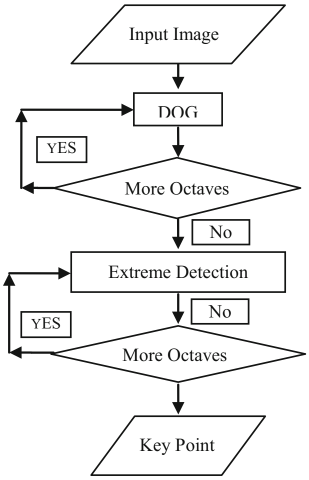


Fig. 4 Flow chart for generating DOG for several octaves

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{3}$$

Once difference of Gaussian (DOG) images has been obtained, key points are identified as local minima/maxima of the DOG images across scales. This is done by comparing each pixel in the DOG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. If the pixel value is the maximum or minimum among all compared pixels, it is selected as a candidate key point.

To determine the difference of Gaussian, only two corresponding points are required and to check whether the point is extremum 26 differences of Gaussian points are required to be spread over three scales around it. Figure 5 shows the scale and DOG for first octave, similarly DOG for second and third octave is calculated.

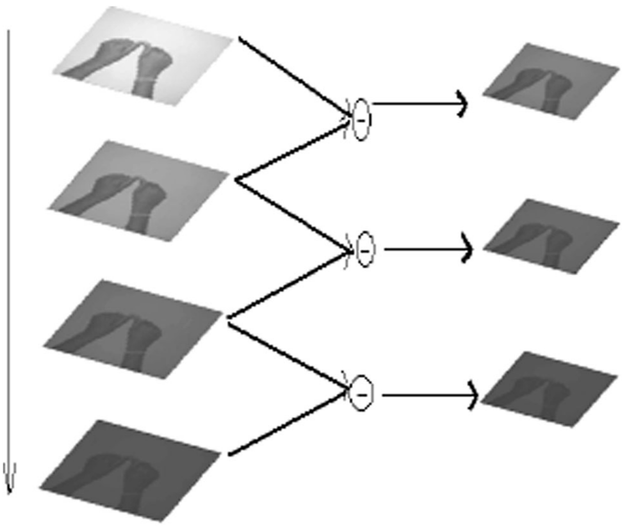


Fig. 5 Scale and DOG for first octave

Key Point Detection

As the input to a SIFT algorithm is a set of N^2 pixels of an N × N image. The small fraction of these pixels typically turns out to be extrema, let $0 < \alpha < 1$ be this fraction. Accordingly αN^2 extrema will move on to the next key point detection phase. Again only the small fraction of these extrema will qualify as a key point, let $0 < \beta < 1$ be this fraction. So nominally there are $\alpha\beta N^2$ key point of that image. In this phase the candidate which lies on the edge of the image may corresponds to point of low contrast. These are generally not useful as feature as they are unstable over image variation. To reject these low contrast or poorly localized extrema, the Taylor expansion approach has been used. In this approach the scale-space function $D(x, y, \sigma)$, shifted so that the origin is at the sample point;

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x \tag{4}$$

where D and its derivatives are evaluated at the sample point and $x = (x, y, \sigma)^T$ is the offset from this point. The location of the extrema, $\hat{x}$, is determine by taking the derivative of this function with respect to x and setting it to zero, giving

$$\hat{x} = \frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x} \tag{5}$$

As suggested by Brown, the Hessian and derivative of D are approximated by using differences of neighboring sample points. The resulting 3 × 3 linear system can be solved with minimal cost. If the offset $\hat{x}$ is larger than 0.5 in any dimension, then it means that the extrema lies closer to a different sample point. The final offset $\hat{x}$ is added to the location of its sample point to get the interpolated estimate for the location of the extrema.

The function value of the extrema, $D(\hat{x})$, is useful for rejecting unstable extrema with low contrast. This can be obtained by substituting Eq. (5) into (4) giving,

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\hat{x} \tag{6}$$

In this paper, all extrema with a value of $|D(\hat{x})|$ less than 0.03 were rejected. For stability it is not sufficient to reject key point with low contrast. The DOG function will have strong response along edge, and therefore unstable to small amount of noise.

A poorly defined peak in the DOG function will have a large principal curvature across an edge. The principal curvatures can be computed from a 2 × 2 Hessian matrix, H, computed at the location and scale of the key point.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{7}$$

The derivatives are estimated by taking difference of neighboring sample point. Let α be the eigenvalue with the largest magnitude and β be the smaller one. Then the sum of the eigenvalue can be computed form the trace of H and their product can be obtained from the determinant.

$$T_r(H) = D_{xx} + D_{yy} = \alpha + \beta$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta$$

In the unlikely event, the determinant is negative, the curvature have different signs so the point is discarded as not being an extremum. Let r be the ration between the largest magnitude eigenvalue and the smaller one, so that $\alpha = r\beta$. Then,

$$\frac{T_r(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{8}$$

This depends only on the ratio of the eigenvalues rather than their individuals values. The quantity $\frac{(r+1)^2}{r}$ is at a minimum when the two eigenvalues are equal and it increases with r. Therefore to check that the ratio of principal curvatures is below some threshold, r, the following relation needs to be checked

$$\frac{T_r(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{9}$$

This is very efficient to compute which can eliminates key points that have a ratio between the principal curvatures greater than r. The final key points detected for first two images were shown in Fig. 6.

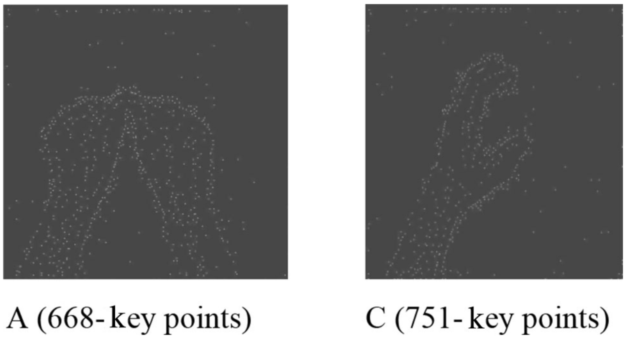
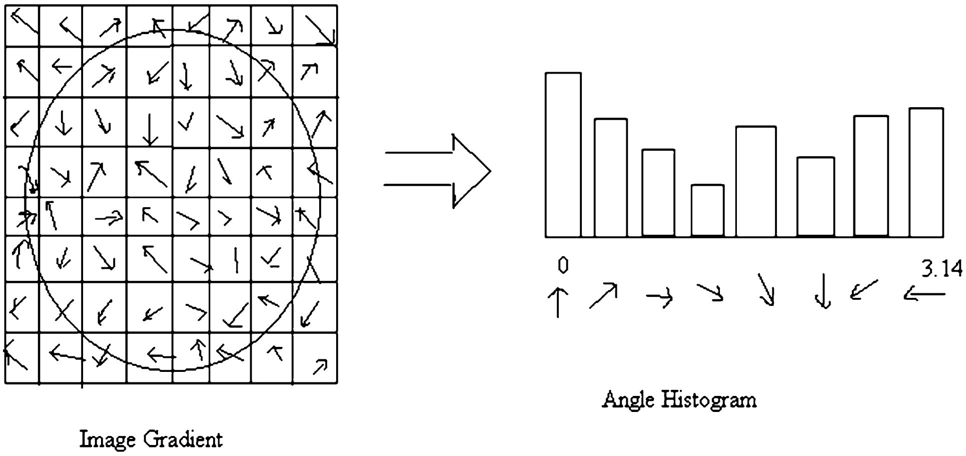


Fig. 6 Number of key points for first two images

Orientation Assignment

This phase assigned orientation to each key point based on local image gradient directions. To achieve this, the Gaussian smoothed image *L(x, y,σ)* at the key point scale σ is taken to perform all computations in a scale invariant manner. For the sample image *L(x, y)* at scale σ, the gradient magnitude, *m*(*x, y*), and orientation, θ(*x, y*), are pre-computed using pixel difference.

Fig. 7 Image gradient and histogram formation



$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (10)$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \quad (11)$$

Figure 7 shows the angle histogram formation using given image gradients. The magnitude and direction calculation for the gradient are done for every pixel in neighboring region around the key point in the Gaussian blurred image *L*. An orientation histogram with 36 bins is formed with each bins covered 10°. Each sample in the neighboring window added to a histogram bin is weighted by its gradient magnitude and by a Gaussian weighted circular window with σ, that is, 1.5 times that of the scale of the key point. The peak in this histogram corresponds to dominant orientations. Once the histogram is filled, the orientation corresponds to highest peak within 80 % of the highest peak are assigned to the key point. If the multiple orientations being assigned, an additional key point is created having the same location and scale as the original key point for each additional orientations. Figure 8 shows the computation for magnitude and orientation done over constant time.

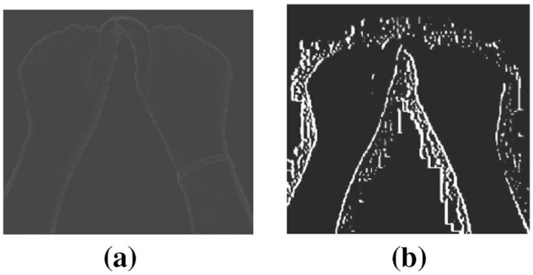


Fig. 8 Orientation assignment. **a** magnitude, **b** phase

Key Point Descriptor

In this phase, the algorithm computes a descriptor for each key point identified. The algorithm computes a descriptor vector for each key point, such that, the descriptor is highly distinctive and partially invariant to the remaining variation, such as, illumination, 3D viewpoint etc. This step is performed on the image closest in scale to the key point's scale. First, the feature descriptor is computed as a set of orientation histogram on 4 × 4 pixel neighborhood. The orientation histogram is relative to the key point orientation data comes from the Gaussian image closest in scale to the key point's scale. As before, the contribution of each pixel is weighted by the gradient magnitude and Gaussian. Histogram contains 8 bin each and each descriptor contains a 4 × 4 array of 16 histogram around the key point. This leads to a SIFT feature vector (4 × 4 × 8 = 128 elements). Further this vector is normalized to enhance invariance to change in illumination. As the dimension of the descriptor is 128, seems high, descriptor with lower dimension than this does not perform as well across the range of matching tasks. Longer descriptor will do better but this rise to an additional danger of increased sensitivity to distortion and occlusion. Therefore SIFT descriptors are invariant to minor affine changes as shown in Fig. 9.

Result and Discussion

This paper mainly focused on the time required for implementing various phases of SIFT algorithm. None of the literature survey focus on the time required for processing of SIFT algorithm. Thus comparison with the existing method could not be done. Table 1 and Figs. 10, 11, 12, 13 and 14 show time required for constructing Gaussian scale space, Differential scale space, finding key

Fig. 9 Computation of key point descriptor

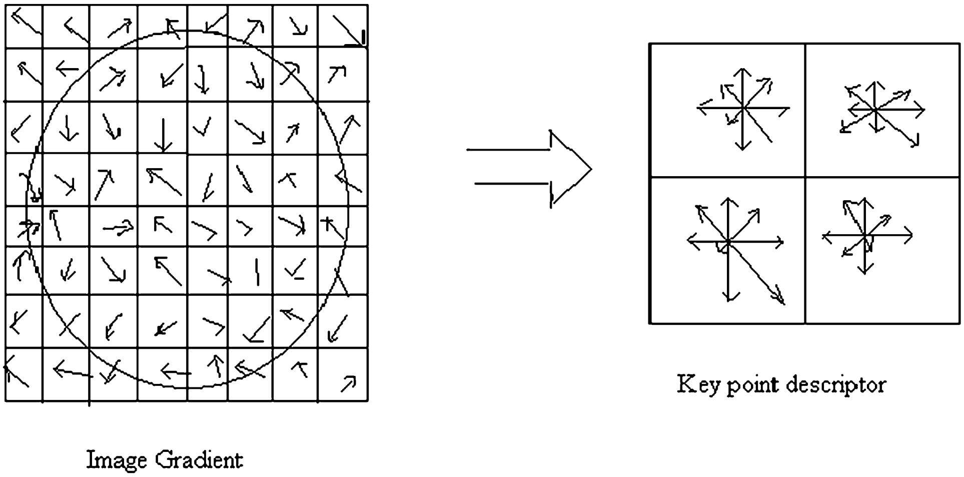


Table 1 Time calculation for various phases of SIFT

ISLG	Time calculation in seconds for				TKPE
	GSSC	DSSC	KP	Descriptor	
A	6.741	0.028	0.490	0.644	668
B	6.699	0.031	0.485	0.780	751
C	6.677	0.029	0.420	0.415	475
D	6.709	0.029	0.509	0.660	669
E	6.698	0.027	0.499	0.567	605
F	6.693	0.028	0.508	0.777	756
G	7.415	0.028	0.397	0.419	485
H	6.752	0.030	0.414	0.471	520
I	6.730	0.027	0.373	0.352	407
J	6.715	0.030	0.288	0.412	457
K	6.759	0.029	0.457	0.500	553
L	6.715	0.028	0.376	0.380	428
M	6.719	0.028	0.458	0.508	544
N	6.746	0.028	0.465	0.578	627
O	6.729	0.027	0.315	0.266	334
P	6.694	0.027	0.475	0.588	630
Q	6.764	0.027	0.488	0.661	677
R	6.725	0.029	0.404	0.458	513
S	6.691	0.029	0.420	0.444	506
T	6.738	0.028	0.462	0.527	566
U	6.712	0.027	0.339	0.290	364
V	6.740	0.029	0.382	0.371	434
W	6.727	0.029	0.360	0.388	440
X	6.700	0.028	0.428	0.575	602
Y	6.743	0.029	0.427	0.400	448
Z	6.763	0.028	0.411	0.544	586

ISLG ISL gestures, *GSSC* Gaussian scale space construction, *DSSC* differential scale space construction, *KP* key points, *TKPE* total number of key points extracted

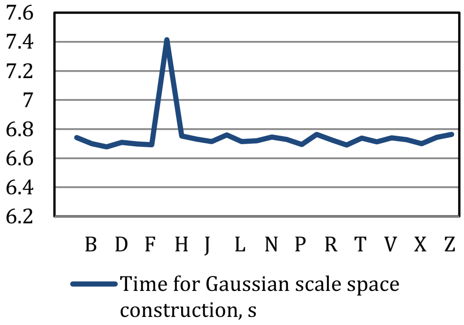


Fig. 10 Time required for Gaussian scale space construction for each ISL gesture

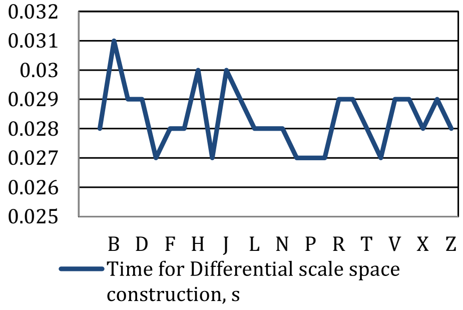


Fig. 11 Time required for differential scale space construction for each ISL gesture

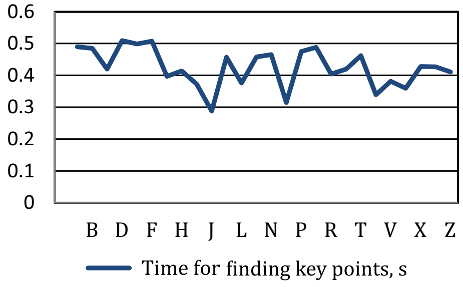


Fig. 12 Time required for finding key points for each ISL gesture

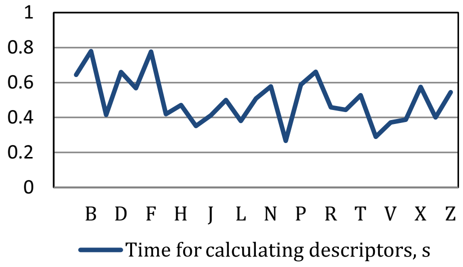


Fig. 13 Time required for calculating descriptors for each gesture

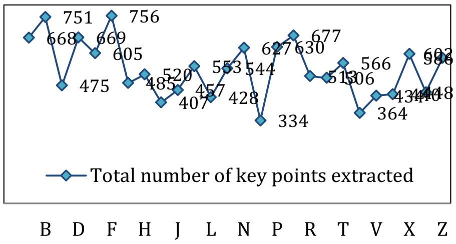


Fig. 14 Total numbers of key point extracted for each ISL gesture

points, descriptors and total number of key points extracted for 26 ISL gestures. The adequate numbers of key points were extracted from each ISL gesture so that it can further been used for matching.

Future Scope

This paper currently focuses on the time required for processing of various phases of SIFT algorithm and also contributed for the number of key points extracted from each ISL gestures images. The future research will be in the direction to change the intensity of each ISL gesture and compute the matching performance of the image based on number of feature extraction from original image and intensity changed image.

Conclusion

The SIFT key points developed in this paper are very useful due to their distinctive features. Large number of key points has been extracted from ISL gesture image. These extracted features are invariant to translation, rotation and scaling. This leads to provide the reliable matching between the images associates with noise and blur. Result shows the computational time required for processing of various phases of SIFT algorithm. Further it also generates the number of key points for all 26 ISLgestures.

References

1. G.R. Sinha, S. B. Patil, Biometrics: Concept and Application, Wiley India Pvt. Ltd (2013)
2. V.J. Arulkarthick, D. Sangeetha, S. Umamaheswari, Sign language recognition using K-means clustered Haar-like features and a stochastic context free grammar. Eur. J. Sci. Res. **78**(1), 74–84 (2012)
3. J. Alon, V. Athitsos, A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Trans. Pattern Anal. Med. Intell. **31**(9), 1685–1699 (2009)
4. M.K. Bhuyan, M.K. Kar, R.N. Debanga, Hand pose identification from monocular image for sign language recognition, IEEE international conference on signal and image processing applications (ICSIPA2011), **3**(12), 378–383 (2011)
5. G. Balakrishnan, P.S. Rajam, Real time Indian sign language recognition system to aid deaf-dumb people, IEEE international conference on computing communication and networking technologies (ICCCNT), 737–742 (2011)
6. P. Chakraborty, S. Mondal, A. Nandy, J.S. Prasad, Recognizing and interpreting Indian sign language gesture for human robot interaction, Int'l conference on computer and communication technology |ICCCT'10|, 712–717 (2010)
7. S. Jothilakshmi, S. Palanivel, V. Ramalingam, A hierarchical language identification system for Indian languages. Digit. Signal Proc. **2**(2), 544–553 (2012)
8. P.V.V. Kishore, P.R. Kumar, A video based Indian sign language recognition system (INSLR) using wavelet transform and fuzzy logic. Int. J. Eng. Technol. **4**(5), 537–542 (2012)
9. P.V.V. Kishore, P.R. Kumar, Segment, track, extract, recognize and convert sign language videos to voice/text. Int. J. Adv. Comput. Sci. Appl. **3**(6), 35–47 (2012)
10. M. Krishnaveni, V. Radha, Classifier fusion based on Bayes aggregation method for Indian sign language datasets. Proc. Eng. **30**, 1110–1118 (2012)
11. A. Magdy, A. Samir, Error detection and correction approach for Arabic sign language recognition, ISSN 978-1-4673-2961, **3**(12), 117–123 (2012)
12. U. Zeshan, M.M. Vasishta, M. Sethna, Implementation of Indian sign language in educational settings. Asia Pac. Disabil. Rehabil. J. **16**(1), 16–40 (2005)
13. S. Majumder, J. Rekha, Indian sign language recognition with global-local hand configuration, IEEE 13th international conference on communication technology (ICCT), 27–33 (2011)
14. T.D. Nguyen, S. Ranganath, Facial expressions in American sign language: tracking and recognition. Pattern Recognit. **45**(5), 1877–1891 (2012)
15. M.P. Paulraj, R. Palaniappan, S. Yaacob, A. Zanar, A phoneme based sign language recognition system using 2D moment invariant interleaving feature and neural network. IEEE Stud. Conf. Res. Dev. **2**(11), 111–116 (2011)
16. A.S. Ghotkar, M. Hadap, R. Khatal, S. Khupase, Hand gesture recognition for Indian sign language, International conference on computer communication and informatics (ICCCI -2012), Coimbatore, India, **2**(4), 9–12 (2012)
17. Y. Quan, Chinese sign language recognition based on video sequence appearance modeling, Fifth IEEE conference on industrial electronics and applications, ISSN 978-1-4244-5046-6/10:1537–1542 (2010)
18. A. Bhosekar, K. Kadam, R. Ganu, S.D. Joshi, American sign language interpreter. IEEE Fourth Int. Conf. Technol. Educ. **6**(12), 157–159 (2012)