Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 58 (2015) 387 – 394

Procedia Computer Science

Second International Symposium on Computer Vision and the Internet(VisionNet'15)

Vision System with Audio Feedback to Assist Visually Impaired to Grasp Objects

Gautam S[a], K.S.Sivaraman[a], Hariharan Muralidharan[a], A. Baskar[b*]

[a] *Student, Dept. of Computer Science and Engineering, Amrita Vsishwa Vidyapeetham (University), Coimbatore, India*
[b]*Assistant Professor, Dept. of Computer Science and Engineering, Amrita Vsishwa Vidyapeetham (University), Coimbatore, India*
**E-Mail:* a_baskar@cb.amrita.edu

Abstract

In recent times, there have been many products that cater to the visually challenged, but only few have been readily available for regular use. The reason being, that the technology driving these products are costly, or complex to use. In this work, we propose a "Vision system with audio feedback to assist visually impaired to grasp objects". Our proposed system eliminates these challenges, in its usability, complexity and functionality. The system is designed to serve the following: (1) Finding a desired object in the scene, in which, the object recognition is done using Weighted Matrix Algorithm from the visual input received from the camera; and (2) Assisting the user to the object's proximity, where, the user is guided using audio-feedback, at every step. This approach is represented taking the example of a commonly found object in our household. We have considered a water bottle. The above approach is found to produce suitable results.

Keywords: Object Recognition, Weighted Matrix, Visually Impaired, Audio Output, Hu Moments;

1. Introduction

The number of visually challenged people is already significant, and is ever-increasing. Owing to their disability, interaction with common household objects is restricted. Technology has assuaged the lifestyle of mankind, whether physically challenged, or visually challenged. Three widely used sensors or methods used to aid the visually impaired are: Audio [10], Visual [3,4,6] and GPS [9]. These have been exploited to the fullest use to ease the condition of the visually impaired. Unfortunately, these endeavours have not been pervasive enough for one or more reasons.

Audio signal has been widely used to facilitate the visually impaired. The following work, Text detection from

doi:10.1016/j.procs.2015.08.041

natural scene images, uses character extraction to provide a reading model for the visually impaired using text-to-speech [10]. The drawback with text-to-speech is that input in the form of text cannot be used by a visually impaired to interact with objects.

GPS sensors alone are seldom used for indoor navigation. They are generally used along with RFID [9] sensors or tags. There have been many efforts and works on indoor navigation [2]. Obstacle avoidance has also been a well-touched area of research. But as stated before, apart from the navigation, further interaction with the object of interest is restricted. However, the usage of these sensors has not been very common, and do not seem feasible for day-to-day activities.

Only a camera as a visual sensor, can best match the functionalities of a human eye. Furthermore, an audio feedback is the most convenient way to assist a visually impaired user. This system makes use of the visual input scene to generate a vocal output. Not many models have combined visual input [5, 7, 8] and Audio output to assist the visually challenged. A vision system with Audio feedback to assist a visually impaired has been implemented in this paper. Using this system, the user can be brought to the vicinity of the object with ease.

The paper has been divided into 4 sections. Section 2 deals with the elaborate implementation of the system. The result and analysis are discussed in detail in section 3. The final section draws a conclusion to this paper.

2. Proposed Methodology

The proposed system combines two modalities, the vision system and the audio system, to assist the visually challenged. Hu Moments are extracted from the images and used as feature vectors. Hu Moments [11] otherwise called as moment invariants are derived from six absolute orthogonal invariants and one skew orthogonal invariant based upon algebraic invariants, which are not only independent of position, size and orientation but also independent of parallel projection. The moment invariants have been proved to be the adequate measures for tracing image patterns regarding the images translation, scaling and rotation under the assumption of images with continuous functions and noise-free. The proposed methodology can be modularized into three steps, namely: 1) Assist the user to view the region of interest (frame containing the target object), 2) Calibrate the object of interest to the centre of the frame, and 3) Instruct the user to reach and grasp the object. The three steps have been delineated as follows.

2.1 Assist the user to view the region of interest

In this step, the system assists the user in identifying the correct Region of Interest (ROI). The ROI is defined as the frame containing the desired object. The ROI is trained manually prior to this step. To train the ROI, we take 100-150 images of the object along with its surroundings. This takes not more than 5 minutes running a python code. From the images, the Hu Moments are extracted. We use Hu Moments as a feature vector (H_i) from which seven invariant features are obtained as follows:

$$H_i = < h_1, h_2, h_3, h_4, h_5, h_6, h_7>$$
$$i \in 1 \text{ to } n$$
$$n \rightarrow \text{total number of training images}$$

The feature vector obtained from the trained images $[H_1,H_2,H_3…H_N]$ has been given to the Weighted Matrix (WM) as proposed in our previous paper [1] and the min (m) and max (M) matrices are modelled. The Weighted Matrix algorithm recognizes the object in two phases, the feature Weighted Matrix Model will be constructed using feature vector and is estimated by Hu Moments in the first phase and in the second phase target objects are identified using Diagonal Rank Matrix.

During the testing, a frame is captured. Hu Moments are extracted from this frame as input vector, T.

The matrices 'm' and 'M' and 'T' are all of the order 1 x 7. Deviation from the maximum(R) and deviation from the minimum(r) is calculated as follows:

$$R = M – T. \text{ Hence, R has an order of } 1 \times 7$$
$$r = T – m. \text{ Hence, r has an order of } 1\times 7$$

$$S = R.r^t$$

The resultant matrix *S* has an order of 1 x 1, i.e., a single element, which is the weighted sum. The sign of the resultant weighted sum decides whether the input frame matches with the desired frame, or not. The audio feedback is summoned at every step. If the sum is positive, then the audio says "Object found in the frame", and proceeds to the next step, else, it says, "Scanning", implying that the exact region of interest is not in view. Figure 1 illustrates the following.

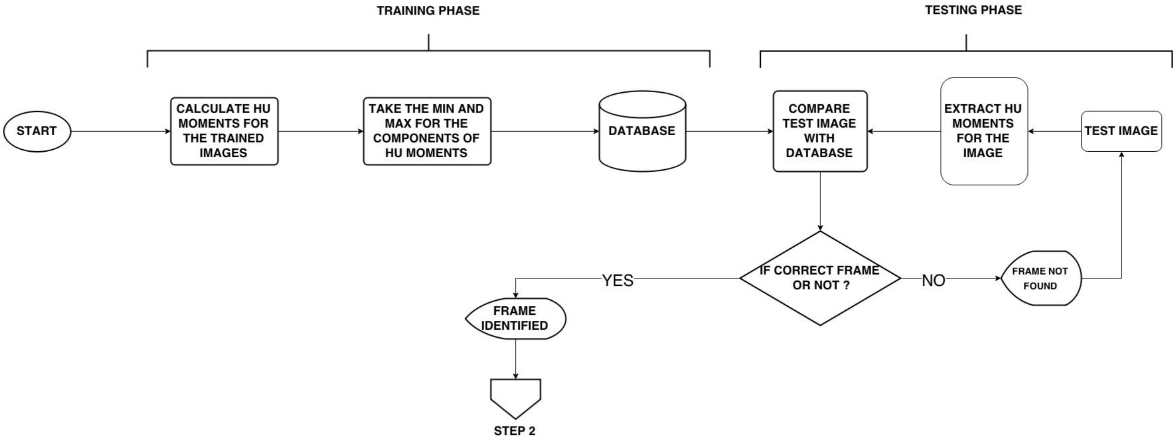

TRAINING PHASE
TESTING PHASE
START
CALCULATE HU MOMENTS FOR THE TRAINED IMAGES
TAKE THE MIN AND MAX FOR THE COMPONENTS OF HU MOMENTS
DATABASE
COMPARE TEST IMAGE WITH DATABASE
EXTRACT HU MOMENTS FOR THE IMAGE
TEST IMAGE
IF CORRECT FRAME OR NOT ?
YES
NO
FRAME NOT FOUND
FRAME IDENTIFIED
STEP 2

Figure 1 To assist the user to view the ROI

A similar processing approach is used for the forthcoming steps. The training image set is processed to obtain the min and max matrices, followed by taking the input frame during testing and applying the WM algorithm to determine the correctness.

2.2 Calibrate the object of interest to the centre of the frame

Figure 2 shows the overview of step 2 in the form of a flowchart. We have now identified the frame which contains the object. The object need not be exactly in the centre of the frame. If the object is in the centre of the frame, it will not be necessary to keep track of the desired object in the forthcoming steps, thereby saving a lot of processing. Guiding the user to the object will be easier if it is kept in the centre of the frame of view. Hence, for aesthetic and efficient reasons, it is important to calibrate the object to the centre of the frame.

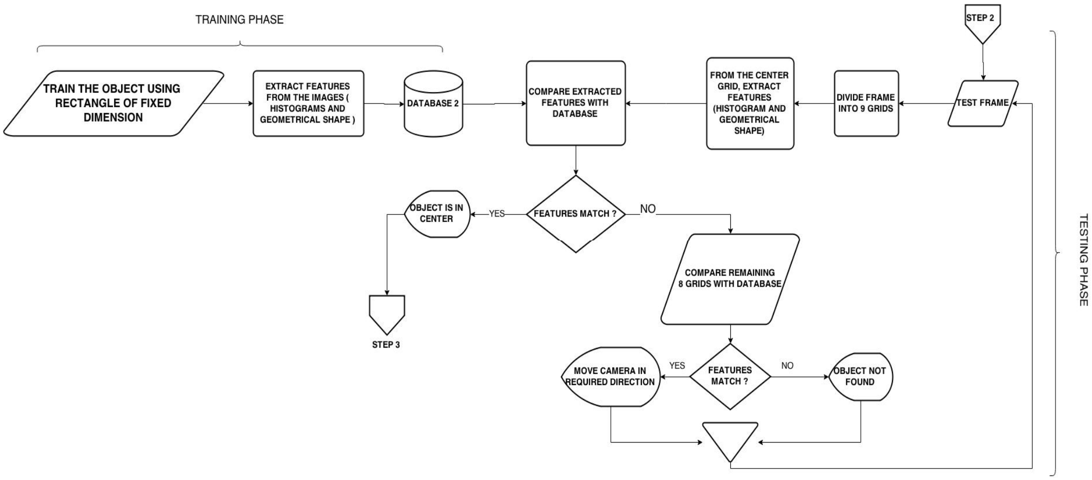

TRAINING PHASE
TRAIN THE OBJECT USING RECTANGLE OF FIXED DIMENSION
EXTRACT FEATURES FROM THE IMAGES (HISTOGRAMS AND GEOMETRICAL SHAPE)
DATABASE 2
COMPARE EXTRACTED FEATURES WITH DATABASE
FROM THE CENTER GRID, EXTRACT FEATURES (HISTOGRAM AND GEOMETRICAL SHAPE)
DIVIDE FRAME INTO 9 GRIDS
TEST FRAME
STEP 2
TESTING PHASE
FEATURES MATCH ?
YES
NO
OBJECT IS IN CENTER
STEP 3
COMPARE REMAINING 8 GRIDS WITH DATABASE
FEATURES MATCH ?
YES
NO
MOVE CAMERA IN REQUIRED DIRECTION
OBJECT NOT FOUND

Figure 2 To calibrate the object to the centre of the frame

The first grid scanned for this step is the centre grid. The features are extracted and processed using the WM algorithm. If the obtained sum is positive, signifying that the grid contains the object, the system alerts the user that the object is calibrated to the centre. In case the first grid does not contain the object, (as per the result of the WM algorithm, if the sum is negative), the other grids are scanned.

The following paragraph explains how the system gives the audio output.

Let *h* and *w* be the height and width of the input frame. Let *G* be the grid with the maximum weight, and *x* and *y* be the centroids of *G.* Let *i* and *j* be the direction of the grid. *i* and *j* are computed as follows:

$$\mathrm{i} = floor\left(\frac{x}{\frac{h}{3}}\right), \qquad j = floor\left(\frac{y}{\frac{w}{3}}\right)$$

$$\text{if i == 0: Top, if i == 1: Centre, if i == 2: Bottom}$$

$$\text{if j == 0: Left, if j == 1: Centre, if j == 2: Right}$$

A pseudo-code is given for the same:

```
1) Stored array: ver = [‘Top’, null, ‘Bottom’]
2) Stored array: hor = [‘Left’, null, ‘Right’]
3) $\mathrm{i} = \left(\frac{x}{\frac{h}{3}}\right),\ j = \left(\frac{y}{\frac{w}{3}}\right)$
4) Print “Look”, ver[i], hor[j].
```

The frame is divided into 9 grids of equal areas, hence, the height and width of the frame, each needs to be divided by 3.The algorithm always starts in the centre of the grid so the case for *i*=1, and *j*=1 can be ruled out. Hence, the possible combinations of output that the system can generate are {“Top”, null, “Bottom”} X {“Left”, null, “Right”}, depending on the respective values of *i* and *j.* Once the user gets the audio feedback from the system, the user shall calibrate his position in such a way that the object moves to the centre of the frame and the system will keep sending feedback until the user reaches the required position. To illustrate further, consider this example where *i*= 0 and *j* = 2, in such a case, the audio output will be “Look Top Right”, hence the user will move in the said direction until the object is aligned to the centre. This is carried out similarly for all cases.

2.3 To lead the user toward the object by instructing to walk forward until a suitable position is reached.

The object has been identified and calibrated to the centre of the frame. But the user is still distant from the object. In this step, we guide the user to the object. As mentioned previously, calibrating the object to the centre has the advantage of guiding the user straight towards the object. To accomplish this, we again make use of Hu Moments as features. We train the top half of the object. This is done, because only the top half of the object can be effectively seen when the user approaches the object. Also, the object slowly creeps to the bottom part of the frame as the user walks forward. As a result, a static window has been maintained at the bottom part of the frame which scans if the top half of the object has reached the bottom part. From the window, the features are extracted, i.e., Hu Moments are computed for the window and processed using the WM algorithm. If the computed sum is positive, then the system sends an audio output saying the user has reached the object and then the system terminates. If the value is negative, the system iterates, and sends the audio message "Walk Forward".

3. Result and Analysis

3.1 Experimental Setup:

The experiments were performed in three steps 1) Check if the user is looking at the correct frame 2) To find the object and calibrate it to the centre of the frame 3)To lead the user to the object. In this section, we have used a bottle as the object of importance. The results have been explained in the following sections, including the screenshot of the running program, along with the magnified output of the text that is read-out, or displayed for the sake of the results.

3.1.1 Check if the user is looking at the correct frame:

This step verifies if the user is looking at the correct frame or not. Hu Moments are extracted from the test image and compared with the minimum and maximum values obtained from step 1 during the training phase. If the values lie between the minimum and maximum values, it can be safely assumed that the user is looking at the correct frame.

If the feature extracted from the test image lies between the minimum and maximum values of the database, the user is said to be looking at the correct frame. If the values from the test image do not lie between the minimum and maximum values of the Hu Moments obtained from step 1 of the training phase, the user is said to be looking elsewhere. The following results illustrated in Fig 3 and 4 show the same.

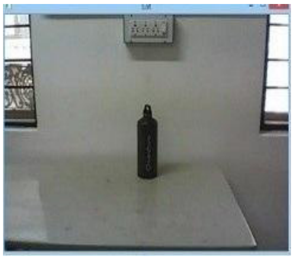
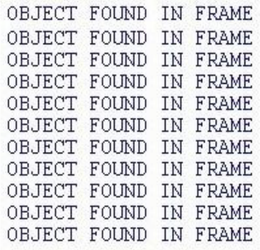


Fig 3(a) Object of interest found



```
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
OBJECT FOUND IN FRAME
```

Fig 3(b) audio system

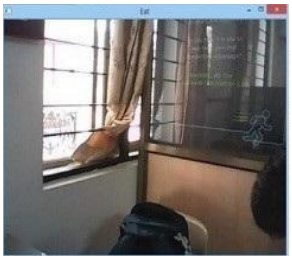
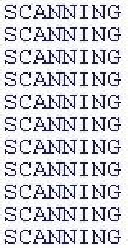


Fig 4(a) Object of interest not found

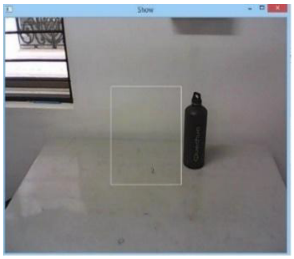


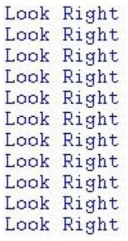
```
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
SCANNING
```

Fig 4(b) audio system

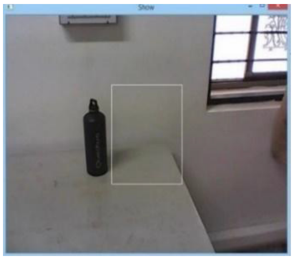
3.1.2 To find the object and calibrate it to the centre of the frame:

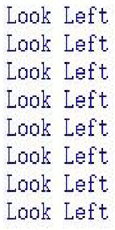
The second phase is to obtain position of the object from the frame identified from the previous step. Object is trained using a bounded rectangle, and Hu Moments are extracted from it. In the testing phase, values from the training phase are compared with the test image. In this paper, results for the object located in the left and right side of the frame is shown. If the object lies to the right or left of the bounded rectangle, we get results as shown in Fig 5 and Fig 6 respectively.





```
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
Look Right
```





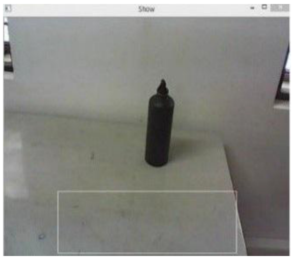
```
Look Left
Look Left
Look Left
Look Left
Look Left
Look Left
Look Left
Look Left
```

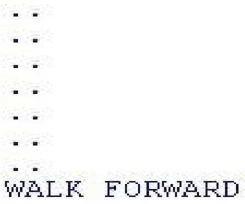
Fig 5(a) Bottle found to the right side of the box Fig 5(b) audio system Fig 6(a) Camera moves from left side to right side. Fig 6(b) audio system

Fig 5 shows the object found to the right side of the frame. If the user moves to the right, the object will be calibrated to the centre of the frame. We get results as shown in Fig 6, for the object situated to the left part of the frame. As can be seen, the screenshots demonstrate the experiment conducted over live video feed and getting the output instantly, adjacent to each screenshot.

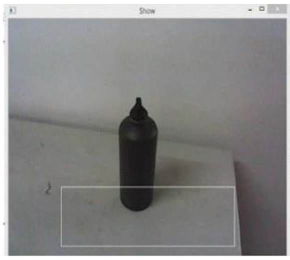
3.1.3 To lead the user in front of the object

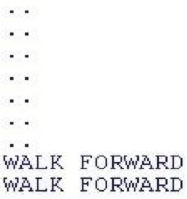
If the top half of the object is not found inside the bounded box, the program instructs the user to walk forward. The user keeps walking forward until the object is obtained, at which point the system alerts that the user has reached the target and terminates the program. Fig 7 to Fig 9 explains step 3. The following pictures contains screenshots from the output of the code. The output text has been magnified for the sake of clarity.





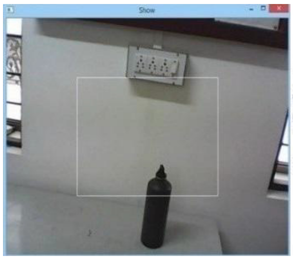
```
. .
. .
. .
. .
. .
. .
. .
WALK FORWARD
```

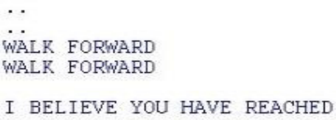




```
. .
. .
. .
. .
. .
. .
. .
WALK FORWARD
WALK FORWARD
```

Fig 7(a) Instruction to walk forward Fig 7(b) Audio System Fig 8(a) User walking forward Fig 8(b) Audio output





```
. .
. .
WALK FORWARD
WALK FORWARD

I BELIEVE YOU HAVE REACHED
```

Fig 9(a) Object has been reached Fig 9(b) Audio System

Fig 7 and Fig 8 illustrate the final step, as the user walks toward the object, as instructed by the system. If the object is identified towards the bottom of the frame as trained, the above loop terminates the program as can be seen in Fig 9.

As the user walks forward, the top half of the object approaches the bottom half of the screen as can be seen in the examples above.

The same experiment was performed on various other objects, which are explained in Table 1. Each of the object was experimented upon, **five** times. The following table shows the results for all the three steps individually for the five trials:

Table i Experimental results for tested objects

OBJECT	SUCCESS RATE – STEP 1	SUCCESS RATE – STEP 2	SUCCESS RATE – STEP 3
Bottle (screenshots)	4 / 5	5 / 5	5 / 5
Coffee-Maker	5 / 5	5 / 5	4 / 5
Pink Mug	3 / 5	2 / 5	4 / 5
School Bag	5 / 5	4 / 5	5 / 5

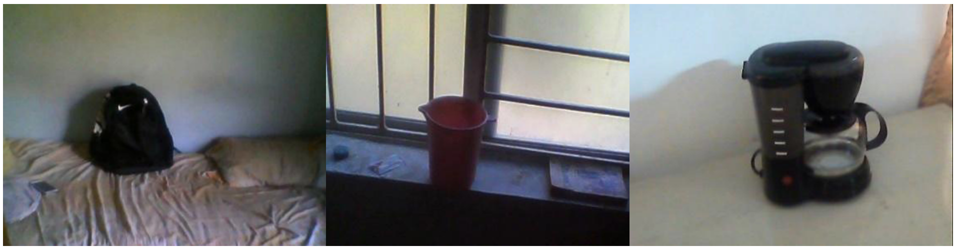


Figure 10 Test objects: Bag, Mug and Coffee Maker

4. Conclusion

In this work we have implemented a vision system with audio feedback to assist the visually impaired to grasp objects From the observations made during the course of this experiment, it was observed that the training data set needs to be crucial for obtaining positive results while testing. The features used to determine also play a necessary role to eliminate dependencies. As a whole, features like Geometric modelling combined with Histogram bin and Hu Moments produce results of a high accuracy under this scenario. The proposed method is implemented in three steps 1) Assist the user to view the region of interest (frame containing the target object), 2) Calibrate the object of interest to the centre of the frame, and 3) Instruct the user to reach and grasp the object. The result proves that the two modules vision and audio combined together will improve object recognition at the same time provide a good platform for assisting the visually challenged, in an effective and efficient manner. The system can be improved when there are many similar looking objects in the same frame; or the lighting/illumination conditions change adversely. It is preferred that the background should be in contrast with the object of interest. Nevertheless, the advantages of this system far outweigh its drawbacks.

References

[1] K.S.Sivaraman, S. Gautam, S. Sarvesh, Archit Khullar, A. Baskar and Shriram K. Vasudevan. *Object Recognition using Weighted Matrix – A novel approach.* Indian Journal of Science and Technology, vol. 8(S7), pp. 278-291, April 2015.

[2] Lisa Ran, Sumi Helal and Steve Moore. *Drishti: An Integrated Indoor/Outdoor Blind Navigation System and Service.* I Pervasive Computing and Communication, 2004. PerCom 2004. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 14-17 March 2004.

[3] W. T. Lee and H. T. Chen. *Histogram-based Interest Point Detectors.* In Proceedings of the IEEE Conference on Vision and Pattern Recognition, 2009, pp. 1590-1596.

[4] H. Zhang, W. Gao, X. Chen, and D. Zhao. *Object Detection using Spatial Histogram Features.* Image Vision Computing, vol. 24, pp. 327-341. 2006.

[5] C. F. Olson. *A General Method for Geometric Feature Matching and Model Extraction.* International Journal of Computer Vision, vol. 45, pp. 39-54, 2001.

[6] Joseph L. Mundy. *Object Recognition in the Geometric Era: A retrospective.* Toward category-level object recognition. Springer Berlin Heidelberg, pp 3-28,
2006.

[7] O. Choi and I. S. Kweon. *Robust Feature Point Matching by Preserving Local Geometry Consistency.* Computer Vision and Image Understanding, vol. 113,
pp. 726-742, 2009.

[8] Dilip K. Prasad. *Survey of the problem of Object Detection in real images.* International Journal of Image Processing (IJIP) 6.6 (2012): 441.

[9] K. Yelamarthi, D. Haas, D. Nielsen and S. Mothersell. *RFID and GPS Integrated Navigation System for Visually Impaired.* 53rd IEEE International Midwest Symposium on Circuit and Systems (MWSCAS), pp. 1149-1152, August 2010.

[10] N. Ezaki, M. Bulacu, L. Schomaker. *Text Detection from Natural Scene Images: Towards a System for Visually Impaired Person.* 17th International Conference on Pattern Recognition (ICPR 2004). Vol. 2, pp. 683-686, August 2004.

[11] Z.Huang, J.Leng. *Analysis of Hu's Moment Invariants on Image Scaling and Rotation.* Proceedings of 2010 2nd International Conference on Computer Engineering and Technology (ICCET). (pp. 476-480). . Chengdu, China. IEEE.