ISSN:0975-9646

Performance analysis of Key Frame Extraction using SIFT and SURF algorithms

Suhas Athani[#1], CH Tejeshwar[*2]

Students, Dept. of ISE, B.V.B. College of Engineering and Technology, Hubballi 580031, Karnataka, India

***Abstract*— Growth of videos in today's Internet usage is extensive. Different types of videos will be available in the Internet which among them are lecture videos. Students can make use of these videos, so there is a need to develop an automated system to search the required content only, rather than wasting the time in viewing the complete video. This can be developed into automated system, required steps are: Frame Extraction, Feature Extraction and Key Frame Extraction. In order to extract the Key Frames, Scale Invariant feature transform (SIFT) and Speed Up Robust Features (SURF) algorithms are used. Accuracy and Robustness are the two main important measures that are considered for performance analysis of computer vision algorithms. This paper presents the performance analysis of SIFT and SURF algorithms in Key Frame Extraction of lecture videos and results show that SURF takes less time when compared to SIFT.**

***Keywords*— Frame Extraction, Feature Extraction, Key Frame Extraction, SIFT, SURF**

I. INTRODUCTION

Enormous popularity of the Internet has made e-lecturing more popular. Students may access e-lectures posted on websites anywhere in the world, at any time they wish. They can also be considered for the sake of note taking and for revision. The students are interacting and learning from these e-lectures. Because of availability of extensive multimedia data on web, it becomes difficult for user to judge whether a video is useful by only glancing at the title and find desired videos without a search function within video archive. The user might thus want to find the piece of information he requires without viewing complete video. Video contains huge amount of information at different time intervals. To get the Knowledge from the videos the problem that need to be focused is the elimination of redundant information. The aim is to remove redundant data which reduces the amount of information that needs to be processed. So Key frame extraction is the initial step in any of video retrieval applications. This technique is also called as video summarizing. Key frame is the frame which represents the prominent content and of the video [5]. Key frames are found by the different methods. Methods include sequential comparison between frames, clustering, reference frame based, event/object based. In sequential comparison method, the first extracted key frame is compared with subsequent key frame and this process is carried out until a different key frame is obtained. Simplicity, low computational complexity are the merits of this type algorithms. In clustering algorithms [6], cluster frames and then choose frames closest to the cluster order derivatives. This is carried out to locate edges and corners on the image and is considered as a good method in finding out key centers as the key frames. The merits of these type of algorithms are that they can use generic clustering algorithms directly. The demerit is that they depend on clustering results. In reference frame method, algorithms select a reference frame and reference frame is compared with all other frames and then key frames are extracted. The merit is that it is simple to understand and implement. The demerit is that, it takes more effort to select reference frame.

A. Methods for Key Frame Extraction

1) Scale Invariant Feature Transform: This algorithm was proposed by Lowe in 2004 to solve the image rotation, scaling, and affine deformation, viewpoint change, noise, illumination changes, also has strong robustness [7]. The SIFT algorithm has four main steps: (1) Detection of Scale Space Extrema, (2) Localization of Key Point,(3) Orientation Assignment and (4) Generation of Descriptors [2] [3].

1.1) Detection of Scale Space Extrema: Building Scale Space model can be considered as an initial preparation for finding interesting points. To create Scale Space, original image is considered and blurred out images are created. This way several octaves of original images are obtained. The size of each octaves image is half the previous one. In each octave, images are blurred using Gaussian blur operator. Gaussian blur is applied to each pixel of each octave. It has mathematical expression given as

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

where,
L is a blurred image.
G is Gaussian Blur operator.
I is an image.
x, y are location coordinates.
σ is the scale parameter.
The * is the convolution operation in x and y

The first stage is to find interest points which are called as key points in SIFT framework. The Laplacian of Gaussian (LoG) can be used to find interesting points in an image. It makes use of second points. But calculating second order derivative is sensitive to noise and expensive. Instead of using Laplacian of Gaussian, Difference of Gaussian is used where the difference between two consecutive scales is calculated. As Difference of Gaussian is simple subtraction it is fast and efficient.

Once DoG images have been obtained, local minima/maxima of the DoG images across scales are considered and they are called as key points. Each pixel is compared 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 highest or lowest among all compared pixels, it is selected as a candidate key point.

1.2) Localization of Key Point: The previous step produce alot of key points. These key points lie along an edge or they don't have enough contrast. In both the cases, features are not useful. For the low contrast features, intensities are checked. If the magnitude of the intensity is less than a certain value, it is rejected. Hence, key point candidates are localized and refined by eliminating the key points where low contrast points are rejected.

1.3) Orientation Assignment: Key point orientation is performed as it provides rotation invariance. Gradient magnitude and orientation are precomputed using pixel differences. Equations 2 and 3 are used to calculate gradient magnitude and orientation:

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

$$\theta(x,y) = \tan^{-1}\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)} \quad (3)$$

After the computation, histogram is created in such a way that 360 degrees of orientation are broken into 36 bins, with each bin covering 10 degrees. This type of orientation histogram is computed for all pixels around the key point. In SIFT, the magnitude gradient of an image has to be blurred by an amount of 1.5 * σ and the window size has to be equal to an amount of 1.5 * σ.

1.4) Generation of Descriptors: Gradient magnitude and orientation computed in previous step is now used to compute the local image descriptor for each key point.

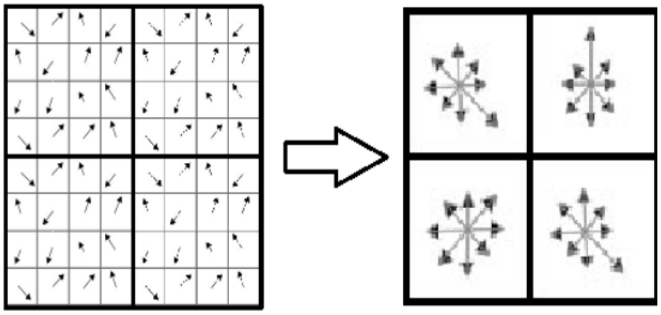

Fig. 1. SIFT Descriptor Generation

In order to get local image descriptor, a 4x4 sample region around a key point is considered. This 4x4 window is broken into four 4x4 window as shown in the right side of the image. For each 4x4 window a histogram of 8 bins are generated [9]. Gradient orientations from 44 are put into respective bins. This is done for all blocks. We get a total of 128 numbers (4x4x8) which are normalised. These 128 numbers form the 'feature vector'. This feature vector will now be uniquely used to identify a particular key point.

2) Speed Up Robust Features: This algorithm was proposed by three people Bay, Tuytelaars and Van [8]. It is a speeded-up version of SIFT. SURF goes a little further and approximates LoG with Box Filter. The algorithm has three main parts: (1) detection of interest point (2) Local Neighbourhood Description and Matching.

2.1) Detection of Interest Points: SURF uses Hessian based blob detector [1] to find interest points. The determinant of Hessian matrix provides two important information. Firstly, it expresses the extent of the response and secondly, it provides the local change around the area [11].

$$H(p,\sigma) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{xy}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix} \quad (4)$$

$L_{xx}(p,\sigma)$,in equation 4 is the convolution of the Gaussian second derivative. Because of use of Box Filters and integral images, there is no need to apply same filter again and again instead Box Filter can be applied directly on the original image and even in parallel. Fig 2. shows Gaussian second order partial derivatives in y-direction and in xy-direction.

2.2) Local Neighborhood Description and Matching: This step is similar to gradient information extraction which is used during SIFT, where descriptors describe the level of intensities with respect to interest point neighborhood. With this information matching of key points is carried out. Use of Haar wavelet responses in x and y direction reduces the time during matching and it increases the robustness.

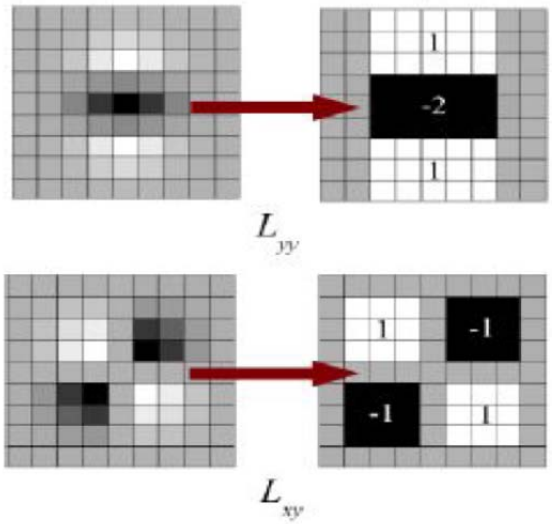

Fig. 2. The Gaussian second orders partial derivatives in y-direction and xy-direction

3) Key Frame Extraction: Key frames are the frames which summarize the entire video. These frames are found by uniqueness among all extracted frames. To find unique frames, dissimilarity between the frames should be calculated [12]. This can be achieved by applying algorithms like SIFT or SURF. The procedure involves following steps. Pre-processing is carried out for the lecture

videos considered as input. Further SIFT and SURF algorithms are used in order to extract key points from these videos. Distance between key points are calculated by using Euclidean distance. The obtained distance is then compared with threshold value and key frame is extracted. Higher the number of matched features more similar the frames are. Slide transition happens during dissimilarity between frames. Match between descriptors are achieved by extracting SIFT or SURF features.

II. Experimental Results

Experiment is conducted on 200 lecture videos collected from YouTube. Key Frames were extracted by applying SIFT and noted down the time and repeatability. Similarly, we extracted Key frames using SURF for same videos. The time and repeatability for SURF are noted down. The table shows the results. Repeatability is calculated by formula,

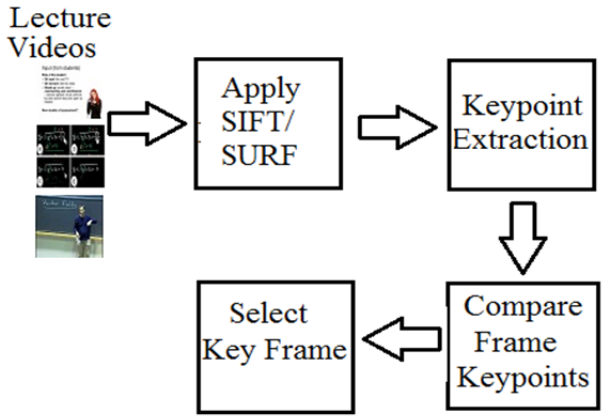

Fig. 3. Diagram representing steps in Key Frame Extraction

$$R = \frac{U+D}{D} \quad (5)$$

where R is repeatability, U is number of unique frames and D is number of duplicate frames.

Table 1. Repeatability comparison of SIFT and SURF algorithms

The repeatability is calculated by formula (5) .The total number of key frames extracted is divided by number of unique frames in the video. The repeatability should be low for the algorithms. If algorithm extracts more duplicates frames, then numerator becomes high, the repeatability of that algorithm will be more. The above table shows the results of conducted experiments on 200 lecture videos.

No of Videos	SURF (Repeatability)	SIFT (Repeatability)
10	0.250	0.190
20	0.268	0.255
30	0.280	0.267
40	0.290	0.278
50	0.259	0.260
60	0.380	0.375
70	0.390	0.396
80	0.420	0.400
90	0.390	0.396
100	0.490	0.43
120	0.320	0.400
140	0.386	0.368
160	0.280	0.254
180	0.389	0.378
200	0.390	0.360

The repeatability is calculated for both SURF and SIFT. The values in the table is inferring that repeatability of both SURF and SIFT are comparably same. The repeatability increases as per the number of videos increases. Fig 4 shows the time comparison of SIFT and SURF against the number of videos subjected to key frame extraction. SIFT generates more features on the images and take more time. Meanwhile, SURF extract less features and take less time for comparing similarity between two frames. SURF is several times faster than SIFT.

Table 2. Efficiency of SIFT and SURF with respect to repeatability.

Algorithm	Repeatability	Time (Hrs)
SIFT	0.360	45
SURF	0.390	28

The Table 2. describes repeatability and time taken by SIFT and SURF for 200 videos. Observing Table 2. we can come to know that SURF has approximately same repeatability as SIFT with better time efficiency.

III. Conclusion

The paper has evaluated the performance of SIFT and SURF in key frame extraction of lecture videos. Based on the experimental results, it is found that the SIFT has detected more number of features compared to SURF but it takes time. The SURF is fast and has good performance as the same as SIFT. Our future work will be running these algorithms on distributed processing frame works like Apache Hadoop.

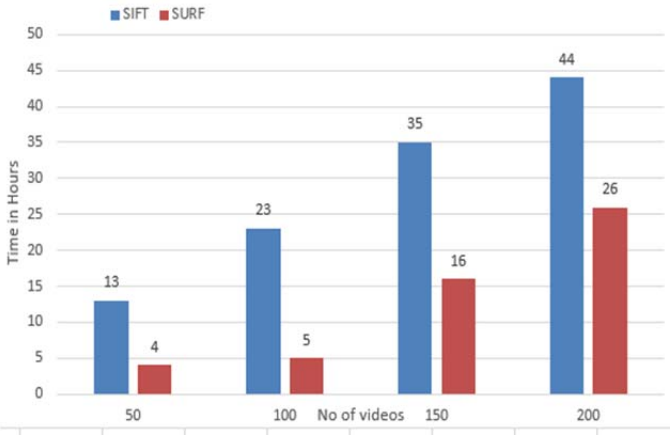

Fig. 4. Time Comparison

References

[1] Tuytelaars, T., Mikolajczyk, K. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision.2007.*

[2] Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision. 60, 2 (2004), 91110.*

[3] Lowe, D.G. Object recognition from local scale-invariant features. *The Proceedings of the Seventh IEEE International Conference on (Kerkyra Greece); 1999. 1150 1157 vol.2.*

[4] K. Mikolajczyk and C. Schmid. *A performance evaluation of local descriptors. In CVPR. volume 2, pages 257 263, June 2003.*

[5] K. W. Sze, K. M. Lam, and G. P. Qiu, A new key frame representation for video segment retrieval, *IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 9, pp. 1148 1155, Sep. 2005.*

[6] Z. Yueting, R. Yong, T. S. Huang, and S. Mehrotra, *Adaptive key frame extraction using supervised clustering, Proceeding in IEEE ICIP, 1998.*

[7] Jian Wu1, Zhiming Cui1, Victor S. Sheng2, Pengpeng Zhao1, Dongliang Su1, Shengrong Gong1. *A Comparative Study of SIFT and its Variants Measurement Science Review, Volume 13, No. 3, 2013.*

[8] Bay, H., Tuytelaars, T., Gool, L.V. SURF: Speeded up robust features. *In Computer Vision ECCV 2006 : 9th European Conference on Computer Vision, 7-13 May 2006. Springer, Part II, 404-417.*

[9] Mortensen, E.N., Deng, H., Shapiro, L. (2005). A SIFT descriptor with global context. *In Computer Vision and Pattern Recognition (CVPR 2005), 20-25 June 2005. IEEE, Vol. 1, 184-190.*

[10] Sebastiano Battiato, Giovanni Gallo, Giovanni Puglisi. SIFT Features Tracking for Video Stabilization. *14th International Conference on Image Analysis and Processing, 2007.*

[11] 11. Dipika H Patel, Content based Video Retrieval using Enhance Feature Extraction. *International Journal of Computer Applications, Volume119 No.19, June 2015.*

[12] 12. D. P. Mukherjee, S. K. Das, and S. Saha. Key frame estimation in video using randomness measure of feature point pattern. *IEEE Trans.Circuits Syst. Video Technol., vol. 7, no. 5, pp. 612620, May 2007.*