MINOR PROJECT

Understanding the role of context in object recognition

Under Supervision of:

Mr. Sarfaraz Masood

Assistant Professor

Department of Comp Engg.

Jamia Millia Islamia, New Delhi

Submitted by:

Sahil Wadhwa (12-CSS-50)

Syed Ashar Javed (12-CSS-71)

CERTIFICATE

This is to certify that this project report, "Understanding the role of context in object recognition" is a bonafide work of Sahil Wadhwa (12-CSS-50) and Syed Ashar Javed(12-CSS-71) in partial fulfillment of Graduate degree of B.Tech Computer Engg. in Jamia Millia Islamia during the year 2015-2016.

The project was carried out under my supervision. The project has not been submitted for the award of any Degree or Diploma as far as my knowledge is concerned.

Mr. Sarfaraz Masood

Assistant Professor
Department of Computer Engg.
Faculty of Engineering and Technology
Jamia Millia Islamia,New Delhi

ACKNOWLEDGEMENT

We would like to thank our mentor, **Sarfaraz Masood** (Assistant Professor, Department of Computer Engg., Jamia Millia Islamia) for providing us with the opportunity to undertake this project. His suggestions in visualizing the project and sustained interest to attain the objectives envisaged are gratefully acknowledged. His invaluable guidance and innovative ideas immensely helped us to surmount difficult situations to bring about successful completion of this project.

We also appreciate our Faculty and Jamia Millia Islamia for providing us work culture and environment that has given is the feel of a professional cooperative life. Special thanks to all our teachers, friends, families and our **Parents** for their valuable support.

Thank you one and all.

SAHIL WADHWA (12-CSS-50)

Department of Computer Engg.

Faculty of Engg. & Technology

Jamia Millia Islamia

SYED ASHAR JAVED (12-CSS-71)

Department of Computer Engg.

Faculty of Engg. & Technology

Jamia Millia Islamia

ABSTRACT

The role of context in object recognition has been extensively examined in both neuro-psychology and computer vision literature. The limitations of the recognition systems in real world setting can be partially attributed to the lack of proper incorporation of context. We try to analyze various methods using which context has been modeled in the past. In the first part of our work, we try GIST, a global image feature, with neural nets for analyzing the problem of object identification by classifying the category and super category of the objects in the image. In the second part we try out conditional random fields for modeling context information by using semantic, spatial and scale context to improve object-label agreement. A GMM modeled on feature detectors through maximum likelihood estimation is also tried for segmentation required for the CRF. The semantic context which refers to the estimation of object co-occurrence and spatial context which is refers to the estimation of objects at a particular spatial location are both taken to be pairwise attributes in images. Possible approaches for future work for current recognition systems which use deep CNNs are mentioned in the future work section.

TABLE OF CONTENT

Introduction

Real world scenes often exhibit a coherent composition of objects, both in terms of relative spatial arrangement and co-occurrence probability. This type of knowledge can be a strong cue for disambiguating object labels in the face of clutter, noise and variation in pose and illumination. Information about typical configurations of objects in a scene has been studied in psychology and
computer vision for years, in order to understand its effects in visual search, localization and recognition performance.

Bar *et al.* examined the consequences of pairwise spatial relations
between objects that typically co-occur in the same scene on human performance in recognition tasks. Their results suggested that (i) the presence of objects that have a unique interpretation improve the
recognition of ambiguous objects in the scene, and (ii) proper spatial relations among objects decreases error rates in the recognition of individual objects.

Some recently developed computational models have appealed to observation (i) in order to identify ambiguous objects in a scene. Torralba *et al.* [suggested] a low level representation of an image called the "GIST" as a contextual prior for object recognition.
Along these lines, other approaches have also considered global image features as a source of context; either by using the correlation of low level features across images that contain the object or across the category.

While the work of employs semantic context1 to improve recognition accuracy by maximizing label agreement of the objects in a scene with respect to co-occurrence, it does not place constraints on the relative locations of the objects.

As an illustration of this idea, consider the flowchart in Figure 1. An input image containing an aeroplane, trees, sky and grass (top left) is first processed through a segmentation-based object recognition engine. The recognizer outputs an ordered shortlist of possible object labels; only the best match is shown for each segment (top right). Without appealing to context, several mistakes are evident. Semantic context in the form of probable object co-occurrence allows one to correct the label of the aeroplane, but leaves the labels of the sky and grass incorrect
(bottom right). Finally, spatial context asserts that sky is more likely to appear above grass than vice versa (bottom left).

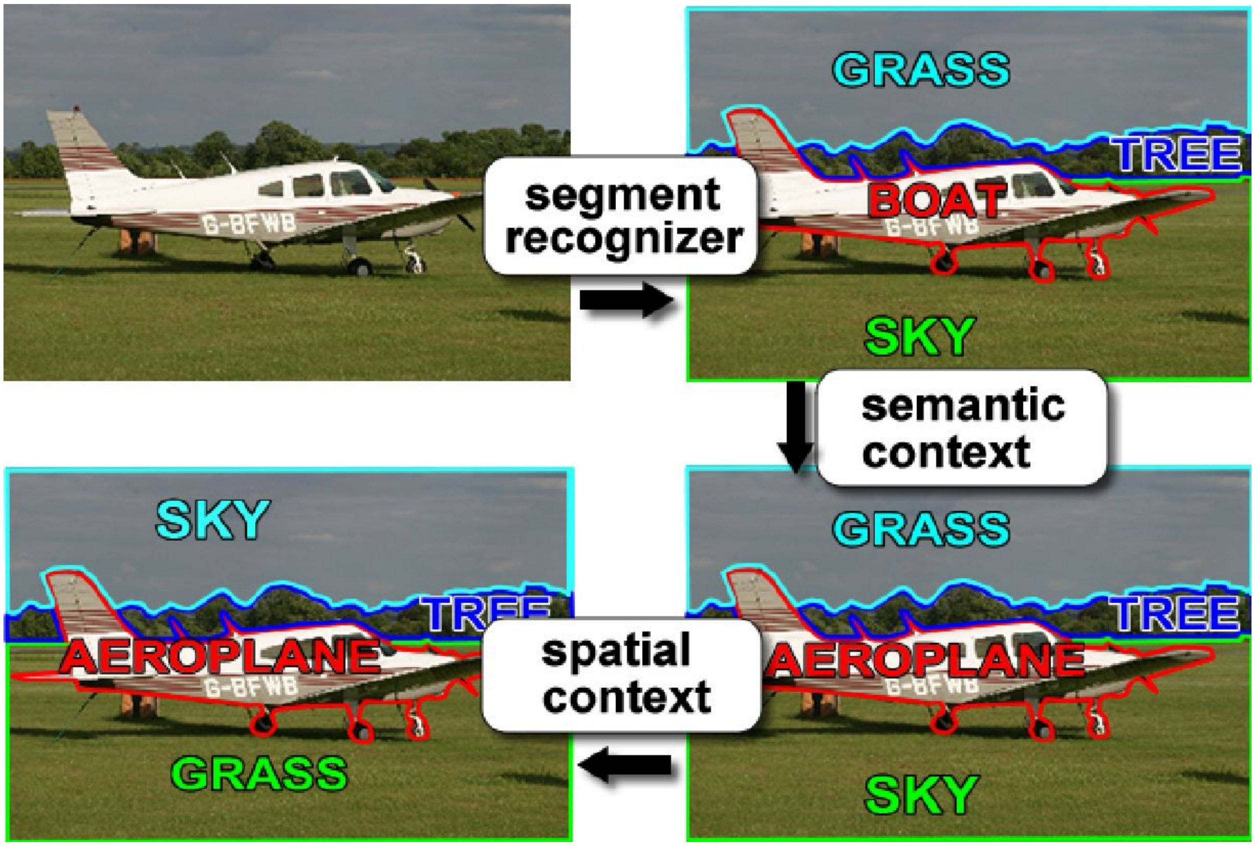


Illustration of an idealized object categorization system incorporating semantic and spatial context. First, the input image is segmented, and each segment is labeled by the recognizer. Next, semantic context is used to correct some of the labels based on object co-occurrence. Finally, spatial context is used to provide further disambiguation based on relative object locations.

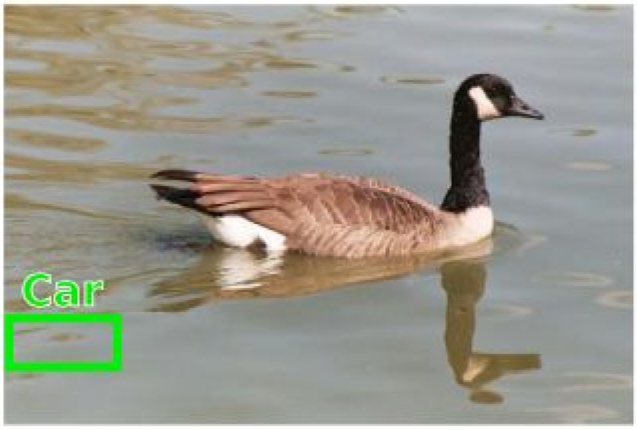
Our approach, named CoLA (for Co-occurrence, Location and Appearance), uses a conditional random field (CRF) formulation in order to maximize contextual constraints over the object labels. Co-occurrence and spatial context are learned simultaneously
from the training data in an unsupervised manner, and models for spatial relationships between objects are discovered, rather than defined a priori as in.

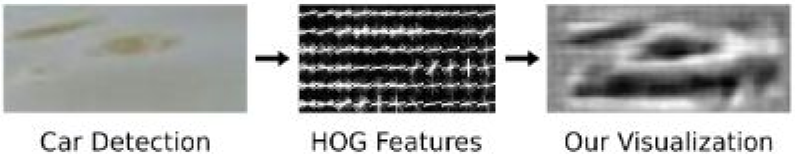
The ability of humans to recognize thousands of object categories in cluttered scenes, despite variability in pose, changes in illumination and occlusions, is one of the most surprising capabilities of visual perception, still unmatched by computer vision algorithms. Object recognition is generally posed as the problem of matching a representation of the target object with the available image features, while rejecting the background features. In typical visual-search experiments, the context of a target is a random collection of distractors that serve only to make the detection process as hard as possible. However, in the real world, the other objects in a scene are a rich source of information that can serve to help rather than hinder the recognition and detection of objects.

In the real world, objects tend to co vary with other objects and particular environments, providing a rich collection of contextual associations to be exploited by the visual system. A large body of evidence in the literature on visual cognition , computer vision and cognitive neuroscience has shown that contextual information affects the efficiency of the search and recognition of objects. There is a general consensus that objects appearing in a consistent or familiar background are detected more accurately and processed more quickly than objects appearing in an inconsistent scene.Hence, it is quite reasonable to include context in implementing a model to improve classification of objects in an image.

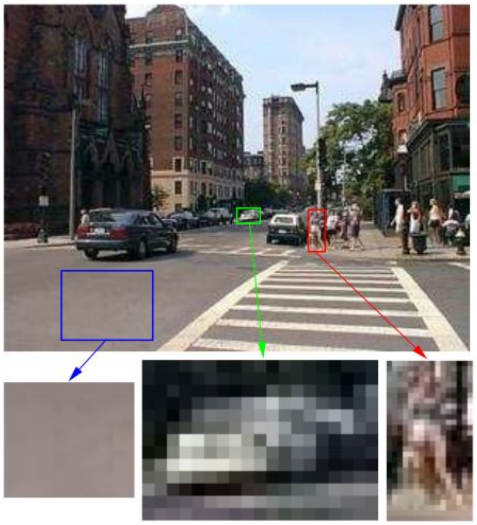
Motivation

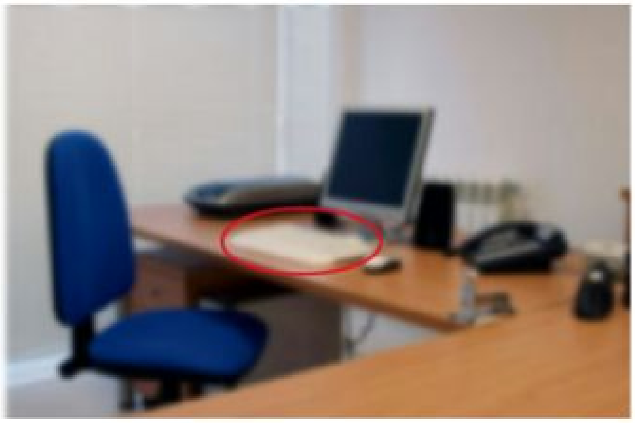
Object context has been studied extensively for aiding object recognition. Studies in neurophysiology have shown that our own visual system processes low frequency information relatively early during recognition. This opens up the scope of contextual priming for obtaining better results and also for reducing space of plausible object hypothesis. Traditionally, object recognition has been done using feature detectors used in various settings like the bag of words model, the part-based model or other discriminative models. But all these methods suffer from certain deficiencies. Objects are difficult to recognize in a feature space as they only club local feature points which mostly are found to not be reversible and one feature representation can correspond to multiple objects, thus causing false positives[1]. The figure below shows the evident misrepresentation of one of the most commonly used feature descriptor, HoG. If contextual priors are added to the recognition system, such false positives can be avoided.





There is also a problem in locally identifying objects as objects often tend to be of low resolution when taken in isolation. Therefore the feature representation lacks quality. The figures below show the isolated objects in real world images and highlights the contextual aid our own visual system gets when identifying such objects with our eyes.





Review of past work

In a well-known series of experiments, Biederman [2] examined performance in a detection task as a function of the number of 'violations' between a target object and the scene in which it was (briefly) presented. The violations varied across properties like spatial and semantic relations and the detection errors were analysed, on the basis of which they suggested five classes of context needed for recognition. These are often simplified into three categories [3], namely semantic context, spatial context and scale context.

[4] models semantic context using Google Sets by modeling the co-occurrence information that relates the joint presence of pair of objects. [5] uses low level descriptors like color and texture over a wide receptive field which are shown to be capable of identifying bicycles, cars and pedestrians using context alone. It also shows how contextual information is marginally useful if the object is unambiguously visible.

In [6] and [7], Bar et al examines the pairwise spatial context on human performance in recognition tasks, concluding that object recognition error rate is decreased by context, especially for objects having a unique representation. [8] uses inter-pixel statistics to categorize objects.

[9] proposes a 3D scene modeling approach using relationship between objects, surfaces and viewpoints to provide a framework which is then used by local object detection. Object co-occurrence and location and appearance are modeled in [10] and [11] on which our work is primarily based.

Apart from context modeling methods, the nature of context is also studies in various papers. The notion of an image-level 'gist' is analysed in [12]and [13]. This global descriptor has been found to effective in scene classification. We also use this feature to analyse its effectiveness on a complex dataset like MS COCO[14].

GIST Descriptor

GIST is a low dimensional representation of a scene[13]. This is termed as a spatial envelope which is built using five different global perceptual features namely, naturalness, openness, roughness, expansion, ruggedness. It has been used for image similarity and scene classification tasks.

It is calculated as follows (as done in the original paper):

1. Convolve 32 Gabor filters at 4 scales & 8 orientations, thereby producing 32 feature maps.
2. Divide each map into a 4*4 grid and average the values inside each block
3. Concatenate the 16 averaged values for 32 maps, which will give a 16*32 dimensional descriptor

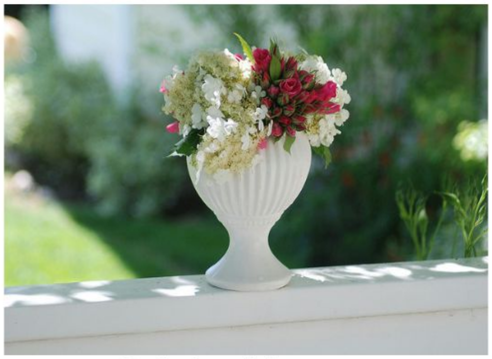
GIST is one of the popular feature detectors used for scene classification and image retrieval. Typically, these features are modeled using SVMs or K-means for classification tasks.

GIST enables the classifier to represent scenes through a global prior which can later be used for object priming. Our eyes recognize cars in bad quality images because of the presence of a road. But often the presence of a road is dependent on the presence of a car and other related objects like pedestrians. To solve this chicken and egg problem, GIST can be used to categorize an image into a super category which can then be used for further recognition in a reduced class space. Though the GIST alone is not good enough for object recognition, we try it as a method for elimination of class labels which show low probability of occurrence.

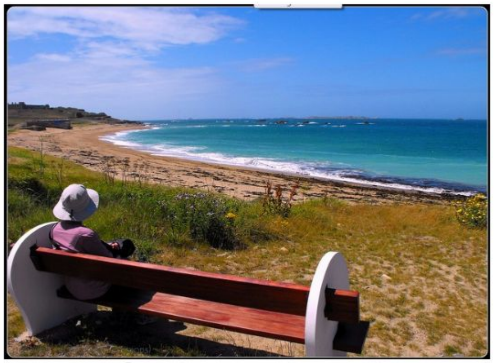
This task could be perhaps better achieved with a bag of words model using various other descriptors but we limit ourself to the study of global features and the use of only contextual information for analysing how they affect recognition.

We compute the features at 4 scales and 8 orientations and use a 8*8 grid with both 8 and 16 pre filtering blocks. We use a multi-layered perceptron with the GIST features as input to classify objects. This procedure will be described in more detail in the next section.

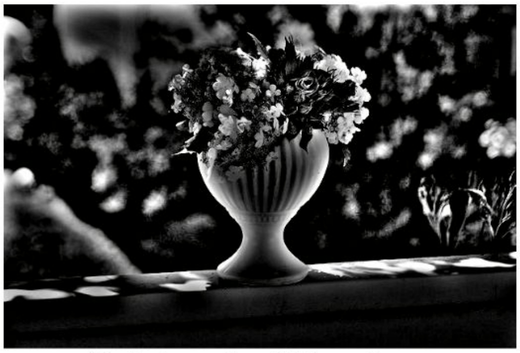
The following figures show the result of a GIST descriptor applied to two sample images. Notice that as the number of pre filtering blocks are increased, the level of detail for the image improves. This may seem as a good thing for a feature detector, but for classification tasks, much of the detail is lost and the neural network may not be able to generalize well with the sparse information present. Therefore we use 8-block filters during classification.



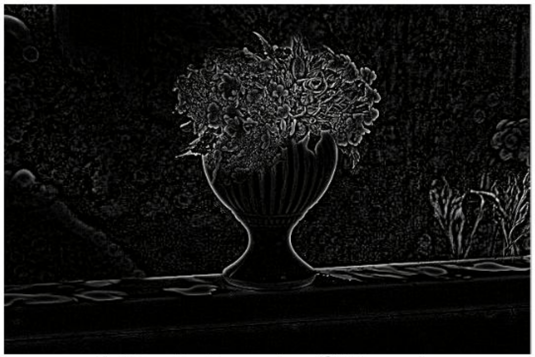
Original Image



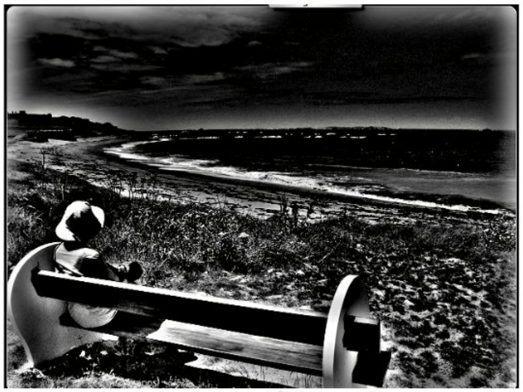
Original Image



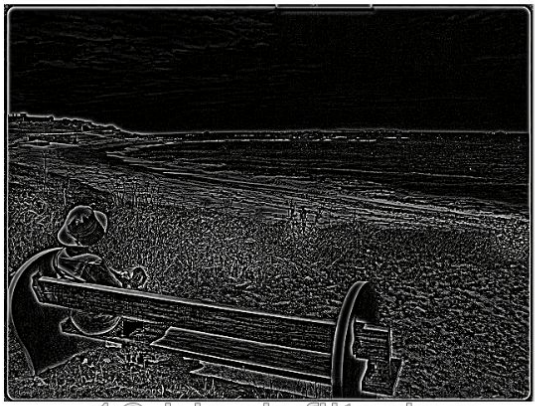
8 block filtering



16 block filtering



8 block filtering



16 block filtering

Training a neural network with GIST descriptor

Since the effectiveness of deep CNNs in object recognition has been established, it is interesting to see how black and white GIST images react to the loss in information and how the error rate of the recognition system is affected. Also, it is expected that objects in similar scenes may be mislabeled as the prime features behind GIST are naturalness, openness, roughness, expansion, ruggedness, all of which determine the nature of scene. Finally, the experiment done on super categories instead of categories may show better results as the objects are relatively abstracted during training on super categories.

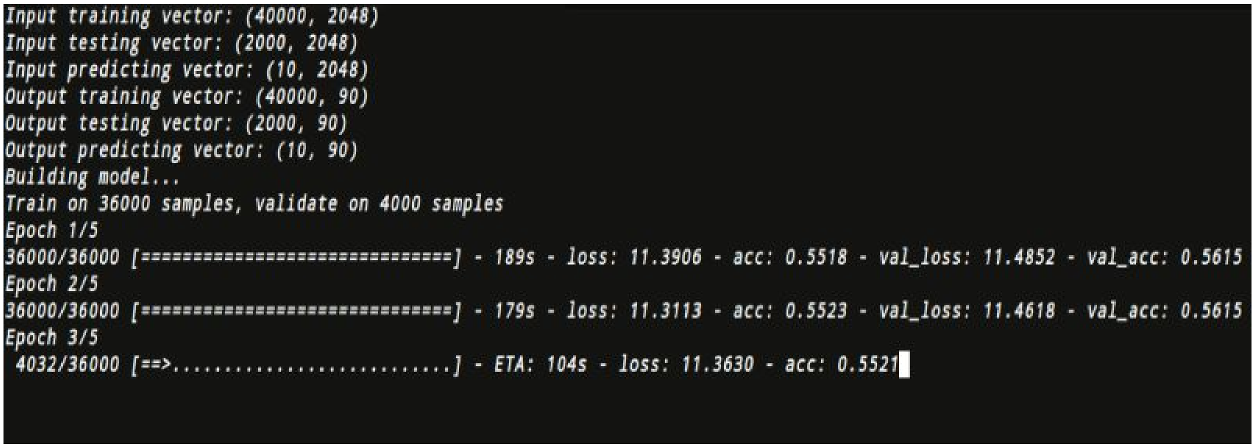
We use MS COCO dataset for training and train it on both object category and super category. Spatial envelope is calculated using 32 feature maps (8 orientations and 4 scales) with averaging across 8*8 grids. Matlab is used for this. The multi layered perceptron is constructed on a Theano based wrapper called Keras. The following table shows the architecture information. ‘// denotes any difference in network for category/supercategory classification.

Input Vector	2048 dimensions
Output vector	90 dimensions/12 dimensions
# of hidden layers	1-3
Error measure	Categorical cross entropy
Optimizer	ADAM
Regularizer	Dropout
Activation function	ReLU units, softmax output layer

60,000 images were used for training with a validation split of 0.2 and a batch size ranging from 32-128. Testing is performed on 5,000 images.

Results

The following figure shows a sample training of a network with different setting.



```
Input training vector: (40000, 2048)
Input testing vector: (2000, 2048)
Input predicting vector: (10, 2048)
Output training vector: (40000, 90)
Output testing vector: (2000, 90)
Output predicting vector: (10, 90)
Building model...
Train on 36000 samples, validate on 4000 samples
Epoch 1/5
36000/36000 [==============================] - 189s - loss: 11.3906 - acc: 0.5518 - val_loss: 11.4852 - val_acc: 0.5615
Epoch 2/5
36000/36000 [==============================] - 179s - loss: 11.3113 - acc: 0.5523 - val_loss: 11.4618 - val_acc: 0.5615
Epoch 3/5
 4032/36000 [==>...........................] - ETA: 104s - loss: 11.3630 - acc: 0.5521
```

The accuracy results are summarized in the table below

Classification label	% accuracy
Object category	59
Object super-category	64

The activations of the final layer are observed to be biased to highly occurring labels like the class ‘person’ which appears in a lot of images.

More importantly, using GIST to eliminate classes, thus reducing class space yields better results. The table summarizes the results of the training.

# of eliminated classes	% loss in accuracy
40	14-17
50	18-20

The mis-classified classes are observed to usually have similar GIST representation
In conclusion, complex datasets like MS COCO having multiple objects in an image in real world scenarios may have difficulty in generalizing well for object classification using GIST alone.

This process can be further done for elimination of more classes by taking top-10, top-20 softmax outputs, but we decided to stop here and move on to the next section because the accuracy of this method was already limiting.

Segmentation and Gaussian Mixture Models (GMM)

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the subpopulation to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without subpopulation identity information.

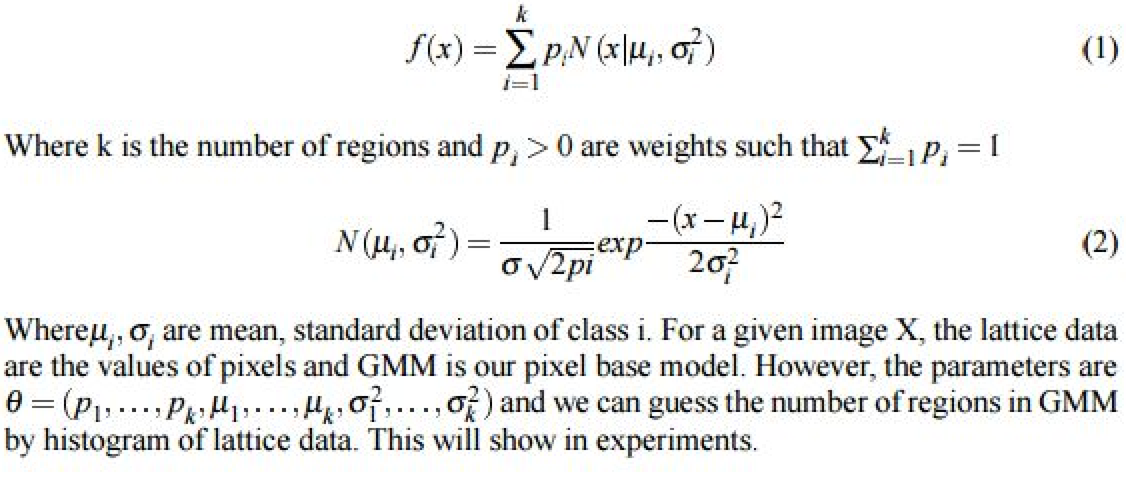
Some ways of implementing mixture models involve steps that attribute postulated subpopulation-identities to individual observations (or weights towards such sub-populations), in which case these can be regarded as types of unsupervised learning or clustering procedures. However not all inference procedures involve such steps.

Mixture models should not be confused with models for compositional data, i.e., data whose components are constrained to sum to a constant value (1, 100%, etc.). However, compositional models can be thought of as mixture models, where members of the population are sampled at random. Conversely, mixture models can be thought of as compositional models, where the total size of the population has been normalized to 1.

Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model.

Image is a matrix which each element is a pixel. The value of the pixel is a number that shows intensity or color of the image. Let X is a random variable that takes these values. For a probability model determination, we can suppose to have mixture of Gaussian distribution as the following form



$$f(x) = \sum_{i=1}^{k} p_i N(x|\mu_i, \sigma_i^2) \qquad (1)$$

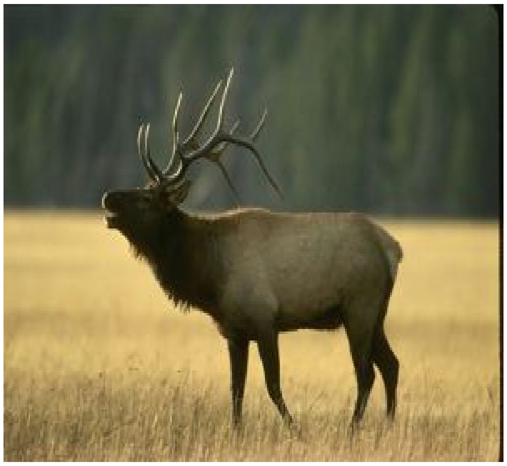
Where k is the number of regions and $p_i > 0$ are weights such that $\sum_{i=1}^{k} p_i = 1$

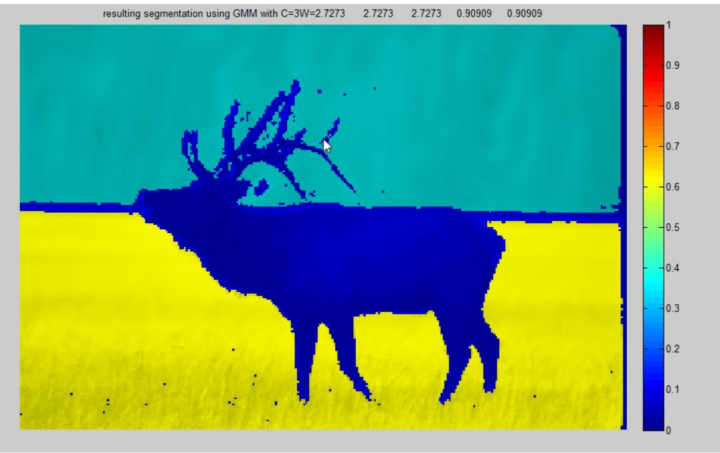
$$N(\mu_i, \sigma_i^2) = \frac{1}{\sigma\sqrt{2pi}} exp \frac{-(x-\mu_i)^2}{2\sigma_i^2} \qquad (2)$$

Whereμ_i, σ_i are mean, standard deviation of class i. For a given image X, the lattice data are the values of pixels and GMM is our pixel base model. However, the parameters are $\theta = (p_1, \ldots, p_k, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2)$ and we can guess the number of regions in GMM by histogram of lattice data. This will show in experiments.

It is also important to note that because the component Gaussian are acting together to model the overall feature density, full covariance matrices are not necessary even if the features are not statistically independent. The linear combination of diagonal covariance basis Gaussians is capable of modeling the correlations between feature vector elements. The effect of using a set of M full covariance matrix Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

GMMs are often used in biometric systems, most notably in speaker recognition systems, due to their capability of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities. The classical uni-modal Gaussian model represents feature distributions by a position (mean vector) and an elliptic shape (covariance matrix) and a vector quantizer (VQ) or nearest neighbor model represents a distribution by a discrete set of characteristic temp.




resulting segmentation using GMM with C=3W=2.7273 2.7273 2.7273 0.90909 0.90909
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

In the above two figures, the second image is the G.M.M output of the first image. The second image contains three classes namely deer, sky and grass. If we increase the number of classes for the first image then the results become more complex.

We have used 7 Gaussians per class in the three-dimensional RGB space.The likelihoods for each pixel are averaged across the segments to obtain a C length vector.The two C length vectors are concatenated and passed through a multi-layer perceptron neural network with C outputs. We used 20 hidden layer nodes in our experiments with a sigmoid transfer function.

Conditional Random Fields

Conditional random fields (CRFs) are a class of statistical modelling method often applied in pattern recognition and machine learning, where they are used for structured prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF popular in natural language processing predicts sequences of labels for sequences of input samples.

CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences and in computer vision. Specifically, CRFs find applications in shallow parsing, named entity recognition, gene finding and peptide critical functional region finding, among other tasks, being an alternative to the related hidden Markov models (HMMs). In computer vision, CRFs are often used for object recognition and image segmentation.

We used a CRF to model the probability of class labels c_1,c_2,...c_k given segments S_1,S_2,...S_k such that the labels are in agreement with one another

$$p(c_1 \ldots c_k | S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i | S_i)}{Z(\phi_0, \ldots \phi_r, S_1 \ldots S_k)},$$

$$\text{with } B(c_1 \ldots c_k) = \exp\left(\sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r(c_i, c_j) \right),$$

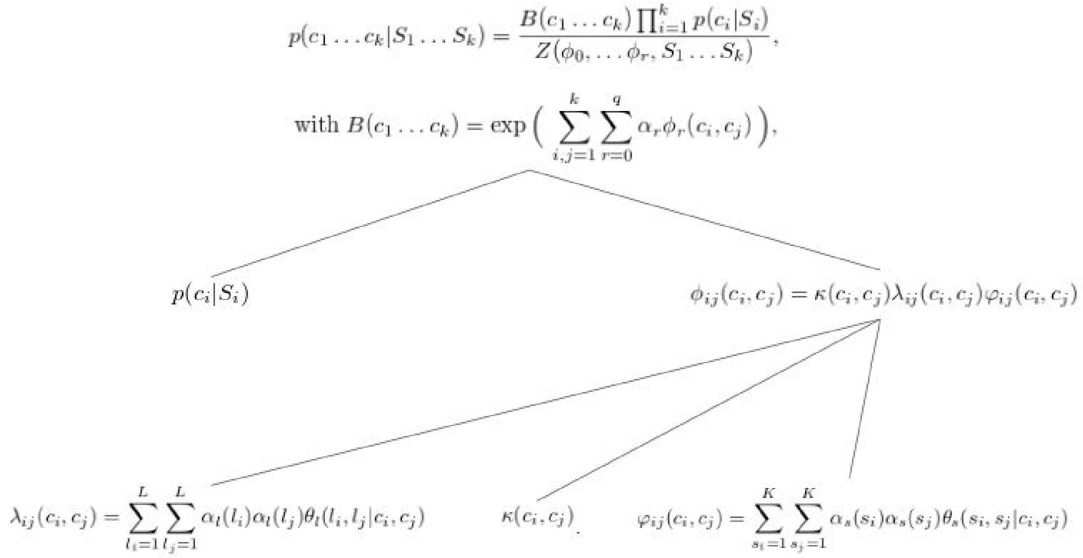
where Z(.) is the partition function, α_r a parameter estimated from training data and q is the number of pairwise spatial relations. A CRF does not require feature independence assumption and can model the conditional probability.

CRF Model

Three terms need to be calculated. First is the class conditional probability, the second is the contextual probability and the third is the partition function

P(C_i|S_i) will be calculated using a CNN and a bag of features model over object instances. Segmentation masking is done and feature extraction over the masks is next

Location and Scale function will be trained on the MLE counts.



$$p(c_1 \dots c_k | S_1 \dots S_k) = \frac{B(c_1 \dots c_k) \prod_{i=1}^{k} p(c_i|S_i)}{Z(\phi_0, \dots \phi_r, S_1 \dots S_k)},$$

$$\text{with } B(c_1 \dots c_k) = \exp\Big(\sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r(c_i, c_j) \Big),$$

$$p(c_i|S_i)$$

$$\phi_{ij}(c_i, c_j) = \kappa(c_i, c_j) \lambda_{ij}(c_i, c_j) \varphi_{ij}(c_i, c_j)$$

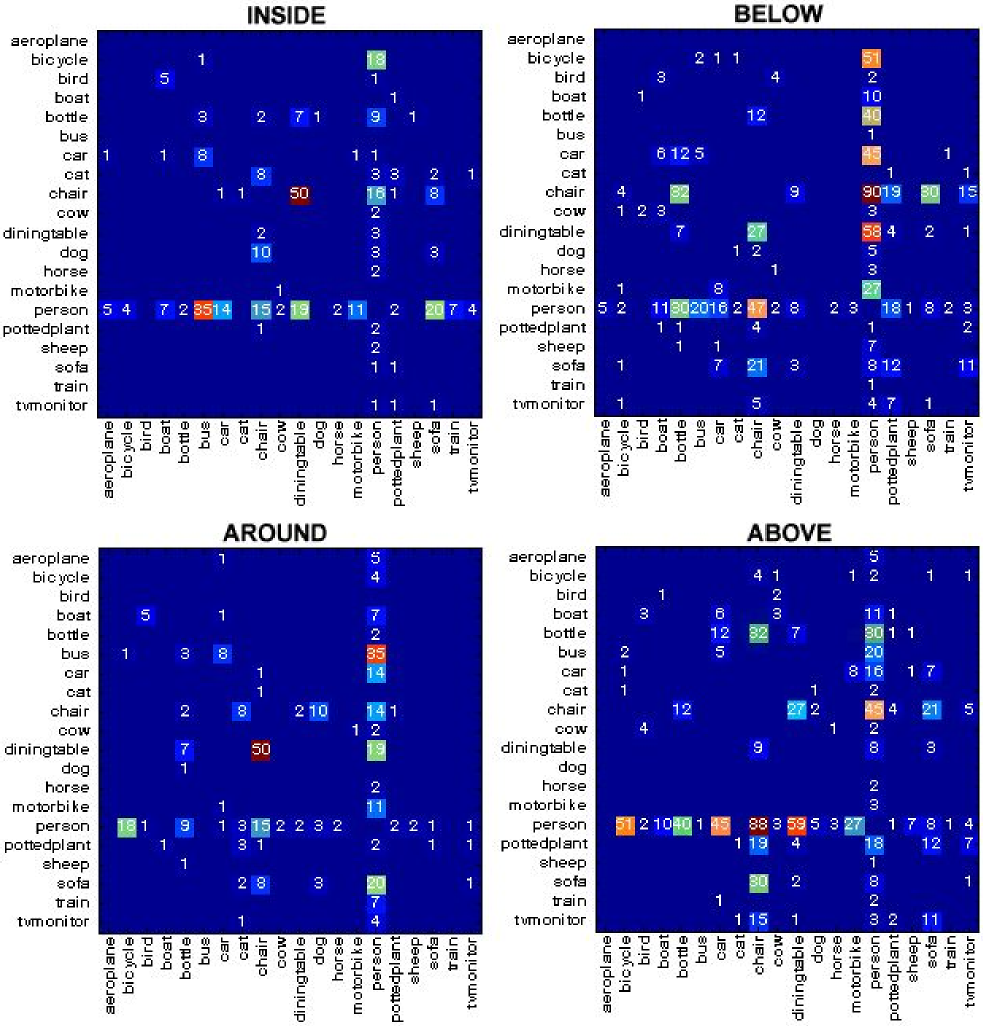
$$\lambda_{ij}(c_i, c_j) = \sum_{l_i=1}^{L} \sum_{l_j=1}^{L} \alpha_l(l_i) \alpha_l(l_j) \theta_l(l_i, l_j | c_i, c_j)$$

$$\kappa(c_i, c_j)$$

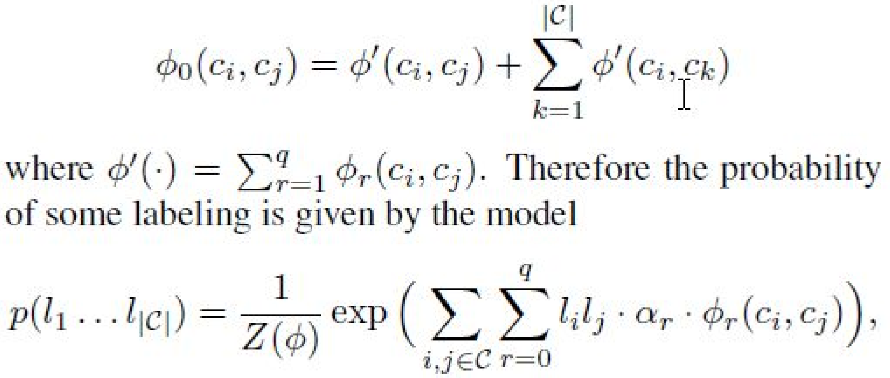
$$\varphi_{ij}(c_i, c_j) = \sum_{s_i=1}^{K} \sum_{s_j=1}^{K} \alpha_s(s_i) \alpha_s(s_j) \theta_s(s_i, s_j | c_i, c_j)$$

Co-occurrence Counts

While the occurrence of category labels are captured by the spatial context matrices above, the appearance frequency – a parameter required for the CRF – is not captured explicitly, since these matrices are hollow. Using the existing spatial context matrices, object appearance frequency can be computed as row sums of all for matrices. Finally, the sum of all four matrices, including the row sums, will result in a marginal (i.e., without regard for location) co-occurrence matrix.



An entry (i, j) in the semantic context matrix counts the number of times an object with label i appears in a training image with an object with label j. The diagonal entries correspond to the frequency of the object in the training set:



$$\phi_0(c_i, c_j) = \phi'(c_i, c_j) + \sum_{k=1}^{|\mathcal{C}|} \phi'(c_i, c_k)$$

where $\phi'(\cdot) = \sum_{r=1}^{q} \phi_r(c_i, c_j)$. Therefore the probability of some labeling is given by the model

$$p(l_1 \ldots l_{|\mathcal{C}|}) = \frac{1}{Z(\phi)} \exp\Big(\sum_{i,j \in \mathcal{C}} \sum_{r=0}^{q} l_i l_j \cdot \alpha_r \cdot \phi_r(c_i, c_j) \Big),$$

with l_i indicating the presence or absence of label i.We wish to find a $\phi(\cdot)$ that maximizes the log likelihood of the observed label co-occurrences. Every time the partition function is estimated, 40,000 points are sampled from the proposal distribution.Therefore we approximate the partition function using Monte Carlo integration.Importance sampling is used where the proposal distribution assumes that the label probabilities are independent with probability equal to their observed frequency.

Future Work

The CRF implementation is done halfway and the remaining two terms need to be calculated for the class probability achieved by the object-label agreement. Also, recent research in deep CNNs have been successful in solving ImageNet with better-than-human accuracy, but more complex datasets still need to be worked upon. Deep CNNs do not model any feature explicitly, but have implicit feature representations a various layers in their feature maps. These layers can be analysed for better understanding to see whether it uses any implicit representation of contextual features and if not, then can incorporating any help in identification.

References

[1] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, Antonio Torralba, HOGles: Visualizing Object Detection Features

[2] Biederman I, Mezzanote R, Rabinovitz J, Scene perception: detecting and judging objects undergoing relational violations.

[3] C. Galleguillos, S. Belongie, Context based object categorization: A critical survey, Comput. Vis. Image Understand.

[4] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context

[5] L. Wolf, S. Bileschi, A critical view of context

[6] M. Bar, Visual objects in context

[7] M. Bar, S. Ullman, Spatial context in recognition

[8] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling appearance, shape and context

[9] Derek Hoiem, Alexei A. Efros, Martial Hebert, Putting Objects in Perspective

[10] C. Galleguillos, A. Rabinovich, and S. Belongie, Object categorization using co-occurrence, location and appearance

[11] Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, Objects in context

[12] Aude Oliva, Antonio Torralba, Building the gist of a scene: the role of global image features in recognition

[13] Aude Oliva, Antonio Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope

[14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., and Zitnick, C. L. ,Microsoft COCO: Common Objects in Context