A step towards development of aid for visually challenged

PHALGUNI^{id}

Department of Electronics and Communication Engineering, National Institute of Technology Karnataka,
Surathkal 575025, India
e-mail: phalguni22@gmail.com

MS received 14 March 2015; accepted 5 October 2016

Abstract. The work done towards developing a visual aid to help visually challenged people is described in this paper. An ultrasonic device is used for measuring the distance to a nearby object and SIFT algorithm based approach is used for object recognition from the captured image. Features of the recognized object as well as the distance to the object are converted to voice output which is available to the visually challenged person for navigation. The system developed and results obtained are described and suggestions are given for further work to be done to develop a better and more compact visual aid.

Keywords. SIFT; blind navigation; object recognition.

1. Introduction

Visually handicapped people who are not born blind and who have lost eyesight due to some deficiencies at a young age prefer not to take help for moving around either in the house or in the town. They know the world by and large and would like to behave as normal people as they can operate a computer, mobile phone, etc. At present there are a large number of tools available that help them overcome many problems. They can operate a computer or a mobile phone with the help of special software with voice feedback options to guide them. There are many types of sticks available for them for blind navigation inside and outside their home, as well as on the roads.

Our aim was to develop an electronic stick with distance sensors and image sensors to give a voice feedback about the scene in front of the blind and help them navigate roads or buy medicine at the medical store with little help. Such a system should do scene analysis and describe the scene as best as possible like another person would do. This is a very complex problem and there are not many robust solutions available.

2. Literature review

Many systems are being developed in today's world that specially cater to the needs of the blind. The electronic stick for blind would help the person to navigate freely without the help of another individual. The most important aspect in our system was to identify the obstacles ahead and give an audio description of the object and its distance from the person. Many approaches have been used till date for

detection purposes like corner detection, which are sensitive to scale changes. An extensive literature survey indicated many drawbacks and problems in the current systems. The detection process needs to be invariant to scale and rotation changes as the frames could be captured at any distance from the target and at an angle. Cluttering and partial occlusion should not affect the matching procedure and this is possible by the use of key-point descriptors. The system must process the visual information similar to the human brain and should not be affected by any of the factors that do not affect human vision. Segmentation-based methods are not suitable in our context as the objects to be identified cannot be uniquely distinguished from the rest of the objects or background in terms of size, shape or colour. Neural networks could be used as another possible approach to the problem, which would require training and extra computation time.

Towards the end we have developed a prototype based on ultrasound distance sensor and image recognition using a webcam and scale invariant feature transform (SIFT) technique proposed by David Lowe [1].

SIFT can be used to detect all objects of different size, shape and colour, including human faces without training by the use of key-point descriptors. Using the key-point descriptors, we can find matches between two images with corresponding objects.

The SIFT method uses the concept of finding features and key-point descriptors, which is similar to the human vision and thinking process. It is invariant to scale and rotation changes, and also partially invariant to illumination changes. The scale invariant features are identified by filtering process in multiple stages. The SIFT keys from an image are used in nearest neighbourhood approach to

identify object models. The ultrasonic sensors give the distance information, which is processed along with the object matching information to give a voice output describing the navigation scenario.

3. Image recognition methods

3.1 Correlation co-efficient method of identifying objects

The real time captured frame was correlated with the database images. The one with the highest correlation co-efficient value was taken as the best match and the object was assumed to be present in the frame [2]. This method is insensitive towards illumination changes. Rotation and scaling do affect the detection process. This method is not suitable for detecting multiple objects in the frame as the correlation values may be high for non targets too and a common threshold for distinction cannot be used. The method also fails when none of the database objects are present in the scene as the image that gives the highest correlation value is assumed to be present by default. The correlation coefficient is calculated as

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \quad (1)$$

where r is the correlation co-efficient, X, Y are location co-ordinates in the image, $\bar{X}, \bar{Y}$ are mean in the x and y direction, respectively, and n is the number of pixels in one dimension. Correlation co-efficient for 2-D greyscale image pairs also give satisfactory performance when the target is at almost a fixed distance and orientation with respect to the camera.

This method is based only on the pixel values of the frame. Background plays an important role in matching, thus giving more false matches. A ball and a bag on the same white background wall may be matched instead of a ball on a dark coloured table. The colour of the object is indirectly considered when it is converted to gray scale format. Blue and yellow have different pixel values.

3.2 SIFT method

The SIFT [3–8] method is used to obtain the key points, feature vectors and their orientations from an image. The image is now described by its feature vectors rather than pixel values. This method when used in object detection is highly invariant to rotation and scaling, and partially invariant to changes in illumination.

The captured image is progressively blurred using a Gaussian function and re-sized. The object of interest may be of any size in the frame as the size depends on the distance of the object from the camera. The image at different sizes and blur levels help in identifying key-points at

different scales with respect to actual size. Corners are important in identifying the features in an image; hence extrema of the difference of frames are calculated at different blur levels. Edges are responsible for false positives and should not be considered in further processing.

Orientations are assigned to key points based on gradients. Normalization at this stage helps reduce the effects of change in illumination, even though the dependence cannot be removed completely. Independence with respect to changes in rotation is achieved by subtracting feature vector orientation from key point orientation.

The matching of pairs of images is done by calculating the dot products of two feature vectors at a time. False matches are reduced by considering ratio of first closest match to second closest match. A valid match should have high ratio as it matches closely with only one of the vectors in the entire set. False match vectors have small difference between the magnitudes of first and second best match vectors.

The entire matching procedure is based on gradients and orientations. Human vision performs matching in a similar way, but many more aspects are considered like colour, shape, probability of occurrence in a given environment and so on. The brain also relies on previous experiences to exclude instances of occurrence of an object. Humans can differentiate between a blue and red ball of similar size and shape, whereas the approach we use tends to identify them to be instances of a single object.

The matching algorithm can be viewed in a simplified manner. Identification of an object, say a bag, is done in steps. The gradients and corners give a sense of shape. Subtraction of feature vector orientations from key-point orientation is similar to keeping the bag in an upright position. Resizing the images helps us to match the bag with different templates of the same bag at different sizes.

Size, shape and position information, on the whole, help us identify the object. The database should be constructed with great care. The database image might look normal to the human eye, but considering matching at pixel level, there might be incorrect results due to a faulty image. The database images can be tested by matching a given image with various objects. A good recognition rate with the object of interest and a poor rate with other objects is the criteria for retaining the image in the database. The results at the end indicate that matching is prominent at the corners and conspicuous edges. Noise-free histogram-equalized images should be used for accurate matching.

3.3 SURF method

SURF (speeded up robust features) gives almost similar performance as that of SIFT and is faster too. The SURF [9] method uses stacks of blurred images unlike SIFT, which forms pyramids by sub-sampling. SURF is based on sums of approximated 2-D Haar wavelet. The SURF method was

slightly more variant to rotation in our system, and hence the SIFT method was preferred.

4. Prototype of system developed

The ultrasonic sensors used give information about the distance up to which the person can navigate without facing obstacles. Two sensors have been used, one at the ground level and another at the camera level, to detect obstacles at two different heights and give a combined information about the navigation situation. This would help the person to identify obstacles like steps that are at different distances at different heights. Face detection is also possible using the key-point descriptors and some of the database images have a set of human faces to be identified. The image capturing device used is Logitech HD Webcam C310. The image captured is approximately at a fixed distance from the ground level, making the identification procedure simpler. RS-232 signals from the sensors are converted to compatible signals to be used with the USB port of modern laptops, which do not have serial ports. The sensors are mounted on a mobile trolley on which the laptop is placed. The maximum frame size for this webcam is 1280×720 pixels for video and up to 5 megapixels for photos.

The trolley can be moved in any direction and the webcam gives the image of the scenario from a single direction. The sensors give the distance of the obstacle in inches within the range of 6–255 in. Hence the distance of 255 in is considered as no obstacle irrespective of the object ahead. The person is instructed to move ahead. Matlab is the platform used in the system.

Another system has been developed on OpenCV platform that has faster matching, but it is not a standalone system that can interact with other software and platforms required to give voice outputs. Experimental set-up of the system developed and the flow chart for the system developed is shown in figures 1 and 2.

4.1 Sensor interfacing

LV MaxSonar EZ1 by MaxBotix was used for the ranging purposes. Sonar Sensors are used for distance measurement based on the principle of short to long range detection and ranging. It provides sonar range information from 6 inches to 255 inches with one inch resolution. The inbuilt PIC microcontroller continuously performs signal processing techniques to extract the distance. Figures 3 and 4 show the serial port interfacing.

Readings could be taken every 50 ms. As the system processing time was much larger than this, the sensor would be ready for measurement in the next cycle without a problem. The TX pin delivers a serial output of baud rate 9600 when calibrated appropriately. This serial range information could be accessed through the serial port read

Figure 1. Experimental set-up with the laptop, webcam and ultrasonic distance sensors on the trolley.

function in Matlab and we can get the distance of the nearest obstacle in inches within a range of 6–255 in. The distance in inches is followed by an R in ASCII format.

4.2 Text to speech conversion

Various descriptions about the surrounding objects were stored in the form of textual data. This was read out as an output of the system. Microsoft Win 32 Speech API was used for the purpose and this was integrated into the Matlab code. There are different voices in various accents available like male, female, child, etc. Another alternative used was storing the description in audio files and playing them one after the other according to the navigation situation.

5. Results

The correlation co-efficient method, though simple, requires online processing of all database images and pairwise matching with the captured frame. The length of the message output that describes the surroundings determines the time required to give the voice output; hence it is variable. For a database of about 30–35 images, the correlation method took 5 s on a Pentium-4 processor (1.6 GHz) to find a match. The efficiency was 20% on an average. This could go up to 50% when the camera was at almost fixed distance and orientation from the target.

Figure 2. Flow chart of the system.

In the SIFT method, the database image descriptors are stored in text files and are loaded at the time of execution to reduce the time required to compute the key points for each of the images. The decrease in computation time accompanies decrease in memory requirements. The SIFT method took about 30 s for image acquisition and matched with an efficiency of above 80%. On an average, about 200 descriptors are extracted for a 480×640 image. Higher resolution images would result in higher number of descriptors and increased accuracy. SIFT is not completely invariant to illumination; hence care must be taken to maintain almost uniform illumination conditions during image capturing process. This would not be a problem in closed spaces with artificial lighting throughout the day. The matching efficiency is over 95% for pairs of images

Figure 3. Sensor interfacing with the serial port showing the power supply and receiver transmitter connections.

Figure 4. The sensor connected to the serial port.

Figure 5. Face detection.

captured on the same day and time due to similar illumination conditions. The colour information in the image is not used in finding the feature vectors unlike the human vision system. The results stated are for single-object detection.

Figure 6. Matching images with a single object.

Figure 8. Matching a single object with a set of different objects.

Figure 7. Multiple object recognition objects and false match in one of them.

The correlation method fails when the sizes of objects are different because the pixel values in the two images will be entirely different. The background also plays a major role in calculation of correlation coefficient. In SIFT-based methods, the background will give less number of key-points when there are no distinctive features present, thus reducing false matches. The SIFT method finds distinct features like corners and orientations which remain the same even with variations in size and rotation. Hence the SIFT method was found to be robust compared with correlation-coefficient-based matching with minimum restrictions on distance and orientation of the camera from the target.

An example for the text that would be converted to voice output is, “The object ahead is a bag and is at a distance of 234 in”.

The overall match rate is slightly higher in our case compared with observations in [1] for SIFT method as the

Figure 9. False matches due to similar feature vectors.

rotation of objects is restricted in our experiment due to the fixed distance and orientation of camera on the mobile trolley. On an electronic stick, the match rate may decrease. Figures 5, 6, 7, 8 and 9 show the pairs of images with lines joining the matched key-points. Figure 9 shows false matches due to similar feature vectors on the objects which lead to incorrect results.

6. Conclusion

The current system is bulky in terms of size and weight. The sensor used provides distance information in the range 6–255 in. This is sufficient in our case. Various other sensors could be used for higher ranges.

The Matlab code could be converted into C code using the latest version of Matlab and then ported into a digital signal processor. The final system could be an electronic stick with compact cameras for hand-held application devices equipped with audio output ports. The system could

be made more robust by including multiple instances of single object in the database, at different angles and distance from the camera, in order to completely avoid sensitivity to rotation and scaling changes. The processing time could be reduced by using improved and optimized SIFT-based algorithms. The electronic stick could then be moved to any height and not be confined to a fixed height from the ground. Combination of SIFT and correlation techniques can also be used to make a decision in case of equal number of match vectors between pairs of images. The use of graphical processing units would make the system much faster due to parallel processing. Electronic sticks would also require larger area for navigation and may not be preferred by some users. The same system can be ported into waist belts without modifications which would be more user friendly compared to electronic sticks.

Multiple object detection is possible when the numbers of key-point matches are high enough and consistent on the presence of database object in the captured frame. Absolute thresholds can be employed to determine presence or absence of an object in a set. This could also make the system more robust and informative. Local texture features that assist human vision systems could be incorporated into the descriptor identification procedure to increase the accuracy of matching. 3D models could be built by capturing the object in an image from different angles followed by the use of algorithms that are based on binocular human vision, making the identification process easier. The information thus obtained could be used to construct a 3-D map of the neighbourhood, and periodic updating of the information collected would help in easier navigation.

Acknowledgements

This research was supported by the International Institute of Information Technology Bangalore, under the Summer

Internship programme 2011. I wish to thank Prof P G Poonacha for his guidance during the internship.

References

- [1] Lowe David G 2004 Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60(2): 91–110
- [2] Szeliski R and Kang S B 1994 Recovering 3D shape and motion from image streams using nonlinear least squares. *J. Vis. Commun. Image Represent.* 5(1): 1028
- [3] Lowe D G 1991 Fitting parameterized three-dimensional models to images. *IEEE Trans. Pattern Anal. Mach. Intell.* 13(5): 441–450
- [4] Lowe D G 1999 Object recognition from local scale-invariant features. In: *International Conference on Computer Vision*, Corfu, Greece, pp. 1150–1157
- [5] Lowe D G 2001 Local feature view clustering for 3D object recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, pp. 682–688
- [6] Mikolajczyk Krystian and Schmid Cordelia 2001 Indexing based on scale invariant interest points. In: *International Conference on Computer Vision*, Vancouver, Canada, pp. 525–531
- [7] Lowe David G 1987 Three-dimensional object recognition from single two-dimensional images. *Artif. Intell.* 31(3): 355–395
- [8] Johannes Bauer, Niko Sunderhauf and Peter Protzel, Comparing several implementations of two recently published feature detectors. In: *Proceedings of the International Conference on Intelligent and Autonomous Systems*, IAV, Toulouse, France, p 122
- [9] Luo Juan and Oubong Gwun, A comparison of SIFT, PCA-SIFT and SURF. *Int. J. Image Process.* 3(4): 143–152