A Blind Assistance and Navigation System using SIFT Algorithm for Indoor Environments

BY

ANIQUA NUSRAT ZEREEN (09201031)

SADIA BINTE KABIR (09201033)

SUPERVISOR

MD. ZAHANGIR ALOM

Signature

Date

Acknowledgement

We thank all who in one way or another contributed in the completion of this thesis.

We are so grateful to the department of computer science and engineering of BRAC University. We also thank to the Professors and Lecturers of the CSE program.

Our special and heartily thanks to our supervisor, lecturer Md. Zahangir Alom who encouraged and directed us. His challenges brought this work towards a completion. It is with his supervision that this work came into existence.

We appreciate our parents for supporting us in every stage of our life.

Contents

List of tables and Figures

Abstract

When blind people are trying to navigate the city streets, they can get assistance from a speaking GPS-enabled Smartphone, just like everyone else. Once they move indoor and lose access to the required satellite signals. While there are some indoor navigation systems that require things like radio-frequency tags to be strategically placed around the building, it's currently unrealistic to expect to find such systems installed in many places. However, we are going to propose blind assistance and navigation system for indoor environment. It required a computer with camera loaded up with digital images of the room. Features and key points of the images have been extracted using Scale Invariant Feature Transform (SIFT)[1] algorithm from parent and child images. Database has been created with the key feature points and sound files for individual child images. Moreover, our Matching algorithm is modified K-D tree[6] known as Best Bin First (BBF) method. It has been used to match the key points of parent images with database to recognize the desire child images. System has been developed to determine the desired object and guide them along it via verbal cues from the database. Experimental results are demonstrated this proposed method is accurate and reliable to solve the promising and challengingBlind Assistance and Navigation Systemusing SIFT Algorithm for Indoor Environments.

Keywords: Blind navigation, SIFT, Blind assistive system, Features descriptors, feature matching.

1. Introduction

According to recent estimates of the World Health Organization, 39 million are blind. The unfortunate fact is in developing countries there are not enough facilities for blind people so that they can interact with the environment. As the price of camera is decreasing day by day, so we can make some efficient device for blind people by using it. Primarily we are using computer with a camera and image processing technique for object recognition which will help to make blind people conscious about their environment.

1.1 Motivation

Blind people do faces a lot of difficulties during moving one place to another. Most of the time there are no people to assist them or look after them. They face problem in searching their necessary things in indoor environment. And sometimes they do not know the location of particular things has changed or not. So we have built a system which will help blind people to assist themselves to move comfortably in indoor environment.

1.2 Related works

There are some Technologies for blind people such as

1. **Color based object detection:**Computer vision system that helps blind people to find lost objects[5]. Which combine color and SIFT-based object detection with sonification to guide the hand of the user towards potential target object locations. This way, this system is able to guide the user's attention.

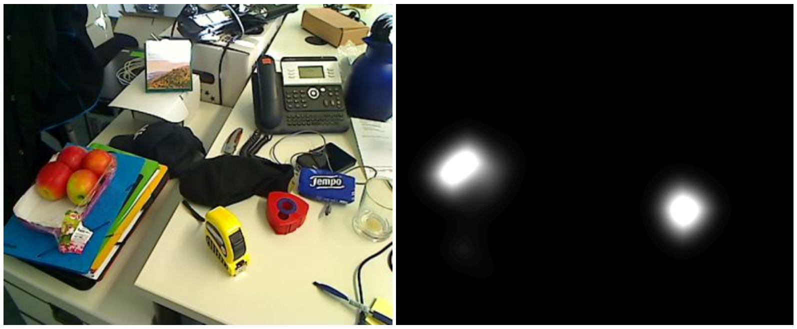


Figure 1: Red color object detection

Disadvantages of 1st system: Depending on light intensity of color at different time the output can be different and inappropriate.

2. **Object detection using GPS system:** Another system represent alternative localization algorithm using a body mounted single camera[2]. Instantaneous accurate localization and heading estimates of the person are computed from images as the trip progresses along a memorized path. A first portable prototype has been tested for outdoor as well as indoor pedestrian trips. Experimental results demonstrate the effectiveness of the vision based localization to keep the walker in a navigation corridor less than one meter width along the intended path.

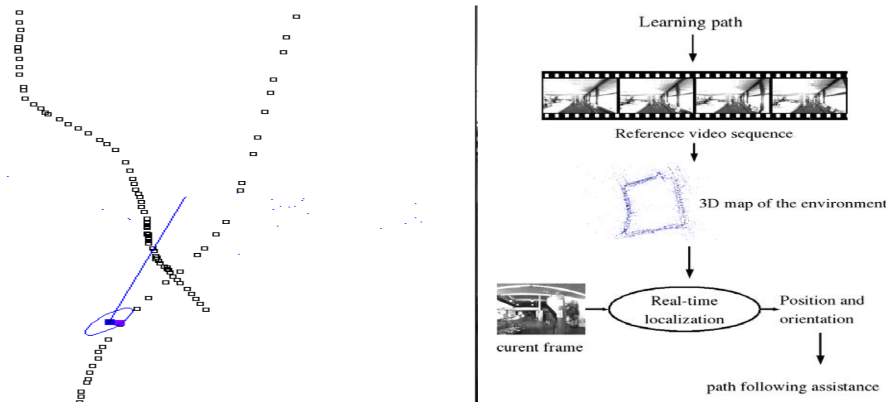


Figure 2: Object detection using GPS

Disadvantages of 2nd system: GPS system to provide accurate information to a traveler, it must be able to receive a signal from a minimum of four satellites. Inclement weather or tall buildings can interfere with signal reception. Another disadvantage of GPS is its dependence on battery power. Even though using a battery enables a GPS system to have portability, a battery can fail without warning. GPS-ready products mentioned above use digital commercial mapping from a third party, and maps become obsolete if not updated periodically. A GPS receiver cannot inform a blind traveler of a drop-off on the sidewalk or a rise in elevation; therefore a blind pedestrian will not be able to use GPS to determine ordinary obstacles that may lie ahead or let him or her know when it is safe to cross the street at a busy intersection. In indoor environment GPS do not work properly.

1.3 The methodology

The problem of blind assistive system is it needs a strong database from input images and video. This approach usually employs machine learning techniques to gain knowledge about objects in an indoor environment. To achieve this, a certain amount of object's images must be collected to process in advance and stored into database in the computer. Those object's images are used as child images to match parent images. In this thesis, we have proposed a set of algorithms to efficiently implement real time object recognition system from arbitrary input object's images and video.

1.4 Outlines

This thesis is organized in 4 chapters including introduction. Chapter 2 attempts to find answer to the first question proposing and investigating the entire algorithm used here step by step, to give a shape to proposed system, blind assistive technology where an input image is processed to detect object. Chapter 3 is concerned about experimental results of this system and it's discussions in detail. Chapter 4 summarizes the main contributions of this thesis. Finally, further research directions are suggested.

2.OverallSystem Implementation

2.1 Problem definition

Real time video has been captured using A4tech 5 mega pixel webcam and it has been processed into frames. These frames have been used as input images for the system. SIFT algorithm, one of the most popular algorithms in the description and matching of 2D image featurehas been used to output key point descriptors of these input images. Later on these descriptors have been used as input of SIFT matching algorithm and output matched images along with sound files.

First of all we have converted the input image which is in RGB, to GRAY image because SIFT is designed for gray images. Images are then smooth out for getting clear edges to detect the key points which are in scale-space extrema step. To eliminate some problematic key points we have used the step key point localization to select key points which are accurate on the basis of scale and sampling, low contrast and edge key points. This is how we have filtered all key points and proceed further with the next step orientation assignment, where we have assigned direction using gradient. This helped SIFT to be invariant to rotation of images. Lastly we computed descriptor vector for each key point such that the descriptor is invariant to the remaining variations like illumination, 3D viewpoint, etc. This step has been performed on the image closest in scale to the key point's scale.

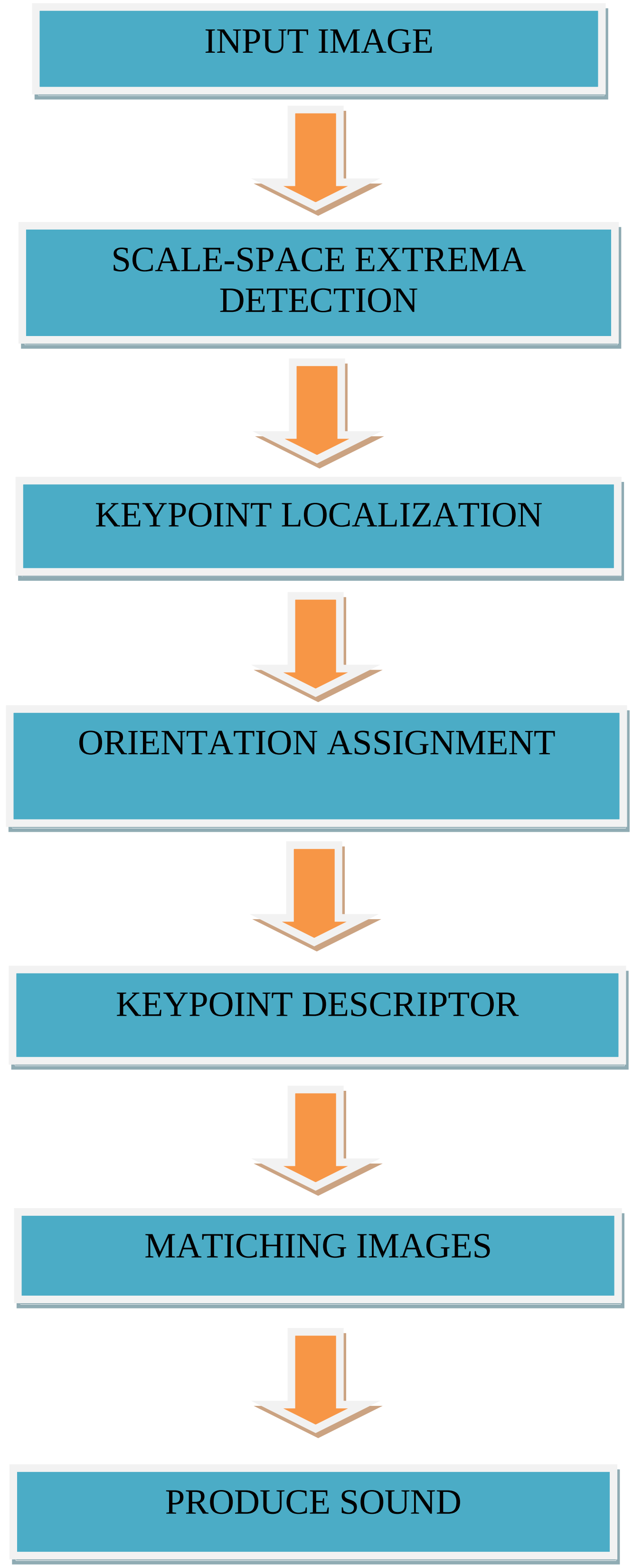


Figure 3: Overall System flow

2.2 Proposed approach

When SIFT algorithm is applied to an object, multiple features are created with respect to different key points on that object. The goal of the descriptor is to uniquely identify the object. Before a feature vector is formed, there is a series of steps, SIFT performs to calculate the vector. These steps give the algorithm its name because it transforms image data into scale-invariant coordinates relative to local features. Following are the major stages of computations, used to generate the set of image features.

2.2.1 SIFT algorithm

Figure 4: Input image

Scale Invariant Feature Transform:

Scale-space extrema detection

This is the step, key points using SIFT are detected. Input image is convolved with Gaussian filters at different scales and difference of successive Gaussian-blurred (DoG) images are taken.

$$D(x, y, \sigma) = L(x, y, k_i\sigma) - L(x, y, k_j\sigma)$$

Where $L(x,y,k\sigma)$ is the convolution of the original image $I(x,y)$ with the Gaussian blur[7] $G(x,y,k\sigma)$ at scale $k\sigma$, i.e.,

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

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

Here K_i is selected as a fixed number of convolved images per octave.

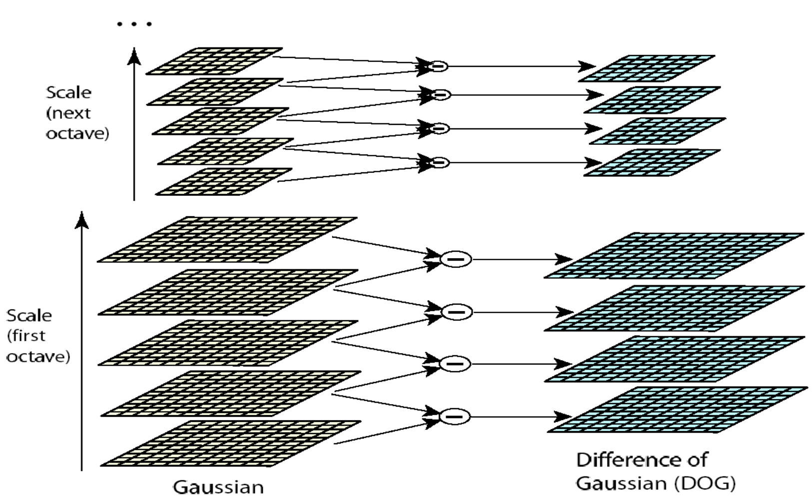


Figure 5: Difference of Gaussian

After DoG,pixels of the input image are compared to its eight neighboring pixels at the same scale and nine corresponding neighboring pixels in each of the neighboring scales; total of 26 pixel comparison has been done. If any pixel gets the maximum or minimum value among all compared pixels, it is selected as a candidate key point.

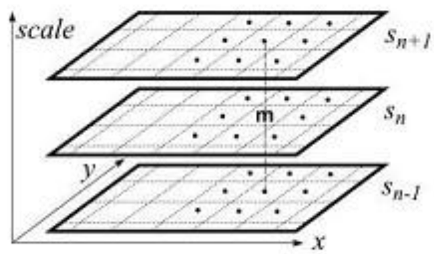


Figure 6: An extrema is defined as any value in the DoG greater than all its neighbors in scale-space

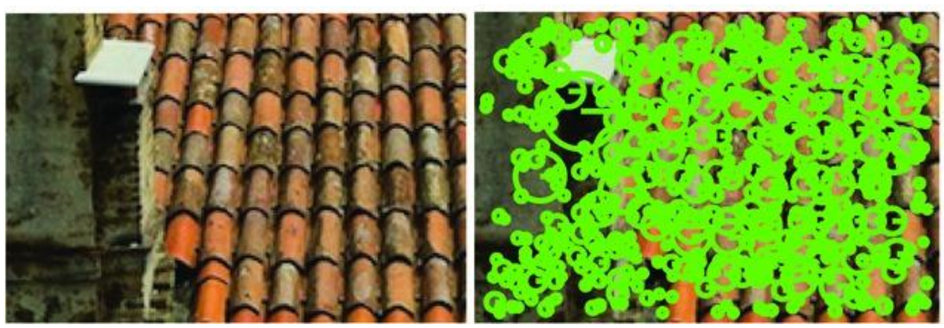


Figure 7: Detected key points

Key point localization

After scale space extrema is detected it outputs too many key points which are unstable. Then the SIFT algorithm discards some problematic key points which are

- **Inaccurate localization due to scale and sampling.**
- **Low contrast key points.**
- **Key points detected on edges.**

Inaccurate localization due to scale and sampling: For accurate position, interpolationof nearby data is computed. The interpolation is done using the Taylor expansion of the Difference-of-Gaussian scale-space function, $D(x,y,\sigma)$ with the candidate key point as the origin.

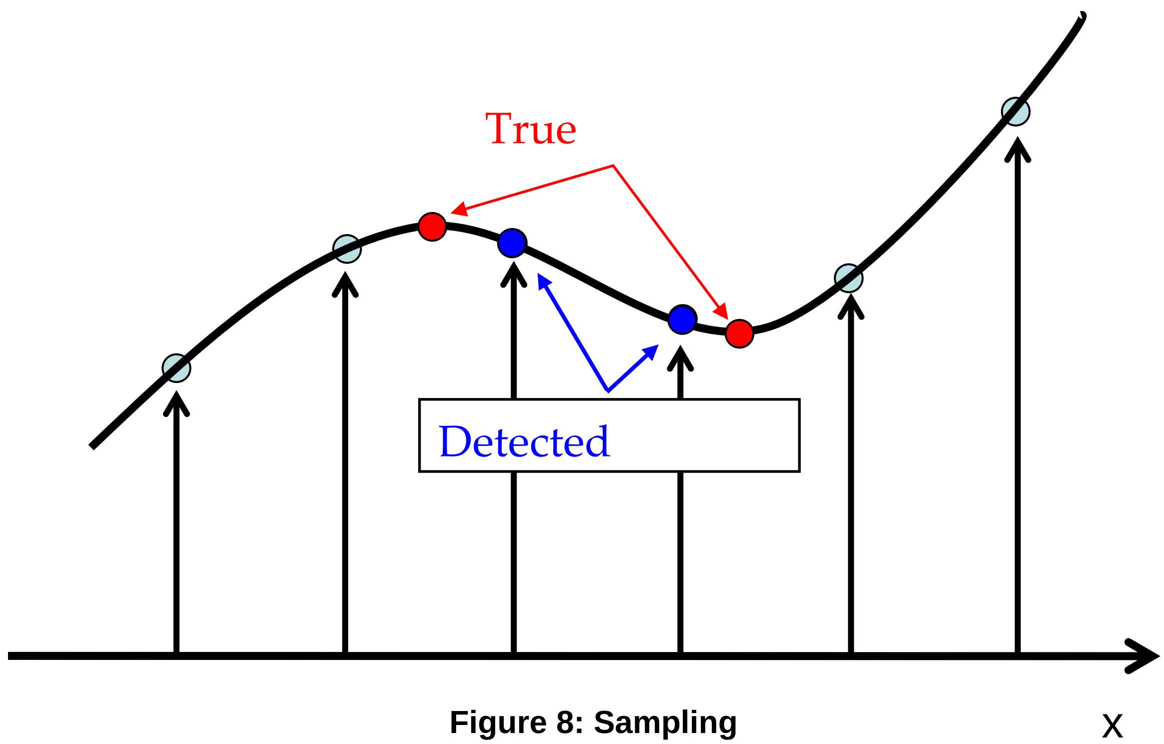


Figure 8: Sampling

This Taylor expansion is given by:

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

Where D and its derivatives are evaluated at the candidate key point and $x= (x,y,\sigma)$ is the offset from sampling point. The location of the extremum, x', is determined by taking the derivative of this function with respect to x and setting it to zero. If the offset x' is larger than 0.5 in any dimension, then that is an indication that the extremum lies closer to another candidate key point. In this case, the candidate key point is changed and the interpolation performed instead about that point. Otherwise the offset is added to its candidate key point to get the interpolated estimate for the location of the extremum.

Minimization to find accurate extreama

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

Ifx′> 0.5, key point is discarded.

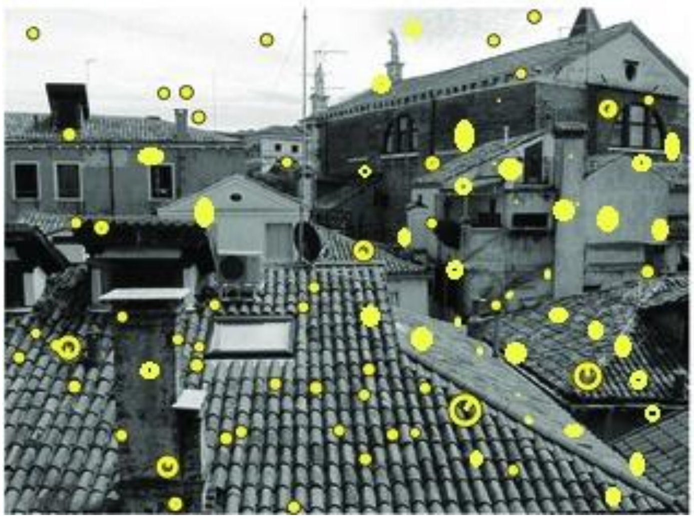


Figure 9.a: After discarding inaccurate localized key points

Low-contrast key points: To discard the key points with low contrast, the value of the second-order Taylor expansion *D(x′)* is computed at the offset *x′*. If this value is less than 0.03, the candidate key point is discarded. Otherwise it is kept.

If |*D(x′)*|< 0.03 then key point is discarded.

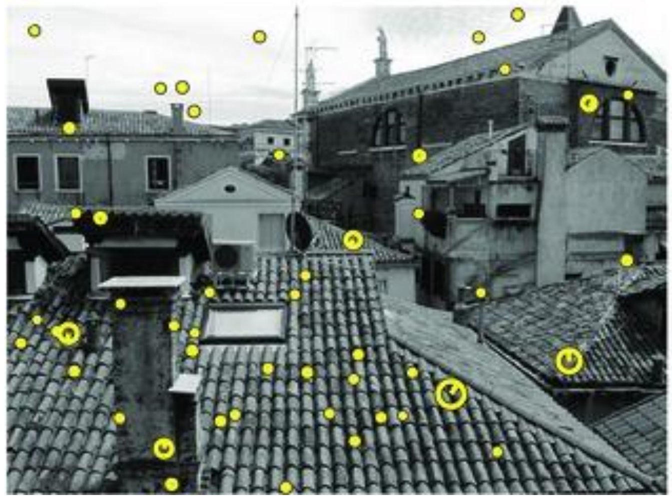
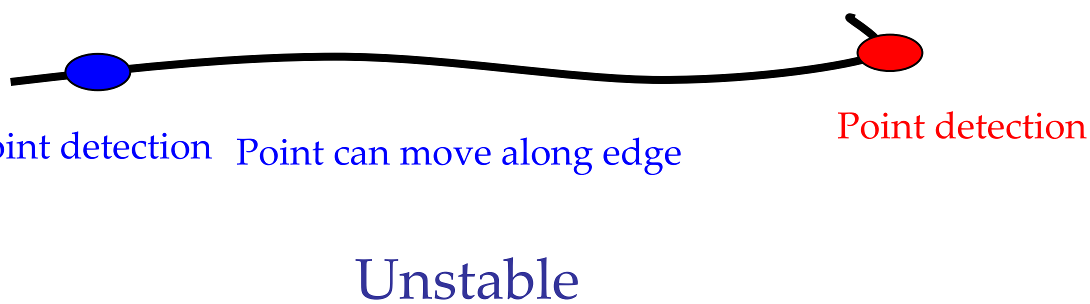


Figure 9.b: After discarding low contrast key points

Key points detected on edges: In order to increase stability, we need to eliminate key points that have poorly determined locations but have high edge responses.



For poorly defined peaks in the DoG function, the principal curvature across the edge would be much larger than the principal curvature along it. These values have been calculated by using Hessian matrix.

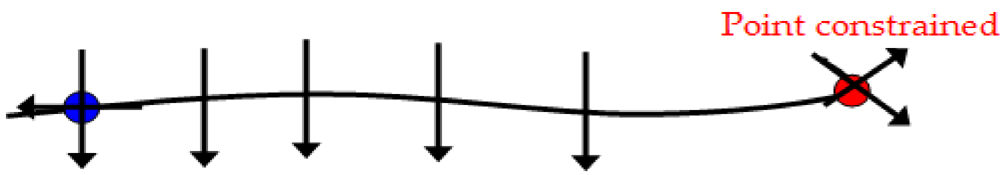


Figure 10: Key points detected on edge

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

The eigenvalues of *H* are proportional to the principal curvatures of *D*. It turns out that the ratio of the two eigenvalues, if α is the larger one, and β the smaller one, with ratio $r = \alpha/\beta$, is sufficient for SIFT's purposes.

The trace of *H*, i.e., $D_{xx} + D_{yy}$, gives us the sum of the two eigenvalues

The determinant, i.e., $D_{xx} D_{yy} - D^2_{xy}$, yields the product.

The ratio $R = Tr(H)^2/Det(H)$ can be shown to be equal to $(r+1)^2/r$, which depends only on the ratio of the eigenvalues rather than their individual values.

$$\frac{\lambda_{\max}}{\lambda_{\min}} < r \Leftrightarrow \frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r}$$

Where if r<10, key point is discarded.

Orientation assignment

In this step, each key point is assigned one or more orientations based on local image gradient directions. For an image sample L(x, y) at scaleσ, the gradient magnitude, m(x, y), and orientation, θ(x, y), are computed using pixel differences:

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

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

The magnitude and direction calculations for the gradient are done for every pixel in a neighboring region around the key point. An orientation histogram[8] with 36 bins is formed, with each bin covering 10 degrees. Once the histogram is filled, the orientations corresponding to the highest peak and local peaks that are within 80% of the highest peaks are assigned to the key point.

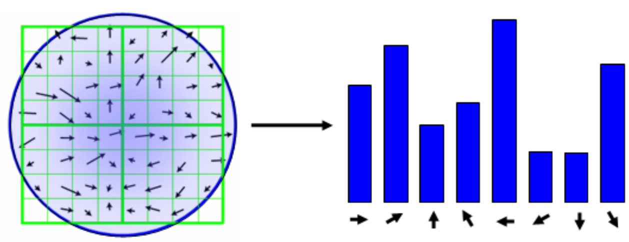


Figure11: Orientation measure

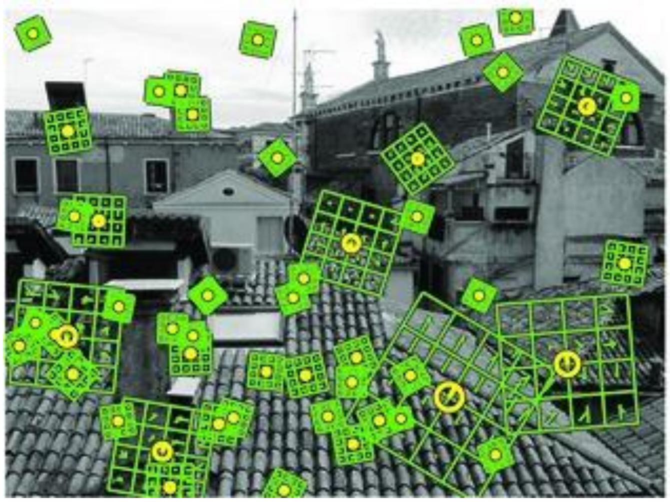


Figure 12: After orientation assignment

Key point descriptor

In this step descriptors are created to help SIFT to be invariant in such illumination, 3D viewpoint, etc. Histograms have been created of 4x4 pixels with 8 bins each. Since there are 4 x 4 = 16 histograms each with 8 bins the vector has 128 elements. First, the vector is normalized to unit length. A change in image contrast in which each pixel value is multiplied by a constant will multiply gradients by the same constant, so this contrast change is canceled by vector normalization. Same goes with brightness, noisiness, sharpness, etc. change in the image. Therefore, the descriptor is invariant to affine changes in illumination. However, non-linear illumination changes can also occur due to camera saturation or due to illumination changes that affect 3D surfaces with differing orientations by different values. These effects can cause a large change in relative magnitudes for some gradients, but are less likely to affect the gradient orientations. Therefore, reducing the influence of large gradient magnitudes by thresholding the values in the unit feature vector to be no larger than 0.2, and then renormalizing to unit length. Thus all the descriptors are stored in database.

2.2.2 Matching algorithm

SIFT Matching algorithm

This section is about the object recognition part. SIFT matching algorithm is used, which is based on modified KD-tree search algorithm, known as Best Bin First (BBF)[6] method. It is used tofind the nearest neighbor key point for a large fraction of the queries, and a very close neighbor key point in the remaining cases. The technique has developed the recognition system, which is able to detect complex objects in real, cluttered or occluded scenes.

The features vector of the query image to the features vectors of all the images in the database has been compared. The method computed Euclidean Distance of the feature vectors for the both images. The smaller the distance, the more similar the images are. To recognize an object it needed at least 3 key points' entries in a bin, bin must be sorted into decreasing order of size. Least-squares solution is performed considering parameters: translation, rotation, scale and stretch for each key point of the query image to the images stored in database.

The affine transformation of a query point $[x\ y]^T$ to an image point $[u\ v]^T$ can be written as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

Where the query translation is $[t_x\ t_y]^T$ and the affine translation, rotation, scale, and stretch are represented by the m_i parameters.

We wish to solve for the transformation parameters, the equation above can be rewritten as

$$\begin{bmatrix} x & y & 0 & 0 & 1 & 0 \\ 0 & 0 & x & y & 0 & 1 \\ & & \cdots & & & \\ & & \cdots & & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u \\ v \\ \vdots \end{bmatrix}$$

This equation showed a single match, but any number of further matches can be added, with each match contributing two more rows to the first and last matrix. At least 3matches are needed to provide a solution.

We can write above linear system as

$$Ax = b$$

The least-squares solution for the parameters **x** can be determined by solving the corresponding normal equations,

$$x = [A^T A]^{-1} A^T b$$

This minimizes the sum of the squares of the distances from the projected query image locations to the corresponding stored image locations.

Using above equations, system outputs two matrices containing matches and scores.

Table 1: Result from SIFT matching algorithm

Matches: image1	0	4	5	23	12	45	7	1	9
Matches: image2	0	8	4	12	4	5	1	6	21
Scores	0	3434	4223	12342	23288	93458	23223	42358	24350

In the ‘matches’ matrix, the number of columns represents the number of ‘matches’ the matching algorithm found between features of two images. The first row stores the key point number of the first image, and the second row is the corresponding ‘matched’ key point number from the second image. So, in the above figure, it shows that key point ‘4’ from the first image matched with key point ‘8’ from the second image, and the Euclidean distance between these two matched key points is 3434. The numbers of matches are ‘9’ since column numbers represent the number of matches. Following figure shows a visual example of this matching. Note that the drawing is not drawn to scale.

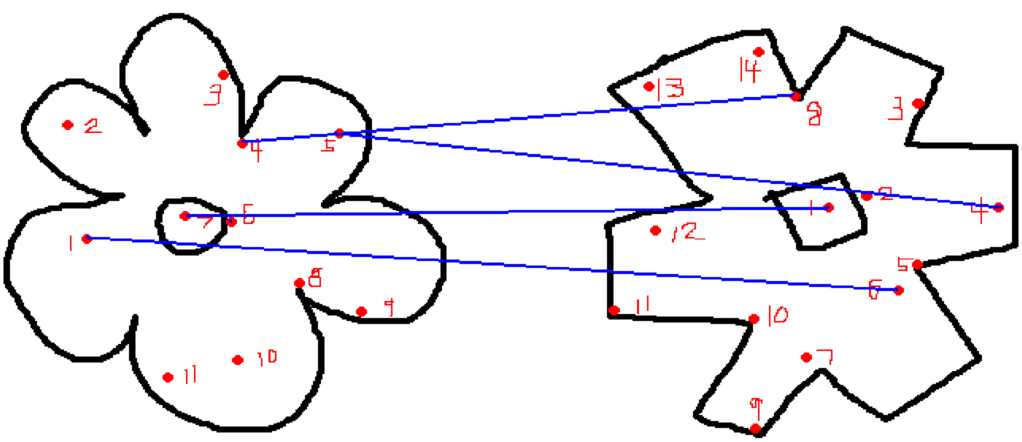


Figure 13: Visual example image of the matching algorithm and numbered key points

Experiment of the input image below

Figure 14: left image is the child image and right image is the parent image to match

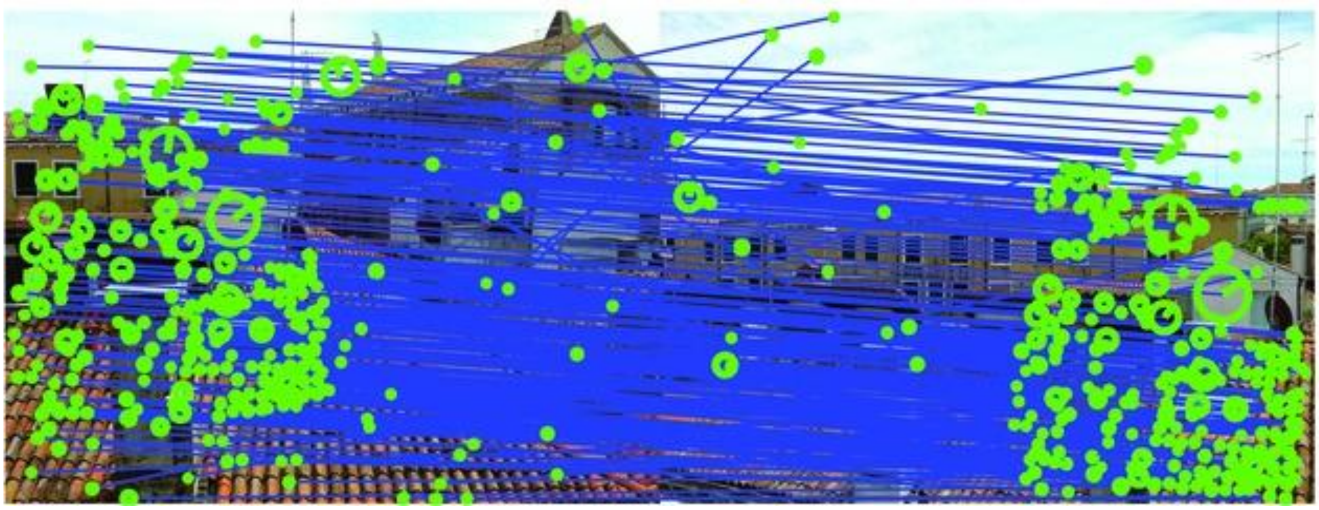


Figure 15: Matched key points using SIFT matching algorithm

2.2.3 Play verbal cues

In this section, system outputs the sound with respect to the stored child image matched from the database. Individual sound file were stored along with each child image in database. Sound files are named as the image files names. Concatenation method has been used to match the stored image file name and the sound file name excluding both file's format characters.

3.Results and Discussion

3.1 Setup for Experiment

Matlab tool has been used for the system implementation. A4tech 5 megapixel webcam has been used for this thesis. The captured video is needed in avi format and then it has been passed to the system which converted that video into frames for further processing. Database is stored with all the objects of a particular indoor environment.

3.2 Discussion on Database

System processes frames from video or still images and detect objects. It detects those objects which are given in database as child images. System plays sound to make the blind people aware of that object. Sound clips are needed in wav format.

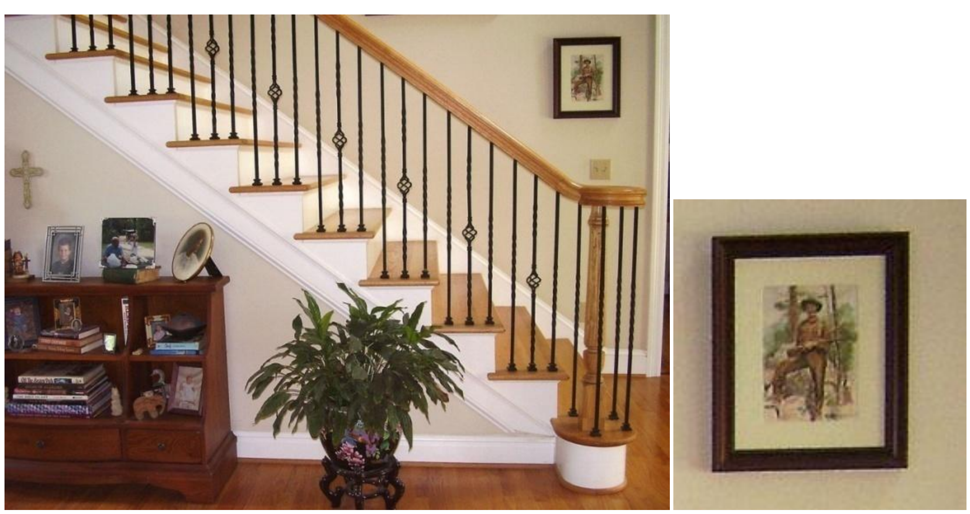
3.3 Discussion on Experimental Results

System has been tested with different types of images. For static images and video a4tech 5 mega pixel webcam has been used. Test images are 1200X900 pixels. Each child image has been tested in six different qualities to show key points matched between child and parent image.

3.4 Advantages of Proposed System

Many people are working on outdoor system for blind people, so they can easily move outside using GPS enabled smart phones. But in the indoor environment it loses access to the required satellite signals. They keep facing problem in indoor environment. This proposed system will help them in some way to face the problem for indoors.

3.5 Performance Evaluation



Parent ImageChild Image

Figure 16: parent and child image

Contrast image Matched image: Key Points 6

Figure 16.a: Contrast Image test

Blur image Matched child image key points 3

Figure 16.b: Blur Image

Occluded image Matched child image key points 5

Figure 16.c: Occluded Image

Dark Image **Matched child image key points 3**

Figure 16.d: Dark image

Noisy image **Matched child image key points 3**

Figure 16.e: Noisy image

Fog image **Matched child image key points 6**

Figure 16.f: Fog image

4.Conclusion and Future works

This chapter represents the conclusions of this work. Further work is also analyzed suggesting various ways through which this present research may continue.

4.1 Conclusion

This thesis has described an efficient and affordable method for blind assistive and navigation system. It has also presented a detailed survey covering a wide variety of blind assistive system and their corresponding methods. In addition, it merged knowledge and proposed some new concepts for blind assistive system with verbal cues in a real time environment. At the end of every chapter an extended summary is included and a discussion of the experimental results in previous chapter showed the details. In conclusion, it is noted that this proposed system is able to detect object in indoor environment where objects of one particular environment is given in database before. SIFT algorithm is used to detect key points of that particular object and object has been recognized using SIFT matching algorithm. Thus blind people can recognize even if any object is missing from the particular environment.

Finally, experimental results have demonstrated this proposed method is accurate and reliable to solve the promising and challenging real-time blind assistive and navigation system.

4.2 Future Works

In our pre-thesis we have worked on two algorithm SIFT (Scale Invariant Feature Transform) and SIFT matching algorithm. In our thesis we have joined these two algorithms and have given a shape to our proposed system which in able to detect objects in indoor environment and give signal by playing sounds as we discuss briefly. As our main target is to assist blind people so we want to enrich our system in future by making it in android compatible and also we wish to work on distance measurement of objects.

REFERENCES

1. Lowe, D.G. 1999. Object recognition from local scale-invariant features. In International Conference on Computer Vision, Corfu, Greece, pp. 1150-1157.

2. Sylvie Treuillet, Eric Royer, Thierry Chateau, Michel Dhome, Jean-Marc and Lavest, “Body Mounted Vision System For Visually Impaired Outdoor And Indoor Way Finding Assistance”, Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Assistive Technology for All Ages,2007,1-3.

3. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., and Wu, A.Y. 1998. An optimal algorithm for approximate nearest neighbor searching. Journal of the ACM,45:891-923.

4. João José, Miguel Farrajota, João M.F. Rodrigues, J.M. Hans du Buf, “The Smart Vision Local Navigation Aid for Blind and Visually ImpairedPersons”, Digital Content Technology and its Applications Vol.5 No.5, May 2011.

5. Boris Schauerte, Manel Martinez, Angela Constantinescu, and Rainer Stiefelhagen “An Assistive Vision System for the Blind that Helps Find Lost Things”, Study Center for the Visually Impaired Students, Engesserstr.

6. Beis, Jeff, and David G. Lowe, “Shape indexing usingapproximate nearest-neighbour search in high-dimensional spaces” Conference on Computer Vision and PatternRecognition,Puerto Rico (1997), pp. 1000–1006.

7. Lindeberg, Tony, “Detecting salient blob-like image structuresand their scales with a scale-space primal sketch:a method for focus-of-attention,” International Journal ofComputer Vision, 11**,** 3 (1993), pp. 283–318.

8.Schiele, Bernt, and James L. Crowley, “Object recognitionusing multidimensional receptive field histograms,” FourthEuropean Conference on Computer Vision, Cambridge, UK(1996), pp. 610–619.

9.http://www.cvip.louisville.edu/wwwcvip/research/publications/Pub_Pdf/2006_2/CSIF

TA%20SIFT%20Descriptor%20with%20Color%20Invariant%20Characteristics.pdf

10. http://en.wikipedia.org/wiki/K-d_tree

11.http://homes.cs.washington.edu/~lfb/paper/iser12.pdf

12.http://uhdspace.uhasselt.be/dspace/bitstream/1942/8035/1/3287.pdf

13.http://stackoverflow.com/questions/1627305/nearest-neighbor-k-d-tree-wikipedia-proof

14.http://en.wikipedia.org/wiki/Scale-invariant_feature_transform#Object_recognition_using_SIFT_features

15.http://stackoverflow.com/questions/13305416/getting-stuck-on-matlabs-subplot-mechanism-for-matching-images-points-for-vlfe

16. http://www.itl.nist.gov/iad/vug/sharp/Publications/NISTIR7625.pdf

17. http://www.mathworks.com/help/matlab/ref/uigetfile.html

18. http://www.mathworks.com/matlabcentral/newsreader/view_thread/306258

19.http://www.hcblind.org/?gclid=CKn1jMa_37ICFQV76wod2XQAYA

20.http://en.wikipedia.org/wiki/Scale-invariant_feature_transform

21.http://en.wikipedia.org/wiki/Scale-invariant_feature_transform#Scale-space_extrema_detection