CS365 Project
Indoor scene classification

Anuja Ranjan
Manav Garg
Advisor : Dr. Amitabha Mukherjee
Dept. of Computer Science and Engineering
{anuja, mgarg, amit} @ iitk.ac.in

10 April,2012

Abstract

Scene classification has now become an active area of research. A lot of work has been done on classifying images into outdoor and indoor categories however, current approaches for scene recognition show a significant drop in performance for the case of indoor scenes. Classifying indoor scenes is a challenging task due to the large variation across different examples within each class and similarities between different classes.Besides spacial properties it requires us to see the objects they contain. Exploiting this idea and following the work of Espinace et al[2], we present a new procedure to classify real world indoor scenes.Here we learn object classifiers with Gist features and use them to find objects in the scene and then classify them.The accuracy obtained are better than those obtained with the previous works.

1 Introduction

One of the distinct features of humans is the ability to differentiate between things i.e. to be able to identify them and to link them with some prior information about the same. Similar is the concept of classification,clustering,etc in artificial intelligence.Learning algorithms give us a method to find features and parameters which would be able to identify and classify different objects.
In the learning method a training set is initially created with representative and labeled images from all categories. Then a learning algorithm is employed, which enables us to come up with parameters which would characterize an image for doing the classification task. Then if a random image is given as an input, on basis of the learned parameters the machine would try to classify the image. This in essence is a generic way in which learning algorithms work, i.e., by learning from a huge set of data and then using this learned information to make predictions about successive inputs.
We consider here a similar basic classification problem.Robotic vision is still quite miserable as compared to human vision.It is quite trivial for us to classify an image of an indoor scene but it is not the same for computers to do so. The problem of indoor classification has come up strongly in the last decade and it is still an open problem. Scene classification is an important task in navigation systems.The use of this classifier could also be in mobile robots which still lack the ability of understanding their surrounding place in indoor environments.In this project we extend the previous works by some improvisations and improve upon the accuracy of the results obtained by earlier methods.

2 Related Work

There are many ways in which previous works have approached this problem. Methods that work well for outdoor scene classification[5] show considerable drop in accuracy for indoor scenes. Earliest methods used low-level features like color or texture for classifying different scene categories. Vailaya et al.[12] improved this approach by the use of heirarchial classification to classify scenes as indoor or outdoor scenes.In the work by Lu et al.[13] they use a mixture model scheme but their method, too gave poor results for indoor scenes.After this more reliable global approaches were used. Color histograms and a k-nearest neighbors scheme was used by Ulrich and Nourbakhsh[14]. Oliva and Torralba[3] use global features(GIST) based on spectral information . Yang et al.[15] uses bag of visual words scheme but they do not give good results for indoor scenes. The main problem with all these approaches has been their inability to generalize from the training data to new scenes.In the work by Bosch et al[8], they provide an elaborate overview of all the methods(up to 2007), that have been used in this field and the problems faced by each.

In a recent work by Torralba and Quattoni[1] regions of interests are extracted from images and compared with those in the prototype image. As such they do not use objects in their approach but they do mention that some indoor scenes are better classified by the objects they contain.In the work by Espinace et al[2,16] they proposed a new approach for indoor scene recognition based on a probabilistic hierarchical representation that uses common objects as an intermediate semantic representation. The main intuition being that one can associate low-level features to objects through object classifiers, and also associate objects to scenes using contextual relations.

3 Methodology

In our approach we shall be detecting objects in the scene and classifying it accordingly.We shall mainly be referring to the work on indoor scene recognition via object detection by Espinace et al.[2].This paper was implemented as the first phase of our project. The method described in this paper is mainly divided into 2 phases.

3.1 Training Phase

We firstly need to build object classifiers for a few object categories.The objects are selected such that they are closely related to the scene categories.As used in the paper we work with 3 Scene categories(Office,Hall and Conference Room) and 4 objects(Monitor,Projector Screen,Door and Clock). Database for these is developed using the Caltech Dataset and random Google Images.Each object classifier was learnt with 150 object images and 500 background images.For each image in the dataset we apply various gabor filters and obtain HOG features for different number of bins.Hence a set of 460 different features are extracted.With each of these features we learn a linear seperator classifier.These are weak classifiers hence we learn a strong classifier from them with the help of AdaBoost algorithm.

3.2 Testing Phase

Given any random image we firstly need to detect the objects contained in them.For this we apply the learnt classifiers on the image with the help of sliding window procedure.Sliding window is a very common technique used in object detection where instead of applying the classifier on the entire image we apply it on various subimages.Here we use 5 different shapes of windows each of 7 different sizes.Once an object is found 3D models are used to compute the probability of matching the geometric properties(Size,Height) of a given object and the information present in a given window.This is basically to improve object classification.Each image in their dataset consists of a visual and corresponding 3D image.The 3D images were taken with the help of SR Ranger Camera. The depth in 3D image is estimated per pixel by measuring the time difference between the signal sent and received.

$$dh_i|\ c_i \approx N(\mu_a,\sigma_a)$$
$$ds_i|\ c_i \approx N(\mu_b,\sigma_b)$$

The distribution for each property is modeled as a pre-learned Gaussian distribution in the code and hence we directly obtain the 3D probabilty as well.If both these probabilities are greater than a given threshold then the object is said to have been found.3D information is used mainly to discard unlikely object locations and sizes.
If S represents the scene category,f and d represent the visual and 3D feature properties, o and c represent the objects and classifiers respectively and w represents the set of windows then once the objects are found then the probability of the scene can be found as

$$\begin{aligned}\mathrm{P}(\mathrm{S}|\ f_{1:w_L},d_{1:w_L}\ ) &= \sum_{o_{1:s}}\sum_{c_{1:w_L}} \mathrm{P}(\mathrm{S}|\ f_{1:w_L},d_{1:w_L},o_{1:s},c_{1:w_L})\mathrm{P}(o_{1:s},c_{1:w_L}|f_{1:w_L},d_{1:w_L}) \\ &= \sum_{o_{1:s}}\sum_{c_{1:w_L}} \mathrm{P}(\mathrm{S}|o_{1:s})\mathrm{P}(o_{1:s}|c_{1:w_L})\mathrm{P}(c_{1:w_L}|f_{1:w_L},d_{1:w_L})\end{aligned}$$

- $\mathrm{P}(\mathrm{S}|o_{1:s})$ represents the a priori probability of the scene given that object.These values have been estimated with the help of large set of test images[4].
- $\mathrm{P}(o_{1:s}|c_{1:w_L})$ represents the classifier confidence which tells the estimate of the confidence with which the classifier can classify the objects correctly.This is again estimated by counting the number of true-positives and false-positives provided by the classifiers on test datasets.
- $\mathrm{P}(c_{1:w_L}\ |\ f_{1:w_L},d_{1:w_L})$ represents the probability of the object being contained in the scene as calculated above.

3.3 Gist Features

Gist features[3], developed by Quattoni and Torralba provide a low dimensional representation of an image.They differ from the traditional methods in the sense that they do no segment the image or use edge/object detection.These are obtained by extracting the following 5 perceptual feature of an image:

- degree of naturalness:structure of scene differs in man-made and natural environments. Straight lines of horizontal and vertical nature dominate man-made scenes, where as natural images show textured zones and undulating contours.

- degree of openness: Openness gives a sense of enclosure. A scene can be enclosed by visual references like in the case of forest, mountains etc or they can be vast like the coastal area.
- degree of roughness: roughness in a scene primarily refers to the size of its major components. It depends on the size of elements at each spatial scale, their abilities to build complex elements and their relations between elements that are also assembled to build other structures.
- degree of expansion: especially the concept of parallel lines giving the perception of depth. so a flat view of a building has low degree of expansion where as a street with long vanishing lines has a high degree of expansion.
- degree of ruggedness: it refers to the deviation of the ground with respect to the horizon. A rugged environment produces oblique contours in the picture and hides the horizon.

The performance of the spatial envelope model shows that specific information about object shape or identity is not a requirement for scene categorization.The method of extraction of these features and the proof of their correctness has been argued in the same paper.

4 Improving previous work

Gist can roughly tell what kind of a scene it is.In a series of work[5,6,7] by Torralba's group these features have shown a much improved accuracy over traditional object detection methods.There is a general consensus that context can be a rich source of information about an object's identity,location and scale in an image.Also object recognition methods based on intrinsic object features can handle many transformations such as displacement, rotation, etc but in situations with poor viewing quality context appears to play a major role in achieving better recognition. Hence we concluded that gist features should improve object recognition here and prove to be a better priors.So we decided to learn a joint probability distribution between gist and the object categories and thus replaced the HOG and Gabor filters with Gist features in the hope of improving the classifier accuracy.
With the same datasets used in the previous classifier we extracted out the gist features of each image and using an SVM built a binary classifier for each of the 4 objects.The acurracy of the binary classifier was increased by randomly changing the parameter of the kernel and selecting the best one.Now given any random scene we first get the probabilities of presence of these 4 objects with the help of gist of the image.And then further improve the classification by taking into account the 3D probabilities as well.If G is the gist feature and d is the 3D feature then given the object probabilities we obtain the scene probabilty as

$$\begin{aligned} \mathrm{P}(\mathrm{S}|\ G,d\ ) &= \sum_{o_{1:s}} \mathrm{P}(\mathrm{S}|\ G,d,o_{1:s},c)\mathrm{P}(o_{1:s},c|G,d) \\ &= \sum_{o_{1:s}} \mathrm{P}(\mathrm{S}|o_{1:s})\mathrm{P}(o_{1:s}|c)\mathrm{P}(c|G,d) \end{aligned}$$

The code for the same was written in MATLAB and used library[10] for gist feature and LIBSVM[11] for learning the classifier.We tried to incorporate the sliding window procedure for determining the objects but the classifier accuracy obtained was very poor hence we sort to applying the classifier directly over the entire image.

4.1 Results

4.1.1 Dataset 1

The first dataset on which we computed Gist features was provided by P. Espinace. It consists of a sequence of images and their corresponding 3D versions obtained using the Swiss Ranger sensor.A few examples are shown below.These images are used to calculate the 3D probability of presence of objects within a scene . If an object is detected in a scene in which it is not present, then it is discarded by comparing its 3D probability relative to a threshold.

Images from P Espinace's dataset

The use of Gist features improved the results in Espinace's paper.Both the methods were run on an office image sequence(208 images),a conference room sequence(156 images) and a hall image sequence(200 images) from Espinace's dataset.Overall object and scene classifier accuracy was highly improved. Out of these images tested on all 4 object categories earlier method showed a wrong object detection i.e. in images where no objects were present few classifiers showed a higher accuracy while images in which the objects were present were not detected.Thus the scenes were wrongly classified.With our method number of such images was considerably reduced.

Table 1: Confusion Matrix:Earlier Method

	Office	Conference Room	Hall	None[1]
Office containing monitor	53%	0%	0%	47%
Conference Room containing Screen	0%	48%	32%	20%
Hall containig door	0%	18%	65%	17%
None[2]	24%	5%	19%	52%
Features used: Gabor, HOG, visual 3D				

[1]Images which could not be classified
[2]Images containing none of the objects

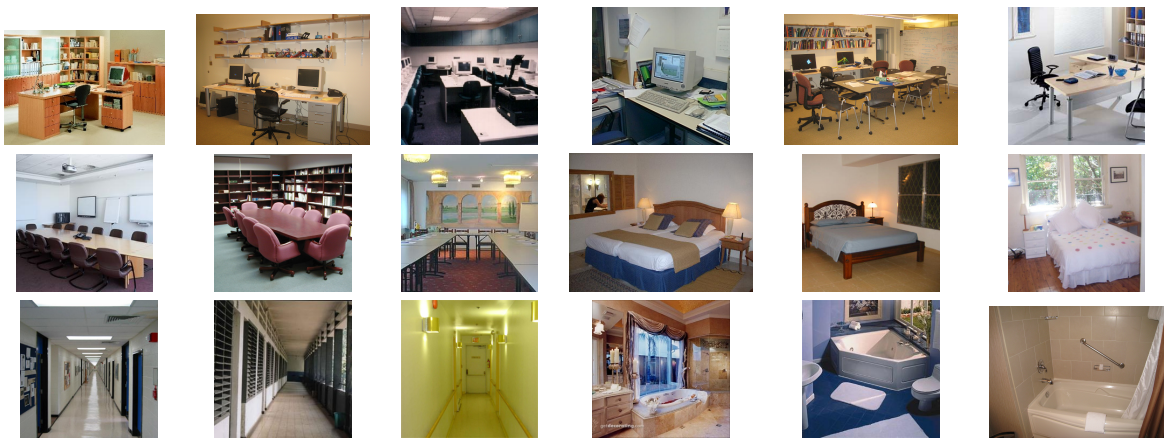
Table 2: Confusion Matrix:Our Method

	Office	Conference Room	Hall	None[1]
Office containing monitor	84%	0%	0%	16%
Conference Room containing Screen	0%	90%	0%	10%
Hall containig door	0%	0%	94%	6%
None[2]	4%	0%	10%	86%
	Features used: Gist, visual 3D			

The results mentioned in the paper are explicitly carried on images in which objects are detected but considering the low object detection accuracy our classifier achieves a better precision.

4.1.2 Dataset 2

The method was also tested on a second dataset obtained from Torralba[4].It contains images belonging to different scene categories collected from different sources like Google, Altavista, Flickr and the LabelMe dataset. The 3D versions of these images were not available as in the previous case.



Images from Torralba's dataset

Given the lack of 3D information in this dataset our classifier accuracy dropped as the confusion between different object classifiers increased.We randomly selected nearly 120 images of different classes containing the objects we have built the classifiers for.Large confusion was observed between screen-monitor and door-monitor which given the difference in their sizes could have been eliminated with 3D information.The confusion matrix for various scene detection is as shown

We also tested our classifier on objects(Monitor,Bed,Bathtub,Clock) with less confusion and took a few unrelated categories(Office,Bedroom,Bathroom).

[1]Images which could not be classified
[2]Images containing none of the objects

Table 3: Confusion Matrix

	Office	Conference Room	Hall	None
Office	64%	0%	20%	16%
Conference Room	38%	62%	0%	0%
Hall	25%	0%	65%	10%

Table 4: Confusion Matrix

	Office	Bedroom	Bathroom	None
Office	70%	22%	7%	1%
Bedroom	8%	79%	13%	0%
Bathroom	5%	6%	86%	3%

The results are an improvement over those in [1] and testing with a few 3D images available we get a better precision than in [2] also.

5 Conclusion

Here in this project we showed how the Gist features can prove to be a better priors as compared to Gabor and HOG features in object detection for indoor scene classification.The importance of using 3D information(one provided by a swiss ranger here) was reemphasized in our work as the information given by the 3D geometrical properties played an important role in improving classifier accuracy. In terms of testing using training and test data we compared our approach with others and it outperforms the alternative methods. This method still cannot classify the images in which none of the objects are present(or detected) or images which do not provide 3D info and this remains a major drawback for us.

Acknowledgments

We are very grateful to the authors P. Espinace, T. Kollar, A. Soto and N. Roy for having provided us with the code and the dataset of their paper[2].

References

[1] Ariadna Quattoni and Antonio Torralba: "Recognizing Indoor Scenes," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*2009.

[2] P. Espinace, T. Kollar, A. Soto and N. Roy: "Indoor scene recognition through object detection," *IEEE International Conference on Robotics and Automation (ICRA),* 2010.

[3] Oliva, A. and Torralba, A.: "Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope",*International Journal of Computer Vision 2001*, 42, 145-175.

[4] Ariadna Quattoni and Antoni Torralba: Indoor scene recognition database: *http://web.mit.edu/torralba/www/indoor.html.*

[5] A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin : "Context-based vision system for place and object recognition" *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October 2003.

[6] A. Torralba: "Contextual priming for object detection" *International Journal of Computer Vision*, Vol. 53(2), 169-191, 2003.

[7] A. Torralba, A. Oliva, M. Castelhano and J. M. Henderson: "Contextual Guidance of Attention in Natural scenes: The role of Global features on object search " *Psychological Review*, Vol 113(4), 766-786, Oct, 2006.

[8] A. Bosch, X. Munoz, and R. Mart, "A review: Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, 778-791, 2007.

[9] P. Agrawal, S. Nayak, S. Dube: "Scene Classification in Images" *EE Course Project Report*, IIT Kanpur, 2009.

[10] Gist Features Code: *http://people.csail.mit.edu/torralba/code/spatialenvelope/*

[11] Chang, Chih-Chung and Lin, Chih-Jen: "LIBSVM A library for support vector machines" *ACM Transactions on Intelligent Systems and Technology*, Vol 2, 1-27, 2011.

[12] A. Vailaya, A. Jain, and H. Zhang, "On image classification: city vs. landscapes," *Pattern Recog.*, vol. 31, pp. 1921-1935, 1998.

[13] Le Lu, Kentaro Toyama and Gregory D. Hager, "A Two Level Approach for Scene Recognition",*IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2005, San Diego, USA.

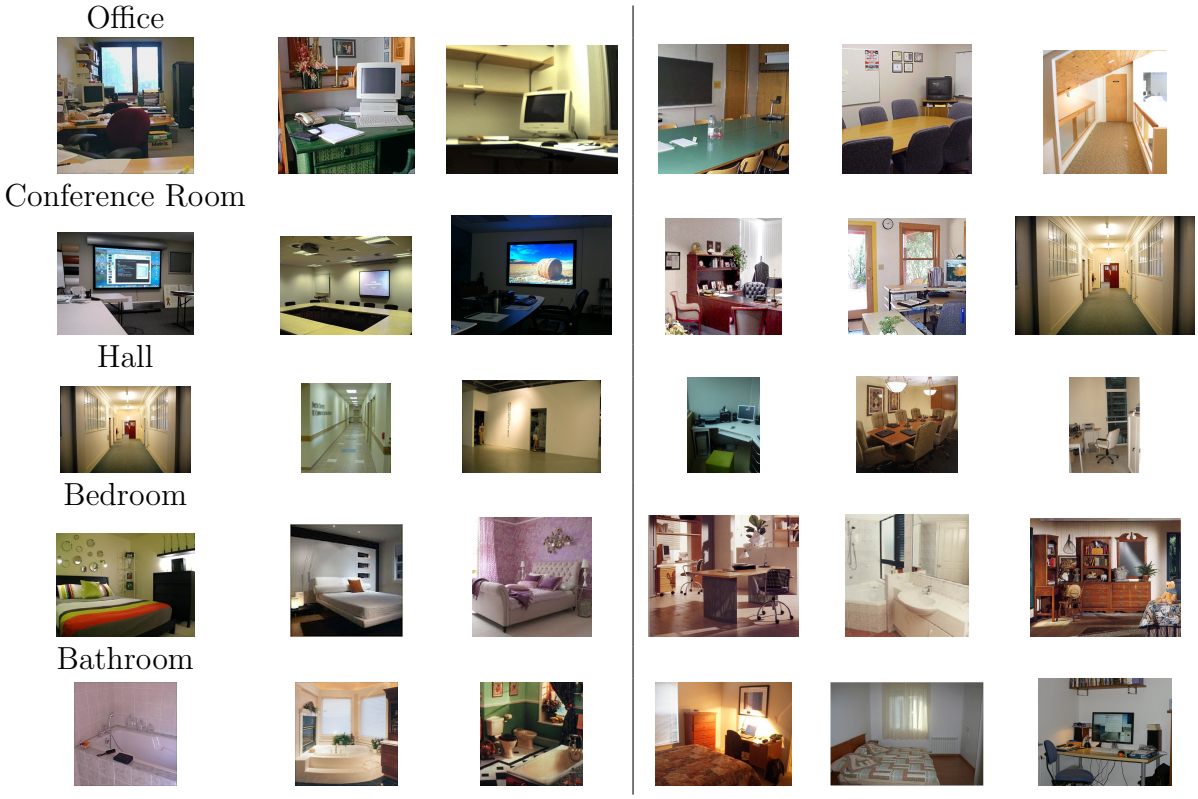
[14] Ulrich and I. Nourbakhsh, "Appearance-based place recog. for topological localization," *IEEE International Conference on Robotics and Automation*,2000.

[15] Jun Yang, Yu-Gang Jiang, Alex Hauptmann, Chong-Wah Ngo, "Evaluating Bag-of-Visual-Words Representations for Scene Classification", *Workshop on Multimedia Information Retrieval(MIR)*, in conjunction with ACM Multimedia 2007.

[16] P. Espinace, T. Kollar, N. Roy, and A. Soto. "Indoor Scene Recognition Through Object Detection Using Adaptive Object Search".*European Conference on Computer Vision(ECCV)*, Workshop on Robotics for Cognitive Tasks, 2010.

Appendix

The code for the implementation of the paper was obtained from the authors.The code is actually of an extension of this paper and hence its scene recognition module is quite different from what is described here.The code works only on a 32-bit LINUX machine and requires OpenCV and nvcc compiler.The compiler was obtained by CUDA Toolkit installation.There could still be error in compilation of openCV commands used in the code, hence you need to look for appropriate version.We have used OpenCV 2.1 here. The code is mainly divided into four modules- ExtactFeatures, BuildClassifier, ObjectDetection and SceneRecognition.There were a few bugs in the code which were ratified and it took time for us to analyze the working of these modules,their input and the form of output.Given the lack of time unfortunately we could not debug the last module and created a new one for our task.The code of our classifier is written in MATLAB and it requires the installation of Visual C++ for LIBSVM if run on Windows.



Classified images for the given scene categories. Each row corresponds to a scene category. The first three columns correspod to images which were classified correctly. The next three columns show three images from test set which were classified not to belong to the given category.