# Indian Institute of Information Technology Allahabad

## A Project Report

## On

## "Multi Object Recognition in an Indoor Environment"

***Submitted By:***

**Abhineet Kumar Singh**
IIT2009148

*Under the Guidance of:*

**Prof. U.S. Tiwary**
Professor
*IIIT-Allahabad*

**December, 2012**

# CANDIDATE'S DECLARATION

I hereby declare that the work presented in this project report entitled "Multi Object Recognition in an Indoor Environment", submitted as the end semester report of 7th Semester of B.Tech. (IT) course at Indian Institute of Information Technology, Allahabad, is an authenticated record of my original work carried out from July 2012 to November 2012 under the guidance of **Prof. U.S. Tiwary**. Due acknowledgements have been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad
Date: 3/12/2012

Name: Abhineet Kumar Singh
Enroll: IIT2009148

# CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Place: Allahabad

**Prof. U.S. Tiwary**

**Professor**

**IIIT Allahabad**

# *Abstract*

The task of recognizing individual objects in a complex real life image containing several object classes is one of the most challenging tasks in the field of computer vision. In spite of several decades of intensive research, modern state of the art object recognition systems are easily outperformed by the detection capabilities of primates, especially human, both in terms of processing time and classification accuracy. This fact has inspired many recent works to try to emulate the way information is processed in the primate visual system on the basis of whatever little is known about this subject. The current work is another attempt in this direction that is based on a hierarchical model of processing in the primate visual cortex. There are two characteristics that are crucial for an accurate and robust object recognition system: invariance to object transformations and selectivity towards specific object features. The objective of this work is to implement a system that can learn both simultaneously, in an incremental manner, across several layers. These layers are organized in a hierarchical manner such that each layer is more robust and sophisticated than the one below it. The performance of this system has been tested on scenes of common indoor environments that are of particular significance for both domestic and industrial robots as well as indoor surveillance systems.

# *Contents*

**List of figures and tables used in this report**

## 1. Introduction

Object recognition in cluttered environments is one of the most challenging problems in the field of computer vision. Most systems that have been developed so far are far out performed by the primate visual system both in terms of accuracy and speed. This is especially true in cases of three dimensional objects with recognition required from arbitrary views of the objects. Object transformations like rotation, translation and scaling together with differences in lighting and illumination levels have likewise proved to be tough to handle for most contemporary systems. On the other hand, the visual cortex of primates, particularly humans, appears to be handling these issues with amazing ease. This is why the most promising approaches to solving these problems appear to be those that attempt to emulate the functioning of primate visual cortex using whatever little is known about the bio-physical processes in the visual stream.

This work proposes to use a hierarchical model of object recognition that learns invariance to various transformations and simultaneously increases its selectivity for specific objects in a step by step manner across several layers of processing. This system is analogous to the way visual information is processed in the initial phase in the primate ventral stream. An important part of this method is the use of scale and position invariant feature detectors that have experimentally been shown to be in quantitative agreement with the selective properties of the cells found in the ventral stream. The features thus extracted are used with a simple linear classifier for classification to demonstrate the inherent variability in these features. This system is tested on several publicly available datasets as well as a dataset of objects commonly found in indoor environments created as a part of this work. The overall objective of this work is to recognize each object class individually in an image containing several object classes together with a cluttered background.

## 2. Aims

2.1. Create a database of images and videos of objects commonly found in indoor environments. The videos should capture full 360 degree view of the objects and the images should be taken from certain pre defined angles also incorporating varying levels of illumination.

2.2. Implement a hierarchical model of object recognition that is based on and experimentally agrees with the processes known to occur in the initial phases of visual processing in the primate visual cortex.

2.3. Use this model to learn various object classes from their images and/or videos and recognize these in an image or video containing several object classes in a cluttered background.

2.4. Compare the results obtained with this model, both in terms of processing time and classification accuracy, with other standard models of object recognition.

## 3. Motivation and Challenges

There are two main motivations for this work: development of a computational model that is consistent with the known properties of the primate visual cortex as a step towards better understanding the way humans recognize objects; and implementation of this model to develop an automatic object recognition system whose performance is superior or comparable to other contemporary systems. Automatic object recognition has several application areas including automated video surveillance systems, environment mapping and localization in robots, content based indexing of videos and images on the World Wide Web and face detection based biometric systems among others.

The task of developing a robust object recognition system presents several challenges, some of which are enumerated below:

3.1. The system needs to be invariant to various object transformations including scaling, rotation and translation as well as different conditions of lighting and illumination.

3.2. The system needs to be able to recognize an object from even those views that are not available in the training set.

3.3. Along with this invariance, the system also needs to have sufficient specificity to recognize individual objects with sufficient accuracy without confusing objects belonging to different classes. Thus a suitable compromise is needed between invariance and selectivity.

3.4. The system is often required to be computationally efficient especially for real time applications like those in surveillance systems and robots.

3.5. The system should be able to work with good accuracy even in the presence of a cluttered background and partial occlusion of the object of interest.

## 4. Literature Survey

There exist a large number of methods for automatic object recognition. These can be classified on the basis of the amount and form of information available about the objects to be recognized as mentioned in [1]. Three main classes of methods can be identified using this criterion:

4.1. Geometry or model based methods: These involve an explicit specification of the object appearance and shape through a three dimensional model (for instance a CAD like model). This model usually specifies only 3D shape related information without any texture, color, reflectance or other surface properties. The recognition task then consists of deciding whether a certain part of an image can be a plausible 2D projection of the 3D object. A survey of several methods in this category is presented in [2]. The main problem with this class of methods is that they are in general good only for objects with well defined geometry that can be specified in terms of simple geometric primitives. The need to create the models manually is another significant drawback.

4.2. Appearance based methods: These do not require any explicit object model to be provided by the user; instead they automatically create representations of the object from images of several views of the object. The model thus created usually relies on surface reflectance properties. This class of methods performs well for unoccluded objects in relatively simple backgrounds but fail to give satisfactory results and also become computationally very expensive for objects in arbitrary or cluttered backgrounds with variations in illumination and occlusion levels. Some appearance based methods are detailed in [3], [4], [5], [6], [7], [8], [9] and [10].

Most of these methods follow a two step process: first the model is extracted from a set of reference images that include images of the object in various orientations and under different lighting conditions, next sub images similar in size to the training images are extracted from the test image and compared with these reference images.

This class of methods also includes histogram based methods in which objects are identified by matching histograms of input image regions to those of the model images. This method was first suggested in [11], [12] and later improved in [13], [14], [15], [16] and [17].

The main problems with this class of methods include the computationally expensive nature of most of these algorithms that require exhaustive sub region extraction from the test image, the very large number of reference images they require to create the model and also their poor performance with images having occlusion and cluttered backgrounds. In fact most of these methods require isolation of the object from its background to give good recognition accuracy.

4.3. Methods based on local features: These methods involve representing objects in terms of localized features extracted from their reference images followed by extracting these same features from the test image and detecting the presence of an object on the basis of the number of local correspondences between its features and those extracted from the test image. Since not all local features need to match for an object to be detected, these methods tend to be robust to cluttered, arbitrary backgrounds and occluded views of the object. Moreover it becomes possible to obtain a view invariant model of an object from comparatively few images since these variations can be modeled by simple affine transformations at the level of these local features.

Several methods have been developed in this class. Some of the important ones include an approach based on selecting transformation invariant anchor points followed by a local histogram based descriptor [18], [19], [20], [21], [22], including the scale invariant feature transform (SIFT) descriptor [23]; choosing corner points as anchor points and describing these with Gaussian derivatives of their intensities [24], [25], [26], [27], [28]; detection of elliptical or parallelogram regions in the image and describing these using a vector of photo metrically invariant generalized color moments [29], [30], [31], [32], [33]; local affine frames (LAF) based approach [34], [35], [36], [37].

The features that are extracted in most of these methods can in general be classified into one of two categories [38]: template based and histogram based. Template based models tend to have high selectivity but low tolerance to variation and thus usually perform well only for detection of a single category of objects like faces [39], [40] or cars [39].Constellation model based methods, described in [41], [42], [43] also fall in this category and have been known to give good performance with very few training samples. Histogram based models, on the other hand tend to be tolerant to object transformations but tend to have low specificity. SIFT based method [23] falls into this category.

The hierarchical model that is proposed to be used in this work also falls into this feature based class of object detection methods. This method is based on a model of object recognition in the primate visual cortex that has been described in [44] and, in considerably greater detail, in [48]. There are also some other hierarchical methods based on this model such as those described in [45] and [46]. Such methods have generally been shown to outperform both template-based and histogram-based methods in terms of finding a good balance between selectivity and invariance to object transformations. Details of the method used in this work have been obtained from [38] and [47].

## 5. Tools and Techniques

Matlab R2011a for implementation of the model.

## 6. Methodology

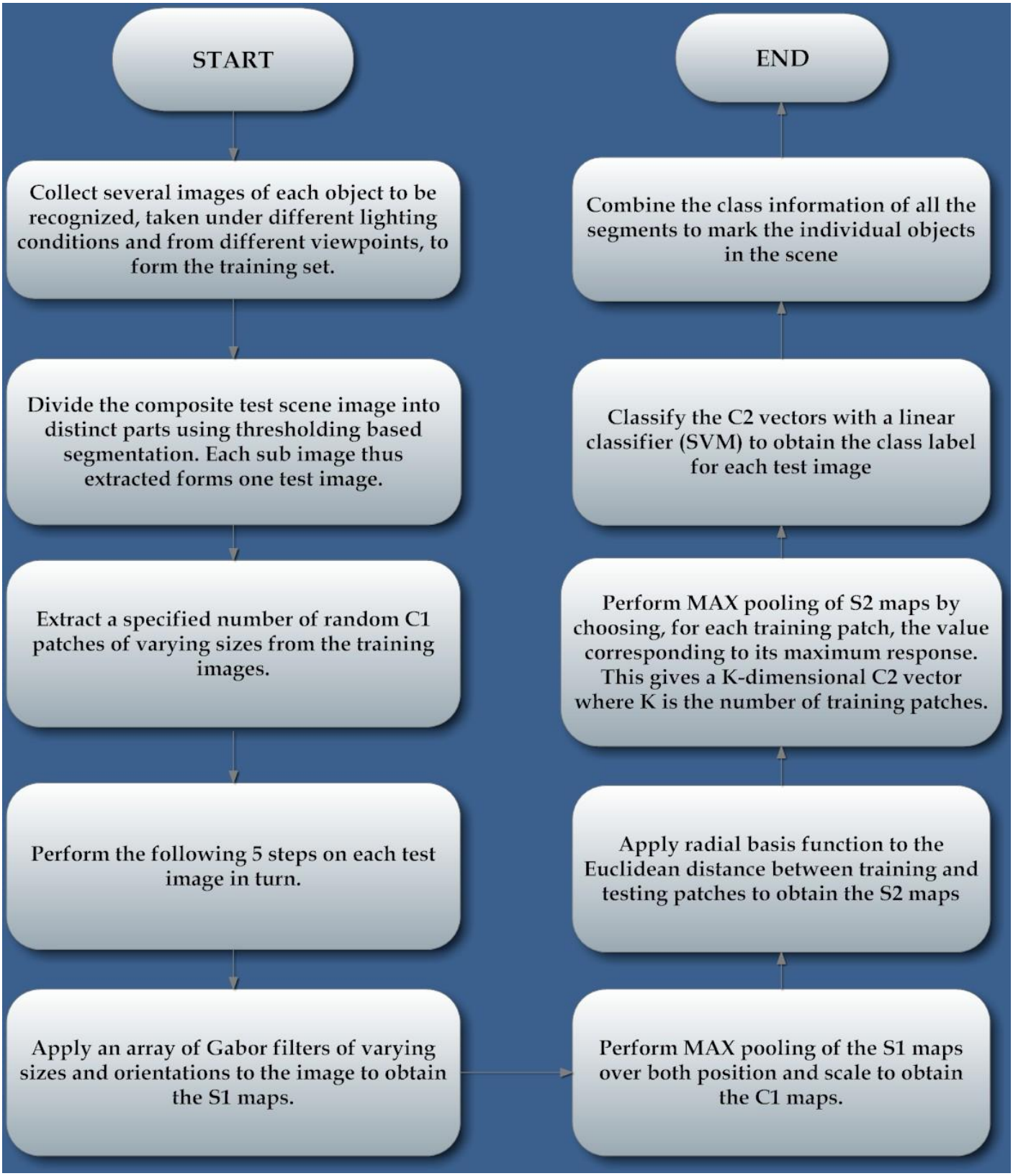


Fig. 6.1 Flowchart of methodology

The methodology used in this work is based on a model of early stages of visual processing in the primate visual cortex that is widely accepted in the neurobiological community. According to this model ([44], [47], [48]), visual processing occurs in a hierarchical fashion by building invariance first to size and position and then to more complex transformations like changes in viewpoint. As we move up the hierarchy, there is an increase in the part of the visual field as well as the actual set of stimuli that elicits a response from a neuron.

This model in its simplest form consists of four layers of computational units (neurons) with simple S units alternating with the more complex C units. The S units are responsible for increasing the selectivity while the C units are involved in building up invariance to various transformations, with simpler ones tackled in the lower layer (C1) and more complex ones in the higher layer (C2). To achieve these targets, the S units combine their inputs with a Gaussian function while C units perform a MAX like pooling operation. This model has proven to be consistent, both qualitatively and quantitatively, with the properties of cells along the ventral stream of the visual cortex ([48]). These two layers of cells (S1, C1 and S2, C2) correspond roughly to the simple and complex cells found in the V1 and V4 sub regions of the ventral stream.

The process of implementation of this model, as has been used in this work, has been summarized in Fig. 1 and is described below in more detail:

6.1. S1 units: These are implemented by processing the grey level input image by an array of Gabor filters, which are mainly used for detecting edges in an image. These filters can be described by the following equation:

$$F(x,y) = \exp\left(-\frac{x_0^2 + \gamma^2 y_0^2}{2\sigma^2}\right) + \cos(\frac{2\pi}{\lambda} x_0)$$

Here, $x_0 = x\sin\theta + y\sin\theta$ and $y_0 = -x\sin\theta + y\cos\theta$ .

In the above equation, γ=0.3, effective width σ, orientation θ and wavelength λ are parameters whose values are chosen so that the tuning properties of these S1 units match those of the corresponding cells in V1, based on the data obtained from testing both of these groups on similar images, as presented in [49], [50], [51] and [52]. A total of 64 Gabor filters are used, varying in sizes from 7 x 7 to 37 x 37 pixels in steps of two pixels, with each such size having filters in 4 orientations from 0 to 135 degrees in steps of 45 degrees. These filters are divided into 8 scale bands with each band containing filters of all 4 orientations and 2 consecutive scales. For example, scale band 1 contains all filters of sizes 7 x 7 and 9 x 9. Thus each band contains 8 S1 maps.

6.2. C1 units: These units correspond to the complex cells of V1 that show some tolerance to position and scale. These work by taking the max over both scales and

different positions for each band. This involves first assigning neighborhood sizes ($N^{\Sigma}$) to each band, varying from 8 x 8 to 36 x 36 in steps of 2. Each band member is then sub-sampled by taking the max over all pixels of each neighborhood and then max over the two corresponding neighborhoods in the two scales with the same orientation in that band. This process is repeated independently for each orientation in each band. Thus each band will have four C1 maps. During the training phase, there is an additional step of extracting K patches ($P_{i=1,2,\ldots..K}$) of varying sizes and all four orientations at random from the C1 maps of all the training images.

6.3. S2 units: These units pool over the responses of C1 units from a local spatial neighborhood from all four orientations by applying the radial basis function on the Euclidean distance between a new input and a stored prototype previously extracted from the training images. This works by computing, for each C1 map:

$$Y = \exp(-\beta \ ||X - P_i||^2)$$

This is carried out for all image patches X and each patch $P_i$ learned during training, for each band independently. Here $\beta$ is a parameter that defines the sharpness of the tuning. One such S2 map is computed for each one of the prototypes $P_i$.

6.4. C2 units: These provide additional shift and scale invariance by computing the max over all scales and positions for each S2 map type, i.e. over all the S2 maps corresponding to a particular patch $P_i$ to obtain a total of K C2 maps. The idea here is to measure the match between a stored prototype and the input image at every position and scale through the S2 units and subsequently use the C2 units to retain only the best match for each prototype, while discarding the rest.

During classification, each image is propagated through all the layers and the standard model features (SMFs) or feature vectors obtained as outputs of the C2 units are passed to a linear classifier based on Support Vector Machines. For recognizing each object class in an image containing several objects, the image first has to be segmented into objects and then the sub-image corresponding to each segment thus obtained is passed through this system to identify the object present in it.

In this work, Otsu's thresholding based method [49] is used for the segmentation step. This produces a mask in which each pixel in the image in assigned one of the given number of classes. This mask is then used to extract polygonal, preferably rectangular, sub regions from the image such that each such sub region contains a majority (or entirety) of pixels belonging to the same class. This has been accomplished in this system by finding, for each class, all the distinct squares within the image that contain more than a specified percentage of pixels belonging to that class. Each such square is the extracted to form a sub image that is passed through the classification system.

## 7. Results and Analysis

The hierarchical model described above has been implemented completely in MATLAB. However the code for Otsu segmentation method has been obtained from [55] since image processing is not the focus of this work. Wrapper code has also been written to integrate the segmentation and object recognition portions of this work and to finally piece together the classification results of individual sub images to produce the final composite image with labeled objects. This last task was not entirely completed since there was an unresolved issue about the alignment of the coordinate axes of the test image and the rectangles and labels that were overlaid on it.

This system has been tested with the following databases:

7.1. Several publicly available datasets including Caltech 101/256 [49][50] databases, ETH80 database [51], Object recognition and 3D stereo databases of the Ponce research group [52] and CBCL StreetScenes database [53]. These databases only contain images of individual objects and no natural composite image containing several of these objects. Thus the testing can either be done by dividing the total image set into two parts and using one for training and the other for testing, or by creating an artificial composite image by arranging images of randomly chosen objects to form a mosaic.

7.2. A custom created databases consisting of two sets of objects, one from my hostel room containing 9 object classes and another from a computer lab containing 4 classes. These sets have their own composite test images with some or all of the training objects arranged in a natural way.

This system has been tested using two different methods:

1. Evaluate its classification accuracy over the sets of object images themselves without using any composite images. For each database, a part (usually half) of the set of object images is used for training and the remaining for testing. This method produces an accuracy figure as the percentage of test images that were classified correctly.
2. Detection of individual objects in a composite image of a scene containing several objects. This process involves first segmenting the scene image, followed by the extraction of rectangular sub images containing individual objects from the overall image. All the available object class images in the database are then used for training and these extracted images are used for testing. Finally, the sub images are joined together along with their class labels to reconstruct the original scene image with the individual detected objects marked therein.

Presented below are some images collected from publicly available databases as well the custom database.

Fig. 7.1 Images of ceiling fans, chairs and lamps from Caltech101 database [49].


Fig. 7.2 Images of backpacks, bread-makers and coffee mugs from Caltech256 database [50].

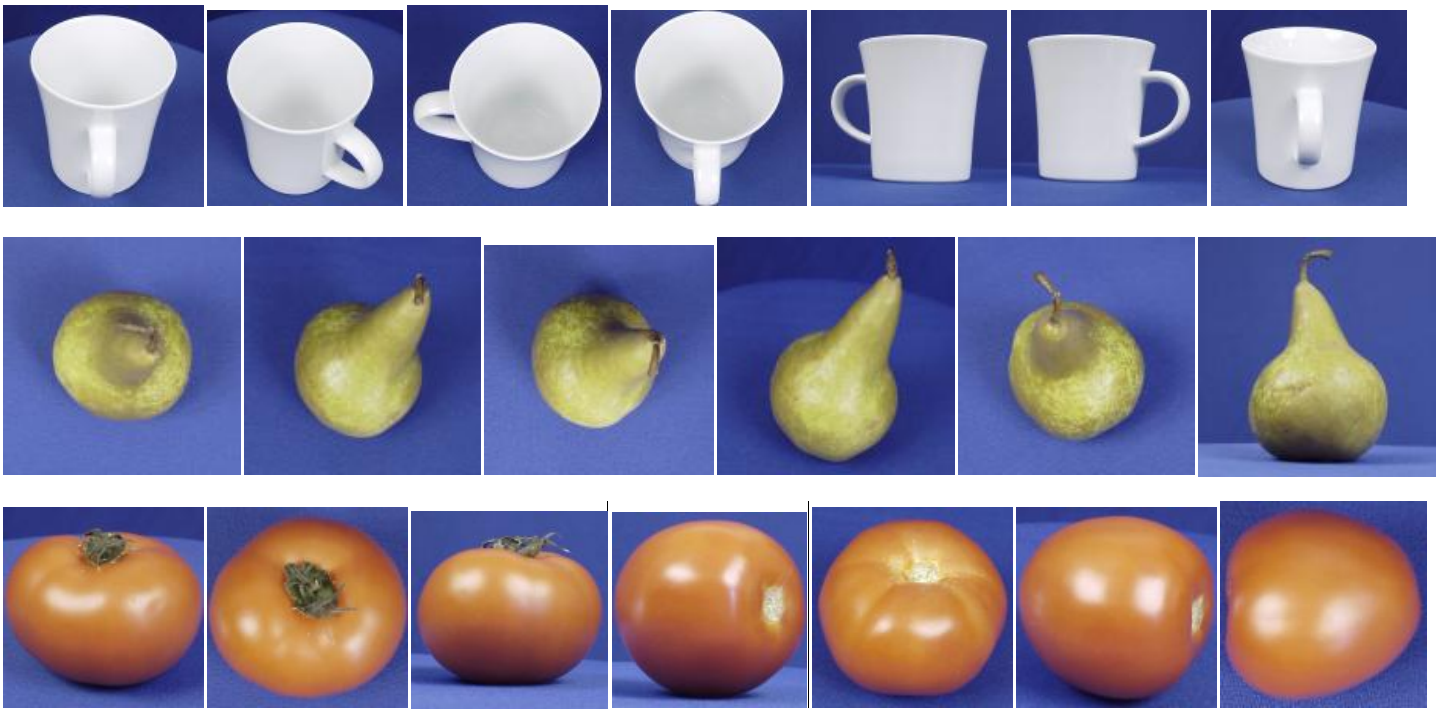

Fig. 7.3 Images of cups, pears and tomatoes from ETH80 database [51].

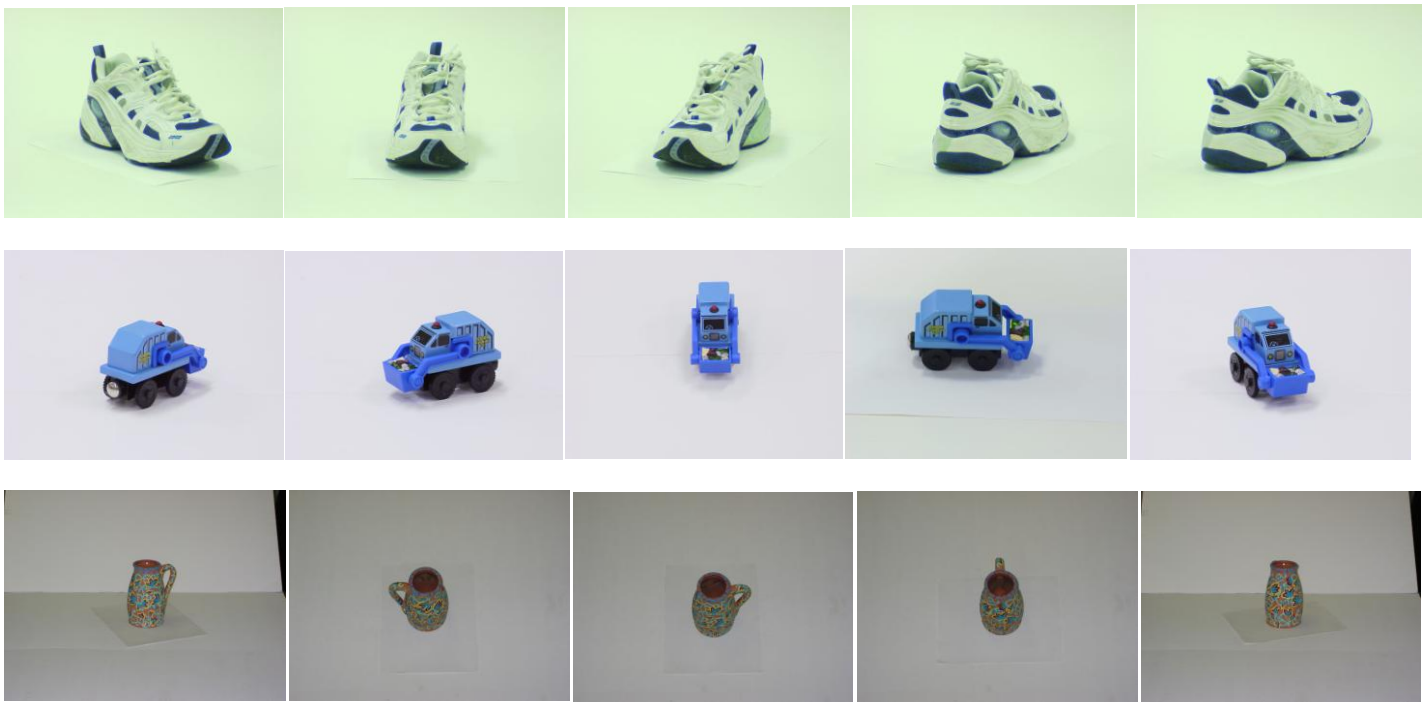

Fig.7.4 Images of a shoe, toy truck and a jug from the 3D Stereo Database of Ponce Research group [52]

Following are images of several objects from the custom created room and lab datasets.

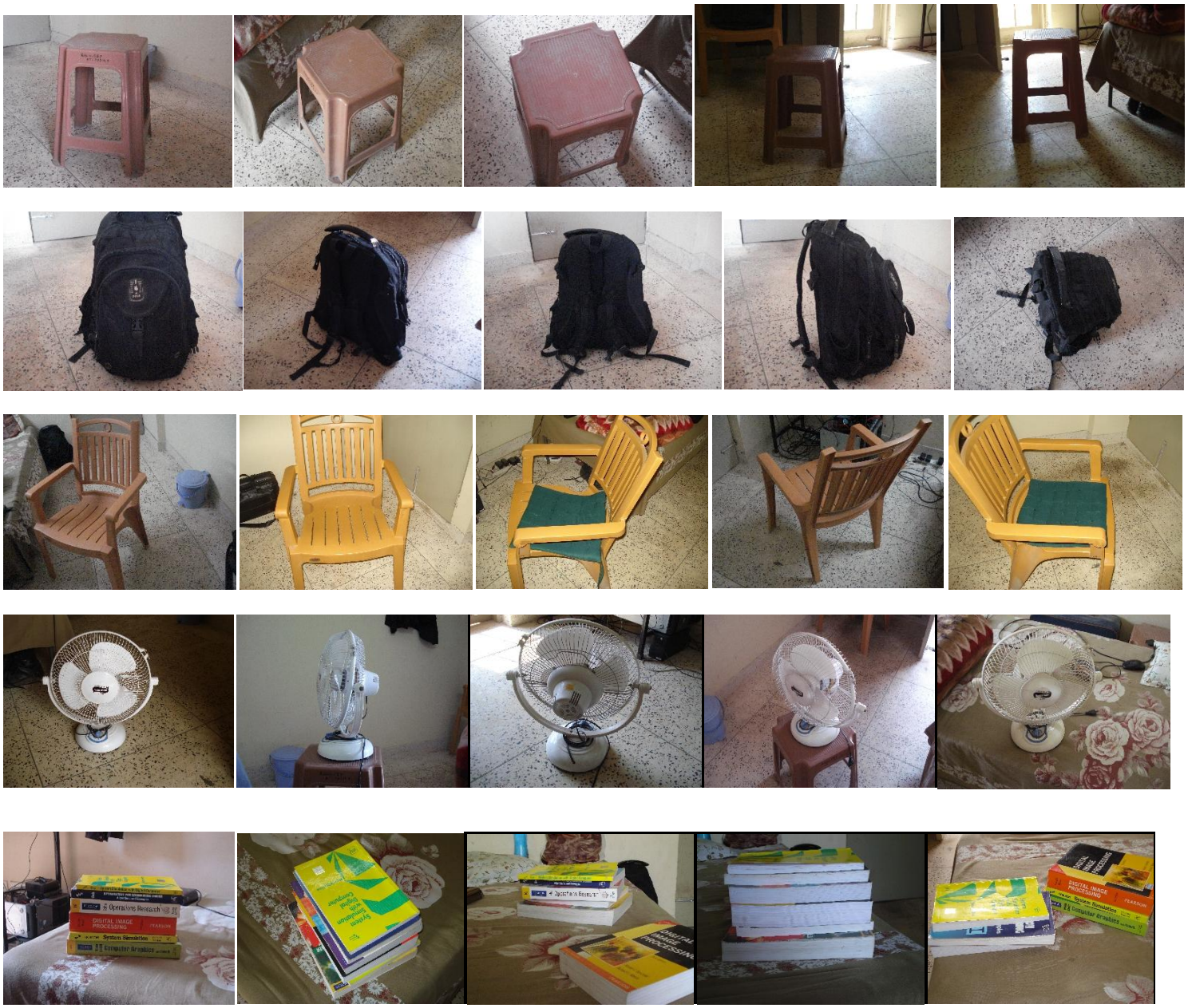

Fig. 7.5a Images of stool, backpack, chair, table fan and books from the custom prepared room dataset.

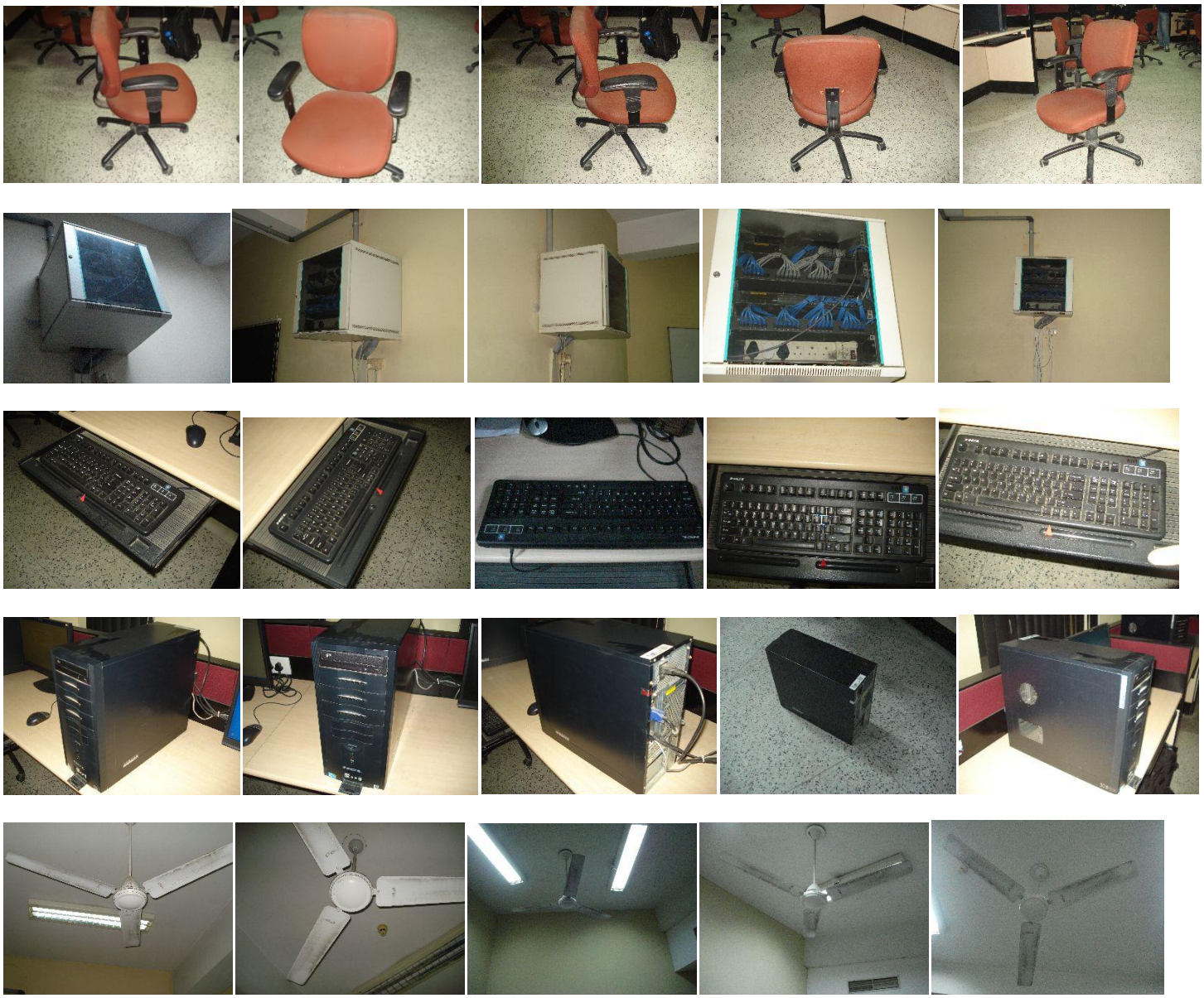

Fig. 7.5b Images of chair, router, keyboard, CPU and ceiling fan from the custom prepared lab dataset.

Following are images of composite scenes containing several objects that will be given as input to the system for object detection


Fig. 7.6 Composite images of room and lab scenes used as input for the object detection part.

Following are the results of the some of the many classification tests that were run:

| **Database** | **No. of Classes** | **No. of C2 patches** | **No. of testing/training images** | **Percent Accuracy** |
|---|---|---|---|---|
| Custom room dataset | 10 | 1000 | 111 | 60.36% |
| Custom room dataset | 10 | 1000 | 100 | 53.00% |
| Custom lab dataset | 6 | 1000 | 36 | 66.67% |
| Caltech 101 | 8 | 1000 | 192 | 80.73% |
| Caltech 101 | 10 | 1000 | 150 | 72.67% |
| Caltech 256 | 3 | 400 | 161 | 78.88% |
| ETH80 | 8 | 400 | 80 | 67.50% |
| ETH80 | 8 | 1000 | 160 | 94.38% |
| Ponce Research Group | 8 | 400 | 81 | 100% |
| Ponce Research Group | 8 | 1000 | 81 | 98.77% |

Table 7.1 Results of classification tests on various databases

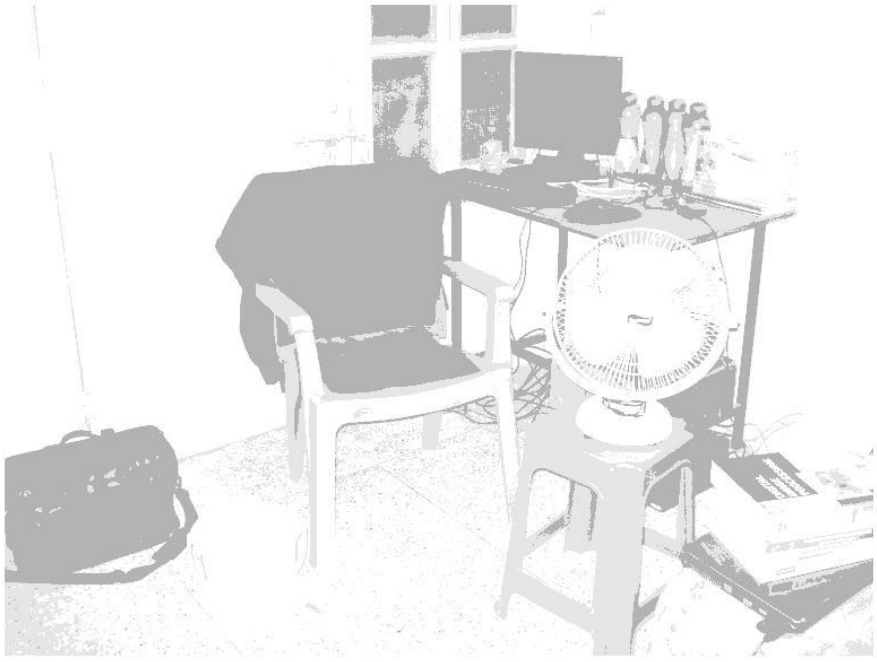

Fig. 7.7 Segmented images of room and lab scenes

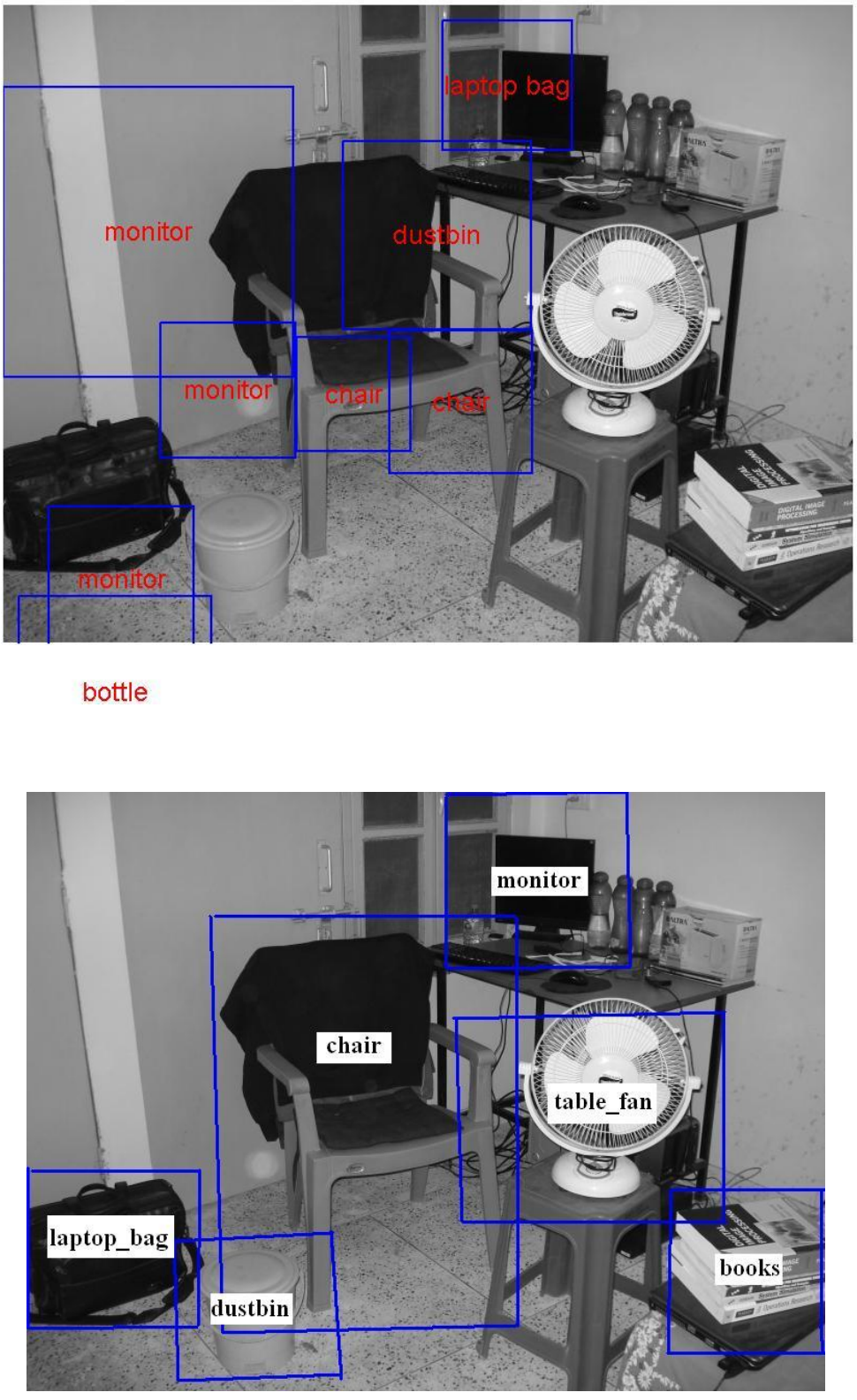


Fig. 7.8a Actual results and the supposed results with the room scene image and plot/label axes aligned

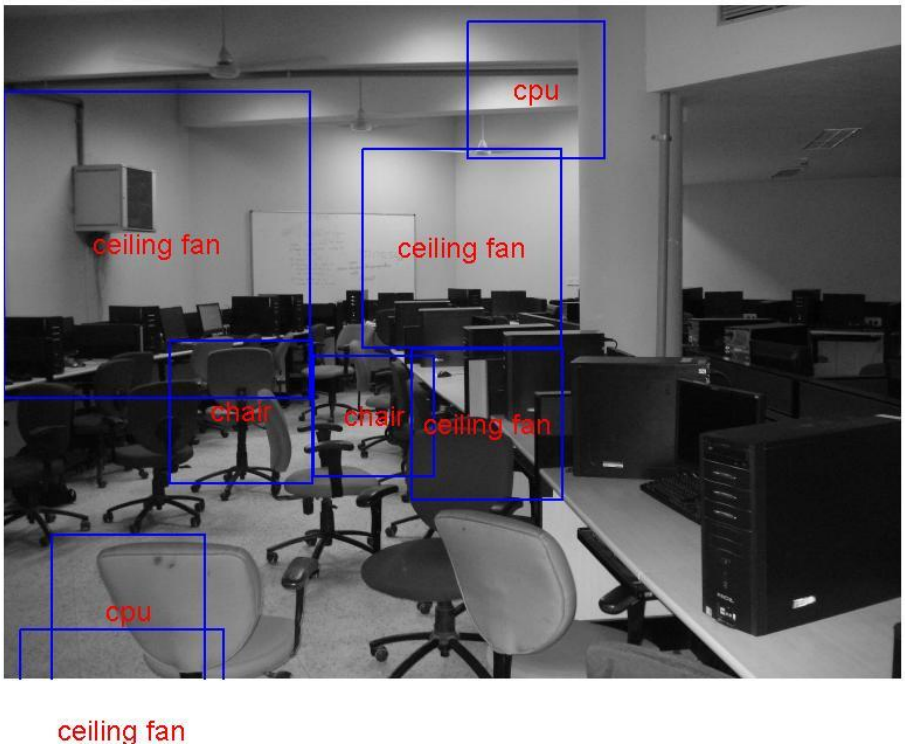


Fig. 7.8b Actual results with the lab scene image.

## 8. Conclusion

8.1. The relatively high accuracy obtained with this method demonstrates the superiority of biologically inspired models of object recognition.

8.2. The difference in the accuracy figures of natural and artificial test images demonstrates the difficulty most automatic object recognition systems have in dealing with unstructured, cluttered backgrounds as well as arbitrary relative positions and orientations of different objects in the scene.

## 9. Future Scope

9.1. This system can be extended to perform real-time object recognition and tracking in video sequences of complex indoor scenes rather than just still images. This will, however, require performance of this system to be improved by a great degree since it is far too slow in its current form to perform recognition or even segmentation in real time.

9.2 The above task can in turn be achieved by using more efficient segmentation technique, faster linear classifier or a more efficient implementation of the base model.

## 10. References

[1] Matas, J. and Obdrzalek, S. "Object Recognition Methods Based On Transformation Covariant Features." *12th European Signal Processing Conference*, 2004.

[2] A.R. Pope. "Model-based object recognition: A survey of recent research." In *Univ. of British Columbia*, 1994.

[3] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman "Eigenfaces vs fisherfaces Recognition using class specific linear projection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 19, Issue 7, July 1997.

[4] T.E. Boult, R.S. Blum, S.K. Nayar, P.K. Allen and J.R. Kender "Advanced visual sensor systems." *DARPA98*, pp. 939–952, 1998.

[5] Moghaddam, B. "Probabilistic visual learning for object detection." In *Proc. IEEE Fifth International Conference on Computer Vision*, pp. 786-793, June 1995

[6] J. Ruiz-del-Solar "Eigenspace-based face recognition: a comparative study of different approaches." *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* Volume 35, Issue 3, pp. 315-325, Aug. 2005.

[7] Nayar, S.K. "Real-time 100 object recognition system." in *Proc. IEEE International Conference on Robotics and Automation*, Volume 3, pp. 2321-2325, Apr. 1996

[8] T. Heseltine, N. Pears, J. Austin, Z. Chen "View-based and modular eigenspaces for face recognition." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 84-91, June 1994

[9] Swets, D.L. "Using discriminant eigenfeatures for image retrieval." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 18, Issue 8, pp.831-836, Aug. 1996

[10] Matthew Turk and Alex Pentland "Eigen faces for recognition." *Journal of Cognitive Neuroscience*, Volume 3, Issue 1, pp.71–86, 1991

[11] M.J. Swain "Indexing via color histograms." in *Proc. Third International Conference on Computer Vision*, pp. 390-393, Dec.1990.

[12] M.J. Swain, Dana H. Ballard "Color indexing" *International Journal of Computer Vision*, Volume 7, Issue 1, pp. 11-32, June 1991.

[13] Brian V. Funt, Graham D. Finlayson "Color constant color indexing." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 17, Issue 5, May 1995.

[14] Glenn Healey, David Slater "Using illumination invariant color histogram descriptors for recognition." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 355-360, June 1994.

[15] Bernt Schiele , James L. Crowley "Object recognition using multidimensional receptive field histograms." *www-prima.inrialpes.fr/jlc/papers/IJCV00-Schiele.pdf*, Sept. 2012

[16] Bernt Schiele "Probabilistic object recognition using multidimensional receptive field histograms." in *Proc. 13th International Conference on Pattern Recognition*, Volume 2, pp. 50-54, Aug. 1996.

[17] Bernt Schiele, James L. Crowley "Recognition without correspondence using multidimensional receptive field histograms." *International Journal of Computer Vision*, Volume 36 Issue 1, pp. 31-50, Jan. 2000.

[18] Jeffrey S. Beis , David G. Lowe "Indexing without invariants in 3d object recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 21 Issue 10, pp. 1000-1015, Oct. 1999.

[19] M. Brown, D.G. Lowe "Invariant features from interest point groups." in *Proc. British Machine Vision Conference*, pp.656-665, Sept. 2002

[20] M. Brown, D.G. Lowe, "Recognizing panoramas." in *Proc. Ninth IEEE International Conference on Computer Vision*, pp. 1218-1225, Oct. 2003.

[21] D.G. Lowe "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, Volume 60 Issue 2, pp. 91-110, Nov. 2004

[22] D.G. Lowe "Local feature view clustering for 3D object recognition." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 682-688, Dec. 2001.

[23] D.G. Lowe "Object Recognition from Local Scale-Invariant Features." in *Proc. International Conference on Computer Vision*, Volume 2, pp. 1150, Sept. 1999.

[24] K. Mikolajczyk "Indexing based on scale invariant interest points." in *Proc. Eight IEEE International Conference on Computer Vision*, Volume 1, pp. 535-531, July. 2001.

[25] Fred Rothganger , Svetlana Lazebnik , Cordelia Schmid , Jean Ponce "3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-

View Spatial Constraints." *International Journal of Computer Vision*, Volume 66 Issue 3, pp. 231-259, March 2006.

[26] C. Schmid "Constructing models for content-based image retrieval." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 39-45, Dec. 2001.

[27] Gy. Dorkó , C. Schmid "Selection of scale-invariant parts for object class recognition." in *Proc. Ninth IEEE International Conference on Computer Vision*, Volume 2, pp. 634, Oct. 2003

[28] C. Schmid, R. Mohr "Combining greyvalue invariants with local constraints for object recognition." *in Proc. Conference on Computer Vision and Pattern Recognition*, pp. 872, June,1996.

[29] T. Tuytelaars "Local Invariant Features for Registration and Recognition." PhD. dissertation, University of Leuven, ESAT - PSI , Belgium, Dec. 2000.

[30] T. Tuytelaars, Luc J. Van Gool "Content-based Image Retrieval based on Local Affinely Invariant Regions." in *Proc. Third International Conference on Visual Information and Information Systems*, pp. 493-500, June, 1999.

[31] T. Tuytelaars, L. Van Gool, L. D'haene, R. Koch "Matching of affinely invariant regions for visual servoing." *IEEE International Conference on Robotics and Automation*, Volume 2, pp. 1601-1606, May, 1999.

[32] T. Tuytelaars, L. Van Gool "Non Combinatorial detection of regular repetitions under perspective skew." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 25, Issue 4, pp. 418-432, April, 2003.

[33] T. Tuytelaars, L. Van Gool "Wide baseline stereo matching based on local, affinely invariant regions." in *Proc. British Machine Vision Conference*, pp. 42-56, Sept. 2000.

[34] Stepán Obdrzálek , Jiri Matas "Local Affine Frames for Image Retrieval." in *Proc. International Conference on Image and Video Retrieval*, pp. 318-327, July, 2002.

[35] Stepán Obdrzálek , Jiri Matas "Object recognition using local affine frames on distinguished regions." in *Proc. British Machine Vision Conference*, Volume 1, pp. 113-122, Sept. 2002.

[36] J. Matas, O. Chum, U. Martin, T. Pajdla "Robust wide baseline stereo from maximally stable external regions." in *Proc. British Machine Vision Conference*, Volume 1, pp. 384-393, Sept. 2002.

[37] Matas, J.; Obdrzalek, T.; Chum, O "Local affine frames for wide-baseline stereo." in *Proc. 16th International Conference on Pattern Recognition*, Volume 4, pp. 363-366, July, 2002.

[38] Thomas Serre, Lior Wolf, Tomaso Poggio "Object Recognition With Features Inspired By Visual Cortex." in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 994-1000, June, 2005.

[39] Henry Schneiderman , Takeo Kanade "A Statistical Method for 3D Object Detection Applied to Faces and Cars." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Volume 1, pp. 746-751, June, 2000.

[40] Paul Viola, Michael J. Jones "Robust real-time face detection." *International Journal of Computer Vision*, Volume 57 Issue 2, pp. 137-154, May, 2004.

[41] Li Fei-Fei , Rob Fergus, Pietro Perona "Learning generative visual models from few training examples." *Computer Vision and Image Understanding*, Volume 106 Issue 1, pp. 59-70, April, 2007.

[42] R. Fergus, P. Perona, A. Zisserman "Object class recognition by unsupervised scale-invariant learning." in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 264-271, July, 2003.

[43] Markus Weber, Max Welling, and Pietro Perona "Unsupervised learning of models for recognition." in *Proc. 6th European Conference on Computer Vision*, Part I, pp. 18-32, July, 2000.

[44] Maximilian Riesenhuber, Tomaso Poggio "Hierarchical Models of Object Recognition in Cortex." *Nature Neuroscience*, Volume 2, Issue 11, pp. 1019-1025, 1999.

[45] Bernd Heisele, Thomas Serre, Massimiliano Pontil, Thomas Vetter, Tomaso Poggio "Categorization by learning and combining object parts." in *Proc.16th Annual Conference on Neural Information Processing Systems*, 2002.

[46] Yann LeCun , Fu Jie Huang , Léon Bottou "Learning methods for generic object recognition with invariance to pose and lighting." in *Proc. IEEE Conference on Computer vision and pattern recognition*, pp. 97-104, April, 2004.

[47] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, Tomaso Poggio "Robust Object Recognition With Cortex-Like Mechanisms." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 29, Issue 3, pp. 411-426, March 2007.

[48] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, T. Poggio "A Theory of Object Recognition Computations and Circuits in the Feedforward Path of the Ventral

Stream in Primate Visual Cortex." Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Technical Report, MIT-CSAIL-TR-2005-082, Dec. 2005.

[49] Chen Yu,Chen Dian-ren, Li Yang,Chen Lei "Otsu's thresholding method based on gray level-gradient two-dimensional histogram" in *Proc. 2nd international Asia conference on Informatics in Control, Automation and Robotics*, Volume 3, pp.282-285, March 2010.

[50] Caltech101 object database.
http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[51] Caltech256 object database.
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

[52] ETH80 database.
http://www.d2.mpi-inf.mpg.de/Datasets/ETH80

[53] Image databases of the Ponce research group.
http://www-cvr.ai.uiuc.edu/ponce_grp/data/

[54] CBCL StreetScenes database.
http://cbcl.mit.edu/software-datasets/streetscenes/

[55] Damien Garcia, "Image segmentation using Otsu thresholding", March 10, 2010, Available at: http://www.mathworks.in/matlabcentral/fileexchange/26532-image-segmentation-using-otsu-thresholding, December 3, 2012