Target Identification using SIFT Algorithm

Darshana. A. Vispute , Sachin. D. Pable
E&TC Dept.
Matoshri College of Engineering and Research Center,
Pune University, Nasik, India

***Abstract*—Robotics is the firmest growing technology in the today's scientific world. Robotics nowadays is being applied to various industries for different applications. Also it replaces human process and where people needed do repetitive job. At these places robots are found ideal replacements. In this paper we will discuss one such robot localization algorithm called as SIFT algorithm. Scale Invariant Feature Transform (SIFT) is an algorithm in Computer vision to sense local features of images and then compute on it. This algorithm was published and patented by David Lowe which can robustly perform identification objects even for the clutter and partial occlusion. This SIFT feature descriptor is invariant to constant scaling, orientation and also invariant to illumination changes. This paper explores about the application of Scale Invariant Feature Transform (SIFT) of Target Identification. It also shows simulation results and performance of SIFT for different variations in scale, rotation and iluumination.**

***Index Terms*—Keypoints, image descriptors, local features, keypoint localization, etc**

I. INTRODUCTION

Due to rapid growth in robotics field, there is need to develop some algorithm for target identification using image processing. In indoor environment target identification proves very helpful for monotonous kind of jobs. Mostly target identification is used on AGV's which are normally used for industry applications. However for localization in indoor environments GPS may not prove really helpful. Also it may fail to work in non network areas of GPS. The goal of this project is to develop a technique that will identify targets which will emulate human learning and address searching capability by using SIFT transform in real time application.

Comparison of images to establish degree of similarity has applications in various domains such as content based image retrieval, robot localization, interactive museum guide, image registration etc. image matching becomes a challenging task because of issues such as illumination changes, partial occlusion of objects, differences in image orientation etc. color histograms, responses to filter banks etc. are global image characteristics which are usually not effective for solving real life image matching problems.

The goal is to design a highly distinctive descriptor for each interest point found which would provide meaningful matches for target identification. It also will simultaneously ensure a given interest point will have same descriptor regardless of object position, illumination in environment and image scale. as a result both the steps detection and description rely on invariance of various properties for effective image matching.

The rest of the paper is organised in total VII sections. Section II explores about the literature review which describes the similar work done earlier by different researchers. Section III describes the SIFT algorithm in brief. In section IV the selection of various SIFT features are defined. Section V shows the implementation and simulation results obtained. Section VI explores about the future work that can be done to improve the efficiency of SIFT. Lastly section VII concludes the paper with SIFT results compared with different invariant parameters.

II. LITERATURE REVIEW

Robotics is the fastest growing technology in today's world. Whenever a robot is designed it also has to be localized to the destination also. Depending upon the requirement of application, robot is improved to perform those functions constantly. Various techniques are used nowadays and most of them are developed through the computer. Also changes can easily be made to functionality of robot using a computer. Localization is a critical issue in robotics. Here we are mainly focusing on robot localization application for indoor environment. There is wide variety of localization techniques available. Out of variety of localizations techniques available SIFT [1] proves to be more effective to implement in real time applications. Using global localization for distinctive visual features also proves to be more helpful localization technique by Lowe [2]. Recently many researchers have turned their attention to extract local features from an image, which are invariant to common image transformations and variations or any image matching scheme. There are two major steps involved in local feature based image matching scheme: First step is to detect keypoints (interest points) from an image in a repeatable way. Repeatability is important at this step as robust matching cannot be performed if the detected locations of keypoints on an object vary from image to image. Second step is to compute descriptors for each detected keypoint.

Image matching techniques using local features for target identification is not new in the image processing field. Sven siggelkow [3] used feature histograms for content based image retrieval, who achieved relative success with 2D object extraction and image matching. Mikolajczyk and Schmid [5] used differential descriptors for approximation of point neighborhood for image matching and retrieval. Van Gool [6] introduced the generalized color moments to describe the shape and intensity of different color channels in a local region of image. Schaffalitzky and Zisserman [8] used Euclidean distance between orthogonal complex filters to provide a lower bound on the Squared Sum Differences (SSD) between corresponding image patches. Ledwich and Williams [11] used the scale information of the SIFT features to improve location discrimination. Tamimia et.al. [4] proposed in their work that using few keypoints and comparing image content against database can improve speed of SIFT approach. Valgren and Lilienthal [13] investigated how two local feature algorithms SIFT and SURF approach can be used for localization in outdoor environment that undergo seasonal changes for almost a year. Sukthankar [12] introduced an alternate representation for local image descriptor for SIFT algorithm, which is more distinctive and compact leading to significant improvement in matching accuracy for both controlled and real world condition. David Lowe [1] proposed Scale Invariant Feature Transform (SIFT), which is robustly resilient to different types of image transforms. Mikolajczyk and Schmid [7] reported an experimental evaluation of several different descriptors where they found that the SIFT descriptors obtain the best matching results.

III. SIFT ALGORITHM DESCRIPTION

Based on requirements of local features mentioned in previous section and reported robustness in [7], selection of SIFT [20] approach is selected. SIFT is an approach useful for detecting and extracting local feature descriptors that are reasonably invariant to illumination changes, noise in image, rotation, scaling, and small changes in viewpoint. Before performing any multi resolution transformation via SIFT, it is first converted to grayscale representation. The complete detail explanation of algorithm is found in [20]. A brief description of algorithm is presented in this paper. The algorithm has basic four major stages as mentioned below:

A. Scale space extrema detection:

The SIFT feature algorithm is based upon finding keypoints within the scale space of an image which can be extracted reliably. We want to find points that give us information about the objects in the image. The information about the objects is in the object's edges. So represent the image in a way that gives these edges as these representations extrema points. These keypoints correspond to local extrema og difference-of-Gaussian (DoG) filters at different scales. There is need to identify the points, whose surrounding patches with some scale are distinctive in nature. An approximation to the scale-normalized Laplacian of Gaussian L is mentioned below :

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \tag{1}$$

Where L(x,y,σ) is the scale space of an image, built by convol-ving the image I(x,y) with the gaussian kernel G(x,y,σ).

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

Each Keypoint is represented as (x, y, σ). The DoG image is reprsented as D(x,y,σ) and can be computed from the difference of two nearby scaled images separated by a multiplicative factor k:

$$D(x,y,\sigma)=\big(G(x,y,k\sigma)-G(x,y,\sigma)\big)*I(x,y)$$
$$=L(x,y,k\sigma)-L(x,y,\sigma) \tag{3}$$

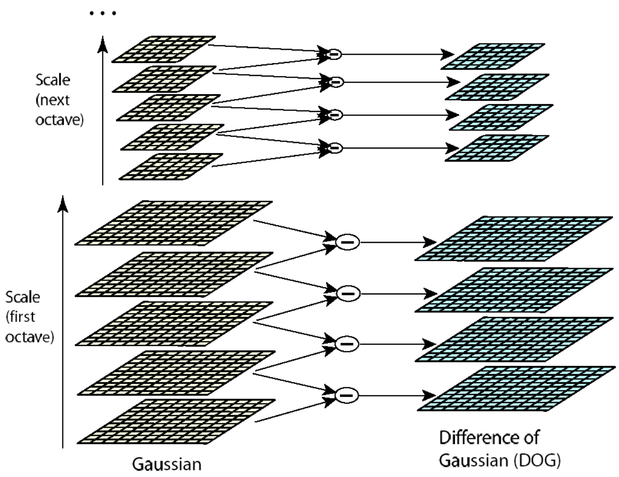


Fig. 1. Difference of Gaussian

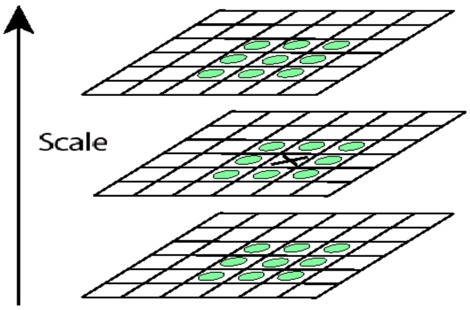


Fig. 2. Extracting keypoints

The convolved images are grouped by octave i.e. 8 components together. An octave corresponds to doubling the value of σ , and the value of k is selected such that the fixed value of blurred images are generated per octave. This also ensures the same number of DoG images are obtained per octave. Keypoints are recognized as local maxima or minima points of the DoG images across different scales. Each pixel in a DoG image is compared to its 8 neighbors at the same scale shown in Fig. 1, and the 9 corresponding neighbors at neighboring scales. If the recognised pixel is a local maximum or minimum then it is selected as a candidate keypoint. X is selected if it is larger or smaller than all 26 neighbors, which is shown in Fig. 2. If the pixel is lower/higher than all its neighbors, then it is labeled as candidate point. Each of these is exactly localized by Taylor expansion series.

B. Keypoint Localization:

There are still many points fom which some of them are not good enough. The locations of keypoints may or may not be accurate. So there is need to eliminate edge points.

Inaccurate localization is mainly caused due to scaling and sampling. Low contrast images are generally sensitive to noise and have strong edge responses. The problem caused due to inaccurate localization is as shown in Fig. 3.

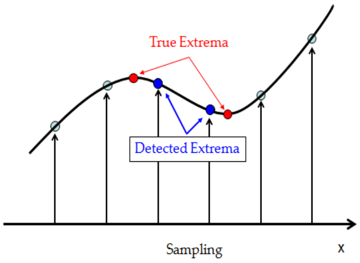


Fig. 3. Figure 1 Mismatch between detection of extrema

The solution is to use the Taylor expansion series as below:

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

The keypoints are then filtered by discarding points of low contrast and points that belong to edges. Equation 5 shows how we can minimize to find accurate extrema.

$$\hat{x} = -\frac{\partial^2 D}{\partial \vec{x}^2}^{-1}\frac{\partial D}{\partial \vec{x}} \qquad (5)$$

If offset from sampling point is larger than 0.5 keypoint should be in a different sampling point. Here first 3D quadratic function is fitted to the local sample points to determine the location of the maximum. The function value at the extremum is used for rejecting unstable extrema with low contrast. The DoG operator has a strong response along edges present in an image, which give rise to unstable key points. A poorly defined peak in the DoG function will have a large principal curvature across the edge but a small principal curvature in the perpendicular direction. The principal curvatures can be computed from a 2x2 Hessian matrix H, computed at the location and scale of the keypoint. H is given by :

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \qquad (6)$$

The eigenvalues of H are in direct proportion with the principal curvatures of D. However eigenvalues are not explicitly computed; instead trace and determinants of H are used to reject those keypoints for which the ratio between principal curvatures is greater than a threshold.

C. Orientation assignment:

Here we assign an orientation to each keypoint, to make descriptor invariant to rotation. The keypoint descriptor can be symbolized in relative with this orientation and therefore invariance to image rotation can be achieved. In this step computation of magnitude and orientation on the smoothed images by Gaussian is also done.

This keypoint orientation is calculated from an orientation histogram of local gradients from the closest smoothed image L (x,y,σ). For each image sample L(x, y) at this scale, the gradient magnitude m(x,y) and orientation θ(x, y) is computed using pixel differences:

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

$$\theta(x,y) = \tan^{-1}\left(\frac{(L(x,y+1)-L(x,y-1))}{(L(x+1,y)-L(x-1,y))}\right) \qquad (7)$$

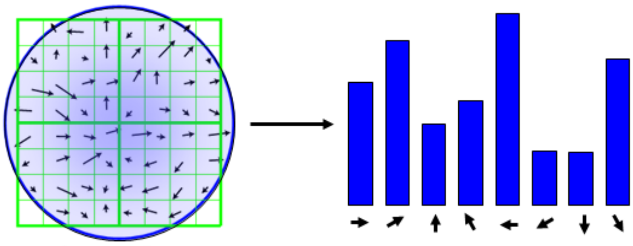


Fig. 4. Figure 2 Histogram of element vector

The orientation histogram has 36 bins covering the 360 degrees of major orientation bins. Each point is added to the histogram weighted by the gradient magnitude m(x, y) and by a circular Gaussian with σ that is 1.5 times the scale of the keypoint. A histogram is designed by quantizing the orientations into 36 bins are as shown in Fig. 4. Peaks in the histogram relate to orientations of the patch. For the exact same scale and location there can be multiple keypoints with diverse orientations. Any histogram peak within 80% of highest peak is assigned to keypoint. The dominant peaks in the histogram are interpolated with their neighbors for a more correct orientation assignment.

D. Keypoint descriptor:

The local gradient of data from the closest smoothed image L(x,y,σ) is also used to create the keypoint descriptor. This gradient image is first rotated with θmax to align it with assigned orientation of keypoint with horizontal direction so as to provide rotation invariance shown in Fig. 5. After this rotation, the region around the keypoint is sectioned into 4*4 square subsections. From each subsections, an 8 bin sub orientation histogram (SOH) is built as shown in Fig. 5.

In order to avoid boundary effects, trilinear interpolation is used to allocate the value of each gradient sample into neighboring histogram bins. Typical keypoint descriptors use 16 orientation histograms aligned in a 4*4 grid. Each histogram orientation histogram has 8 orientation bins each created over a support window of 4*4 pixels. Finally, the 16 resulting SOHs are converted into 128-D vector. These vectors are normalized to unit length to achieve invariance against illumination changes. This vector is called as SIFT descriptor and is used for similarity measuring between two SIFT features.

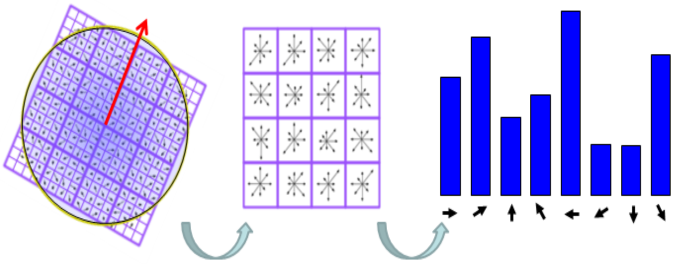


Fig. 5. Element vector 128(4*4*8)

IV. SELECTION OF FEATURES AND MATCHING OF KEYPOINTS

The following key requirements needed to be considered while selecting a local feature for images used in this project:

A. *Invariance:* The feature should be supple to the changes in illumination, image noise, uniform scaling rotation, and minor changes in viewing direction.

B. *Distinctiveness:* The features should provide precise object detection with the lowest possibility of mismatch.

C. *Matching performance:* Whenever input image is given system for identifying target, it should be relatively easy and fast to extract the features and compare the images with the database of local features.

Implementation of target identification can be done on real time basis. Using camera with the system and setting it on continuous video mode is to be performed. Then from that continuous video frames are needed to be grabbed at a specific interval of time. This grabbed frame each time will act as a input test image for matching purpose.

Whenever the input test image is given each of its keypoint is compared with keypoints of image present in the database. At first the Euclidean distance is calculated between each invariant feature descriptor of the test image each invariant feature descriptor of the database image. However the two keypoints with the minimum Euclidean distance (closest neighbors) may not match necessarily because many features from an image may not have correct match in the database of images either because of background clutter or may be the feature was not detected at all. Instead the ratio between the closest neighbors and distance between the second closest neighbors is computed. If the ratio value is greater than 0.6, then the match is rejected. Each time the number of matched keypoints is shown, whenever the reach is maximum we can terminate the execution of the algorithm and system is set for robotic localization and navigation purpose. Navigation is not the scope of this project.

V. SIMULATION RESULTS

The approach described above has been implemented using MATLAB. This implementation can be classified into two aspects: matching and inference. During matching phase locally invariant features (keypoints, orientation, scales and descriptors) from input test images are retrieved using SIFT algorithm and stored in a file. During inference the main objective is to recognize a input test image. A set of local invariant features are retrieved for the test image during inference phase and compared using the metric explained in section V.

An important aspect of SIFT is that it generates large number of features for wide range of scales and locations. The number of features generated mainly depends on image size and content, as well as algorithm parameters. If the obtained test images are of higher resolution then down sampling is necessary to reduce the number of keypoints. Fig. 6 depicts about the working of the target identification system. Camera initialization is done with winvideo adapter in manual trigger mode. Loading of object from database is done following with capturing of image from camera. SIFT features are only extracted when the image are converted to gray. Comparison of these features is performed and matching keypoints are decided. If the value obtained is above threshold then target is said to be identified. If not final destination then load next images from database and again repeat the steps. Factors to be considered while implementing a system in real time:

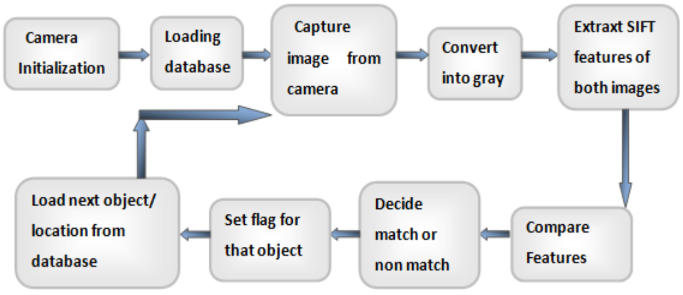

Camera Initialization
Loading database
Capture image from camera
Convert into gray
Extraxt SIFT features of both images
Compare Features
Decide match or non match
Set flag for that object
Load next object/ location from database

Fig. 6. Block diagram of system

1) *Computational expense:*
It should be computationally inexpensive so modern PC has enough power to run it.

2) *Moving background rejection:*
Misclassification can easily occur if the area of moving background is large compared to the object of interest.

3) *Tracking through occlusion:*
Many algorithms still fail to track the image if it is occluded for longer period of time.

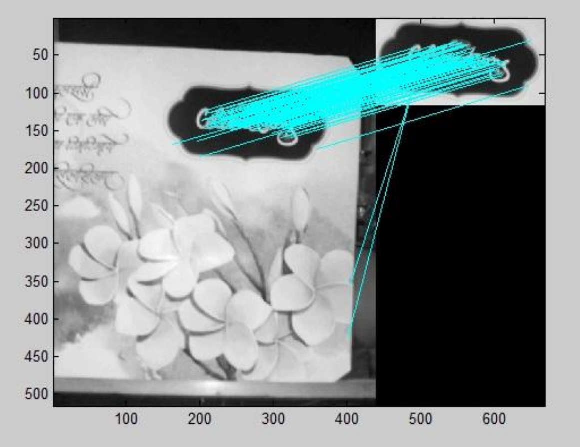
4) *Adapting to illumination variation:*
Real time applications would inevitably have variation in scene illumination, so that a target identification algorithm needs to overcome illumination changes.

5) *Analyzing object motion:*
This could be difficult for non rigid objects such as humans if objects view is not in the right perspective for the algorithm.

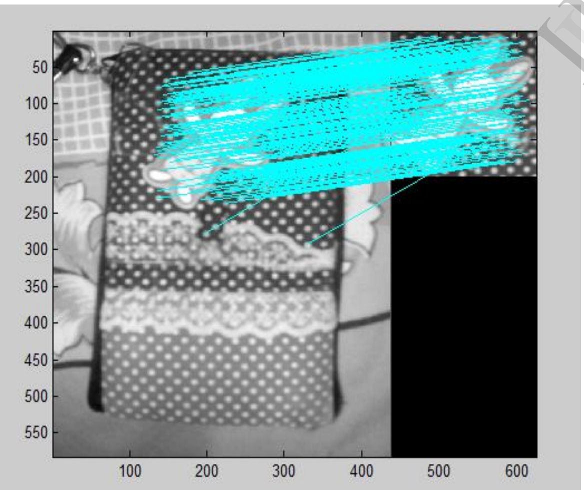
6) *Adapting to camera motion:*
Detecting moving entities from video streams still remains a challenge in this research field.

The SIFT algorithm was implemented first on images from the database only. Scene image and object image are the two images that are to be compared. Object image consist of main image which is stored in database and match of this object image is to be traced in scene image. Initially for simulation scene image is also stored in the database. For real time implementation scene image is that image which is captured by camera. After certain interval of time frame from an video is grabbed, and thus that frame becomes the scene image.

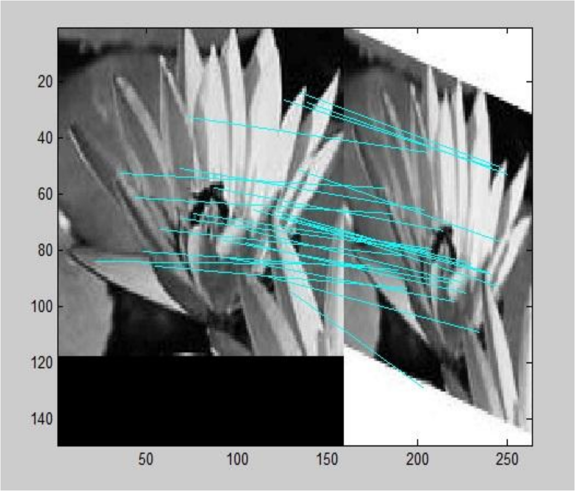
Figure 7 shows the simulation results that are obtained application of SIFT algorithm to the images from the database. This results also depict that SIFT features are invariant to illumination changes as shown in Fig 7 (a). The keypoints found in this were 132 from which 125 keypoints were matched. Scale variation and rotation variation results are shown in Fig 7(b) and Fig 7(c). The keypoints for this result were 404 from which 228 matches were found. For rotation variation 167 keypoints were found and 28 keypoints were matched. This algorithm was also implemented on real time basis which captured live images and then the keypoints were matched successfully with frame interval of 2.17 seconds.



(a) Illumination variation



(b) Scale variation



(c) Rotation variation

Fig. 7. Simulation results obtained for different invariant parameters

Detail analysis can be shown in the graphs which are compared against each invariant parameter and matching points. Fig 8 shows the variations in matching points obtained for different illumination results for scene image. It can be observed that as the illumination increases the no. of matching points is decreased.

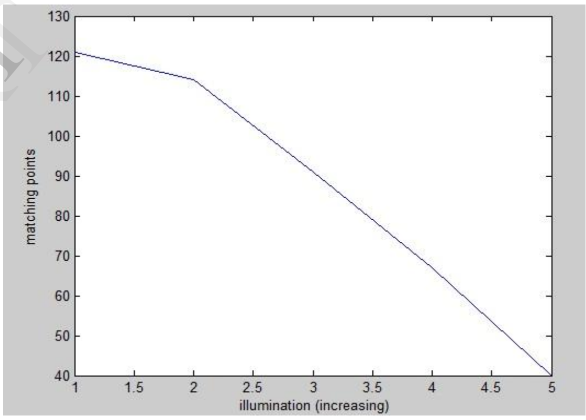


Fig. 8. Illumination variation

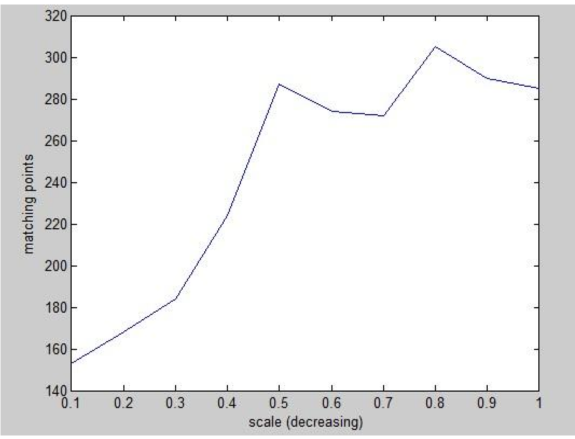


Fig. 9. Scale variation

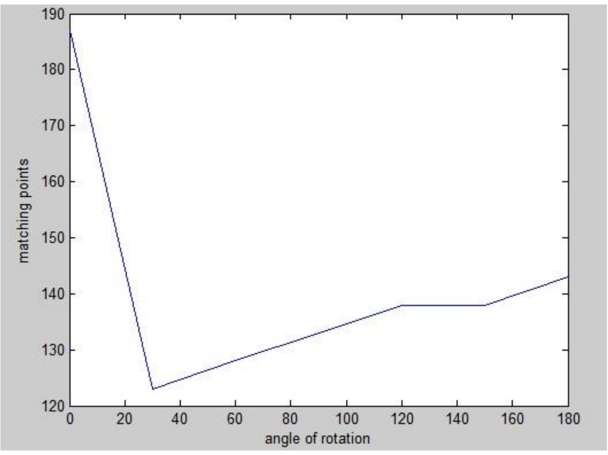


Fig. 10. Rotation variation

Figure 9 shows the graph of scale invariation i.e size of the scene image is varied with constatnt size object image, which represents as the scaling percentage is increased the matching points obtained are decreased. Fig 10 shows the rotation of image for various angles. Scene image is rotated from 0-180 degress for observing results for rotation invariance of SIFT. Worst results obtained were for 30 degrees of rotation of scene image. Also performance of SIFT was checked on image in which all above parameters were varied and still the results were obtained that 68 matches were found from 132 keypoints, which depicts the robustness of SIFT algorithm.

TABLE I. COMPARISON OF MATCHING POINTS

Variation parameters	Best case	Worst case
Illumination	Intensity : 1 Matches : 121	Intensity : 5 Matches : 40
Scale	Size : 0.8 Matches : 305	Size : 0.1 Matches : 153
Rotation	Angle : 0 degrees Matches : 305	Angle : 30 degrees Matches : 153

Table 1 depicts the comparison of matching points obtained for best case and worst case. As a result there is need to develop some technique which would enhance the performance of SIFT algorithm. Various techniques can be used to improve the performance of SIFT which are explored in next Section. According to application the suitable technique can be chosen.

VI. FUTURESCOPE

In order to improve the performance of the SIFT algorithm in various aspects, further research can be done. Some improvement parameters are mentioned below:

A. Super resolution of input images:

Super-Resolution (SR) is a technique by which a number of Low Resolution images are combined into a single High Resolution image. This has a greater resolving power. Super Resolution is not only useful to enhance the resolving power of an image; also it can reduce the aliasing noticeably. At an initial level it may take a longer time for execution for higher resolution images but improvements can be done to increase the execution speed.

B. Use of DWT and SVD:

This method use to enhance the quality of an input image [14]. The enhancement is done both with respect to resolution and contrast. To upsurge the resolution, the method is to use DWT. These transform converts the given input image into four sub-bands, from which one is of low frequency and the rest all are of high frequency. The High Frequency components are interpolated using predictable interpolation techniques. Then we use IDWT to associate all of the interpolated high frequency and low frequency components. To upsurge the contrast level, use of SVD and DWT can be done.

C. Global features:

An image may have many keypoints that are similar to each other locally. These multiple similar areas may produce indistinctness while matching of local descriptors. The local invariant features of SIFT can be improved by computing global features of image.

D. Difference of mean:

As mentioned in [8], instead of using the DoG of images for finding out signal space extrema, we can also use Difference of Mean (DoM) images to approximate DoG. For finding out DoM of images first need is to compute an integral image. An integral image can be computed from an input image I as mentioned below:

$$J(x, y) = \sum_{x'=0}^{x}{}' \sum_{y'=0}^{y} I(x, y) \tag{8}$$

The mean of rectangular region of an integral image can be computed very efficiently and doesn't depend on size of the region.

VII. CONCLUSION

This project proved to be a compelling exploration of applications of image processing techniques. SIFT is the state-of-the-art algorithm for extraction of locally invariant features and this project makes us understand about the multi resolution image analysis and its application in target identification. Efforts for this project resulted in robust target identification implementation which performs really well in real time applications. Also results for different invariant parameters of SIFT were observed and showed graphically. Comparison of these parameters with matching points was done which proved to be more helpful to find best case and worst case for SIFT algorithm.

REFERENCES

[1] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Computer Science Department University of British Columbia Vancouver, B.C., Canada, lowe@cs.ubc.ca, January 5, 2004.

[2] S. Se, D. Lowe, J. Little,"Global Localization using Distinctive Visual Features," Proceeding of 2002 IEEE, intl. conference on robots andsystem, Switzerland.

[3] S. Siggelkow."Feature Histograms for Content-Based Image Retrieval".PhD Thesis, Albert-Ludwigs-University Frieiburg, December 2002.David G. Lowe.Distinctive Image Features from Scale-

Invariant Keypoints. International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[4] H. Tamimia, H. Andreasson, A. Treptowa, and T. Duckettc, A. Zella," Localization of Mobile Robots with Omnidirectional Vision using Particle Filter and Iterative SIFT," 2006 Elsevier, Robotics and Autonomous Systems vol. 54, pp. 758–765,2006.

[5] Mikolajczyk, K., Schmid, "An Affine Invariant Interest Point Detector". In: ECCV, (2002) pp. 128-142.

[6] V. Gool, T. Moons, and D. Ungureanu."Affine photometric invariants for planar intensity patterns". In ECCV, pp. 642-651, 1996.

[7] Mikolajczyk K., Schmid, C.: "Indexing Based on Scale Invariant Interest Points". In: ICCV, (2001) pp. 525–531.

[8] Schaffalitzky F., Zisserman, "A.: Multi-view Matching for Unordered Image Sets", In: ECCV, (2002) pp. 414-431.

[9] D. G. Lowe. "Object recognition from local scale-invariant features". International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157.

[10] M. Grabner, H. Grabner, and H. Bischof. "Fast approximated SIFT". Proc. ACCV 2006, Hyderabad, India.

[11] L. Ledwich, S. Williams," Reduced SIFT Features For Image Retrieval and Indoor Localization," ARC Centre of Excellence for Autonomous Systems School of Aerospace Mechanical and Mechatronic Engineering University of Sydney, NSW, 2006, Australia.

[12] Y. Ke1, R. Sukthankar,"PCA-SIFT: A More Distinctive Representation for Local Image Descriptors", School of Computer Science, Carnegie Mellon University; 2 Intel Research Pittsburgh.

[13] C. Valgren, J. Lilienthal," SIFT, SURF & Seasons: Appearance-based Long-term Localization in Outdoor Environments," AASS Research Centre, Dept. of Computer Science, Orebro University, SE-70182, Orebro, Sweden.

[14] Ganesh sai P., Habibullah K ,, Bhavana.k 3,Muralidhar. , Tulasi sai K.," Image enhancement using Wavelet transforms and SVD," International Journal of Engineering Science and Technology (IJEST), Vol. 4 No.03 March 2012, pp. 1080-1087.