Patch-based Object Recognition Using Discriminatively Trained Gaussian Mixtures

Andre Hegerath, Thomas Deselaers, and Hermann Ney
Human Language Technology and Pattern Recognition Group,
RWTH Aachen University – D-52056 Aachen, Germany
{hegerath, deselaers, ney}@i6.informatik.rwth-aachen.de

Abstract

We present an approach using Gaussian mixture models for part-based object recognition where spatial relationships of the parts are explicitly modeled and parameters of the generative model are tuned discriminatively. These extensions lead to great improvements of the classification accuracy. Furthermore we evaluate several improvements over our baseline system which incrementally improve the obtained results which compare favorable well to other published results for the three Caltech tasks and the PASCAL evaluation 05 tasks.

1 Introduction

Recently, part-based models in general and patch-based models in particular have gained an enormous amount of interest in the computer vision community [13, 6, 16]. This approach offers some immediate advantages like translation invariance and robustness against occlusion because the parts can be modeled more or less independently. Thus, an object that is partly occluded can be classifed correctly as long as the visible parts can be recognized.

In this paper, we present an approach that uses Gaussian mixture densities to model and recognize objects in unconstrained images. The parameters of these generative models are refined discriminatively. In addition to the appearance of the extracted patches, we consider absolute and relative positions of the patches.

Previous work on part-based object recognition can be divided wrt. the type of modelling (generative vs. discriminative) and whether spatial information is used or not. Although recently, some other groups proposed to take advantage from a mix of generative and discriminative methods [17, 15, 7] most approaches in this area are either generative like the star model [6] or discriminative [16, 9, 13]. In [16] a comparison of generative and discriminative models is given and the conclusion is drawn that neither approach is sufficient for large scale object recognition. Our approach directly addresses this point as we start from a probabilistic generative model that is refined discriminatively.

While in the star model [6] and in the constellation model [5] spatial relationships are explicitely modelled, other successful approaches disregard them [13, 1].

In contrast to these approaches, our approach consists of a generative model where some of the parameters are refined discriminatively. An advantage over other mixed models is that only the training step is modified but the classification itself is done using a

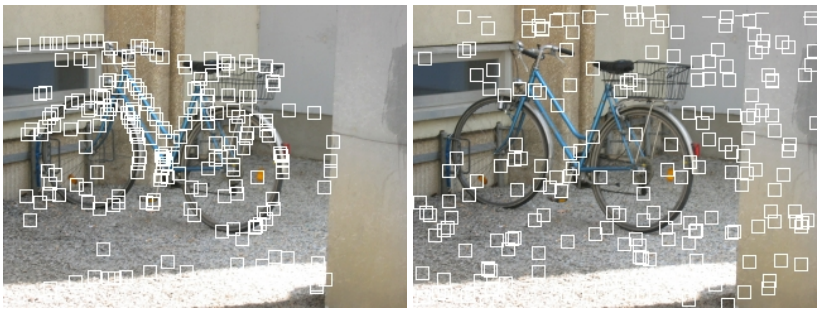

Figure 1: Patches extracted around interest points (left) and randomly chosen points (right).

generative model. Thus it is an effective combination of the advantages of both worlds: *Generative approaches* have the advantages that they can handle missing or partially labelled data, that new classes can be added incrementally, and that compositionality can easily be handled. Discriminative methods directly model the decision function and do not consider data that may be irrelevant for the classification decision and thus lead to better predictive performance in many cases and they are usually very fast in predicting the class label for an unseen observation [16].

Spatial information (absolute positions as well as relative positions) can be incorporated into the model and improve the recognition performance. This approach obtains very good results for the rather easy Caltech tasks and for the more difficult PASCAL Visual Object Classes Challenge Tasks. In particular we have the best result published so far for the considered Caltech tasks and very competitive results for the PASCAL tasks.

2 Feature Extraction

We use square image patches of different sizes as features that are extracted from the images around interest points. We do not use other descriptors as e.g. the SIFT descriptor [12] as we want to focus on the model and not on the features. For dimensionality reduction the patches are PCA transformed keeping 40 coefficients. The feature extraction points are obtained from *a)* a wavelet-based interest point detector proposed by Loupias et al. [11] and *b)* randomly chosen points in the images. The interest points capture regions of the image where "*something happens*" and the random points also capture homogeneous regions. Figure 1 shows an image with both types of extraction points marked.

At these extraction points, we extract patches of different sizes (7×7, 11×11, 21×21, and 31×31 pixels). This multi-scale extraction allows us to represent object parts of different sizes and it allows us to handle scale changes to some degree. The same approach was used in [2] to obtain partial scale invariance.

Figure 2 shows the process of deriving a feature vector from an image: first, the extraction point of the patch is determined. Then, the patch is extracted, resulting in an $n\times n$ subimage. Its pixels form a vector of n^2 gray-level components which is reduced to 40 coefficients by PCA transformation.

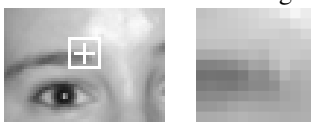

Figure 2: Deriving feature vectors from images.

3 Gaussian Mixture Models

Gaussian mixture models are a *generative* model: for each object class a class-dependent mixture $p(x \mid k)$ is used. To decide which object is depicted in an image, Bayes' decision rule is used:

$$\begin{aligned} r(\{x_1^L\}) &= \operatorname*{argmax}_k \left\{ p(k|\{x_1^L\}) \right\} = \operatorname*{argmax}_k \left\{ p(k) \cdot p(\{x_1^L\}|k) \right\} \\ &= \operatorname*{argmax}_k \left\{ p(k) \cdot \prod_{l=1}^{L} p(x_l|k) \right\}, \end{aligned} \tag{1}$$

where $\{x_1^L\}$ denotes the set of patches $x_1,\ldots,x_L$ extracted from image X. An alternative to the above given decision rule is to classify each patch individually and combine the decisions e.g. using sum rule to one classification decision. Then, we can consider the posterior probability of X, $p(k|\{x_1^L\})$, to be proportional to the sum of the posterior probabilities of the individual patches, $p(k|x_l)$:

$$\begin{aligned} p(k|\{x_1^L\}) &\propto \frac{1}{L} \sum_{l=1}^{L} p(k|x_l) \\ &= \frac{1}{L} \sum_{l=1}^{L} \frac{p(k) \cdot p(x_l|k)}{\sum_{k'} p(k') \cdot p(x_l|k')} \end{aligned} \tag{2}$$

In both cases, the feature vectors are assumed to be independent.

For *detection* tasks, where it has to be decided whether an object of interest is contained in an image or not, a different decision rule has become quite common, which allows us to calculate the *equal error rate (EER)*. Here, an image is *accepted*, i.e. classified to contain the object of interest, if the probability for the "positive" class exceeds a certain threshold probability p_T, otherwise it is *rejected*:

$$r(\{x_1^L\}) = \begin{cases} 1 & \text{if } p(k=1|\{x_1^L\}) \geq p_T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Here, "1" denotes "acceptance" and "0" denotes "rejection". The threshold probability p_T is set such that the false positive rate equals the false negative rate.

In our model, the class-dependent distributions $p(x_l|k)$ are modeled by Gaussian mixture densities and we distinguish two cases:

Untied mixtures (class-dependent) :

$$p(x_l|k) = \sum_{c=1}^{C_k} p(c|k) \cdot p(x_l|c,k) = \sum_{c=1}^{C_k} p(c|k) \cdot \mathcal{N}(x_l|\mu_{ck}, \Sigma_{ck}) \tag{4}$$

Tied mixtures (class-independent) :

$$p(x_l|k) = \sum_{c=1}^{C} p(c|k) \cdot p(x_l|c) = \sum_{c=1}^{C} p(c|k) \cdot \mathcal{N}(x_l|\mu_c, \Sigma_c) \tag{5}$$

In the untied case, further classes can be added easily by estimating the corresponding class-dependent distributions for these classes without the need to reestimate the distributions for all other classes.

The class-dependent distributions are trained by maximum likelihood training using a top-down EM clustering approach known as the Linde-Buzo-Gray algorithm [10].

Discriminative Training. To improve recognition performance, the parameters of the generative model can be refined using discriminative methods. In contrast to the common maximum likelihood approach, where the class representation is optimized, here we are interested in optimizing the discrimination performance. We propose to tune the parameters by maximizing the posterior probability (also known as maximum mutual information (MMI) criterion) instead of maximizing the likelihood. Following this approach it is possible to refine all parameters of the Gaussian mixture models, but here we only refine the mixture weights $p(c \mid k)$. This approach can be compared to the method presented in [1] where maximum entropy training is used to train the discriminativeness of clusters of patches.

Combining Eq. (2) with Eq. (4) or Eq. (5), the dependency of the posterior probability $p(k|\{x_1^L\})$ on the mixture weights $p(c|k)$ becomes obvious:

$$\begin{aligned} p(k|\{x_1^L\}) &= \frac{1}{L}\sum_{l=1}^{L} p(k|x_l) = \frac{1}{L}\sum_{l=1}^{L} \frac{p(k)\cdot p(x_l|k)}{\sum_{k'} p(k')\cdot p(x_l|k')} \\ &= \frac{1}{L}\sum_{l=1}^{L} \frac{p(k)\sum_{c=1}^{C_k} p(c|k)\cdot p(x_l|c,k)}{\sum_{k'} p(k')\sum_{c=1}^{C_{k'}} p(c|k')\cdot p(x_l|c,k')}. \end{aligned}$$

We define an auxiliary function F of the mixture weights as the sum of the logarithmic posterior probabilities for the correct class k_n with the set of feature vectors $\{x_1^L\}_n$ over all N training images:

$$F\left(p(c|k)\right) = \sum_{n=1}^{N} \log p(k_n|\{x_1^L\}_n). \tag{6}$$

Maximizing F is known as *MMI* training: The derivative of F wrt. the mixture weights $p(c|k)$ is calculated and the mixture weights are iteratively updated using gradient descent.

$$p(c|k) \leftarrow p(c|k) - \varepsilon \cdot \frac{\partial F\left(p(c|k)\right)}{\partial p(c|k)} \tag{7}$$

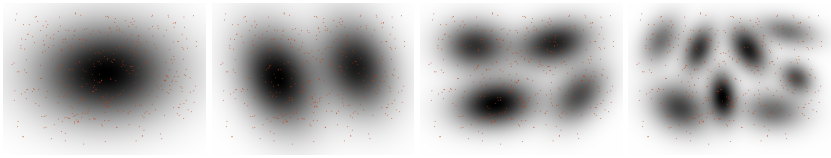

Figure 3: 1, 2, 4, and 8 probability distributions for patches of a given density.

If ε is chosen small enough, convergence towards a local maximum is guaranteed. The $p(c|k)$ are initialized using the Maximum-Likelihood estimates, i.e. the relative cluster sizes are chosen. This discriminative updating of the mixture weights gives higher weights to densities with higher discriminatory relevance and lower weights to densities less relevant for the classification.

4 Spatial Information

While many approaches to object recognition ([1,3,13]) ignore spatial relations between the parts completely, we believe that the incorporation of the position information allows for substantial improvement in recognition. We present two extensions to the Gaussian mixture model incorporating relative and absolute patch positions respectively.

4.1 Absolute Patch Positions

Let x_l be the feature vector accounting for the appearance of the l-th patch of an image as before. Further, let y_l be the position of the patch. Let x'_l be the combination of both vectors, we propose to model the emission probabilities $p(x'_l|c,k)$ of the mixtures by products of the emission probabilities for appearance and position, $p(x_l|c,k) \cdot p(y_l|c,k)$, such that the class-dependent distributions $p(x'_l|k)$ becomes:

$$p(x'_l|k) = \sum_{c=1}^{C_k} p(c|k)\ p(x_l|c,k)\ p(y_l|c,k) = \sum_{c=1}^{C_k} p(c|k)\ \mathcal{N}(x_l|\mu_{xck},\Sigma_{xck})\ \mathcal{N}(y_l|\mu_{yck},\Sigma_{yck})$$

for untied mixtures and the analogous form for tied mixtures. The distribution for $p(y_l|c,k)$ is estimated from the positions of all patches from which c is estimated. To account for object parts occurring more than once in an object, multiple position probability distributions can be estimated per cluster, thus the emission probability of the patch position becomes a mixture density itself:

$$p(y_l|c,k) = \sum_{i=1}^{I_{ck}} p(i|c,k)\ p(y_l|i,c,k) = \sum_{i=1}^{I_{ck}} p(i|c,k)\ \mathcal{N}(y_l|\mu_{yick},\Sigma_{yick}).$$

A visualization of different numbers of probability distributions for patches of a given density is shown in Figure 3, where dark areas account for a high patch position probability and light areas for a low probability.

4.2 Relative Patch Positions

A problem in using the absolute position of patches is the loss of invariance wrt. translation: if an object appears in all training images at a fixed position, it cannot be recognized

at different positions anymore. To overcome this problem, we propose using *relative* positions instead of absolute positions. For objects of the same scale, the relative positions of their parts remain constant regardless of the absolute position of the object. In the following, we extend the proposed model to incorporate relative positions.

Let x_l be the l-th patch of the image X, let $\{z_1^{L-1}\}_l$ denote the set of position differences to all other patches in X: $\{z_1^L\}_l = \{z_{1,l}, \ldots, z_{l-1,l}, z_{l+1,l}, \ldots, z_{L,l}\}$, where $z_{\lambda,l}$ denotes the position difference between the l-th and the λ-th patch of X: $z_{\lambda,l} = y_l - y_\lambda$.

Now, let x_l'' be our new feature vector consisting of the appearance x_l and the relative position information $\{z_1^L\}_l$ of the l-patch. The decision rule is then reformulated as

$$r(\{x''^L_1\}) = \underset{k}{\operatorname{argmax}} \left\{ p(k) \prod_{l=1}^{L} \sum_{c=1}^{C_k} p(c|k) \;\; p(x_l|c,k)^{1-\alpha} \;\; p(\{z_1^L\}_l|c,k)^{\alpha} \right\}$$

where α is a weighting factor and the class- and cluster-dependent probability for the set of position differences, $p(\{z_1^L\}_l|c,k)$, is modeled as a product of the probabilities for the individual position differences:

$$p(\{z_1^L\}_l|c,k) \propto \left(\prod_{\lambda \neq l} p(z_{\lambda l}|c_\lambda, c_l, k) \right)^{\frac{1}{L-1}}$$

We apply maximum approximation to determine c_l and c_λ: $c_l = \underset{c}{\operatorname{argmax}} \{p(x_l|c,k)\}$, $c_\lambda = \underset{c}{\operatorname{argmax}} \{p(x_\lambda|c,k)\}$ to reduce computation time. For each pair c_l, c_λ of (appearance) densities, a set of $J_{c_\lambda c_l}$q Gaussian distributions over the relative positions is estimated from training data:

$$p(z_{\lambda,l}|c_\lambda, c_l, k) = \sum_{j=1}^{J_{c_\lambda c_l}} p(j|c_\lambda c_l k) \mathcal{N}(z_{\lambda,l}|\mu_{z c_\lambda c_l k}, \Sigma_{z c_\lambda c_l k})$$

5 Databases and Experimental Results

We use two different databases to experimentally evaluate the performance of the proposed approach: the Caltech database and the PASCAL visual object classes challenge tasks.

Caltech databases. The Caltech tasks were introduced by Fergus et al. [5]. The task is to determine whether an object is present in an image or not. For this purpose, several sets of images containing certain objects (airplanes, faces, and motorbikes) and a set of background images not containing any of these objects[1] are given. The images are of various sizes and for the experiments they were converted to gray images. The airplanes and the motorbikes task consist of 800 training and 800 test images each, the faces task consists of 436 training and 434 test images. For each of these tasks, half of the images contain the object of interest and the other half does not. An example image of each set is shown in Figure 4. Many different groups have published results for these data. The Caltech tasks turn out to be quite easy, because the objects to be recognized dominate the images, have little scale variances, and are fully visible.

[1] `http://www.robots.ox.ac.uk/~vgg/data`

Figure 4: Example images from the Caltech database.

Figure 5: Example images from the PASCAL database.

PASCAL databases. In contrast to these tasks, the PASCAL database[2] is far more challenging. This database was used for the PASCAL Visual Object Classes Challenge 2005. It contains images of four different object classes: bicycles, cars, motorbikes and people. Unlike the Caltech database, no background class is given, for each of the four object classes, the remaining three form the negative class. While the whole database contains 684 training and 689 testing images, the different classes are not uniformly distributed. Example images for the four object classes are shown in Figure 5. We use the images from the PASCAL 1 tasks.

Recognizing the objects in the images of the PASCAL database is much harder. The objects appear at very different scales, where the smallest of the objects make up only a negligible part of the image at all. The objects are also shown from different view points. For example, some cars are shown from the right, others from the front, and even images showing cars from above are present. Furthermore, some objects are partially occluded by other objects and thus are not fully visible. To complicate recognition further, a few objects appear rotated, and others are depicted at different lighting conditions. These circumstances turn out to be tough for object recognition.

Experimental Results In the following we describe the experiments we performed with the Gaussian mixture models. We start from a baseline system that we use to tune the parameters of our system, which can obtain very competitive results. Then, spatial information is added and discriminative training is applied, which leads to a significant gain in classification performance.

For the baseline experiments, we extract 200 patches around interest points per image

[2]available at `http://www.pascal-network.org/challenges/VOC/voc2005/index.html`

of the size 11×11 pixels. These features are PCA transformed keeping 40 dimensions. For each class a mixture of 256 densities is estimated. In the first step, these parameters are optimized:

Extraction points: The choice of the extraction points is changed and in informal experiments different types of extraction points have been tested. We found that using more points usually leads to better results and that 200 random points in addition to the 200 wavelet-based salient points are sufficient to gain nearly optimal results [8].

Extraction size: The use of patches of different sizes simultaneously leads to better results than using one single size alone. Thus, at each extraction position, patches of four different scales are extracted (7×7, 11×11, 21×21, 31×31).

The results of these experiments can be seen in the first three lines of Table 1. It can clearly be seen that in all cases the use of more extraction points and multiple sizes strongly reduces the equal error rates.

Starting from this improved baseline, the effect of spatial information is tested (cf. Table 1). Absolute and relative positions lead to improvements over the previous results. Interestingly, despite the loss of translation invariance, absolute patch positions outperform relative patch positions. One explanation for this observation may be that the tasks addressed do not strictly require translation invariance or that the necessary translations are sufficiently represented in the training data and can thus be recognized correctly in the test data.

Due to these results and the fact that the incorporation of relative patch positions requires a lot of additional time during training and classification, we use absolute patch positions for the remaining experiments.

Initial informal experiments with the discriminative refinement of the cluster weights have shown that the method presented in Section 3 did not improve the equal error rates in the experiments. Still, we observed that the confidences for the classification of the patches were strongly improved. To overcome this problem, we apply a variant called "*falsifying training*" that has been used in the domain of speech recognition before [14]. Therefore, in Eq. (6), the sum does not range over all training images N, but only over those images which have been classified worst in the previous training iterations, i.e. those images for which the probability for the correct class is least. In our experiments, we take the 20% worst classified images in each training iteration and from Table 1 it can be seen that the results were improved.

Table 1 also compares our results with the best results other groups have reported for the Caltech and the PASCAL database. It can be seen that our method clearly outperforms all other methods for the Caltech tasks. For the PASCAL tasks, our method is nearly as good as the best result from the 2005 PASCAL visual object classes challenge [4].

Figure 6 gives example classification results of our method where the most discriminative patches are depicted by green and red squares drawn into the image. While green squares denote patches with high probability for the object class, red squares denote patches with high probability against the object class.

6 Conclusion

We have presented a novel part-based approach to object recognition using Gaussian mixture densities, which is completely based on probability theory. Starting from a generative

Table 1: Results (EER) of other approaches compared to our method.

	Caltech			PASCAL			
Spatial relationships	Airpl.	Faces	Motb.	Bicyc.	Cars	Motb.	Peop.
This work, baseline	1.5	3.2	3.5	11.0	11.1	10.6	22.6
+ Random points	1.3	0.5	3.5	11.3	9.1	9.3	19.0
+ Multiple scale	0.8	**0.0**	2.3	8.9	9.1	7.4	16.7
+ Absolute patch positions	0.5	**0.0**	0.8	2.6	6.3	3.7	13.1
Relative patch positions	0.8	**0.0**	1.5	7.9	7.5	5.7	13.1
+ Discr. training	**0.5**	**0.0**	**0.3**	**1.6**	**5.1**	**3.0**	**8.6**
Boosting hypothesis [13]	2.5	**0.0**	5.7				
Discr. trained histograms [2]	1.4	3.7	1.1	13.2	7.5	6.0	13.9
PCA SIFT [18]	1.7	0.3	1.0				
Star model [6]	6.4	9.7	2.7				
Discr. sc. inv. descriptors [3]				7.0	**3.9**	**2.3**	**8.3**
SIFT features [12]				31.3	20.7	27.8	42.9

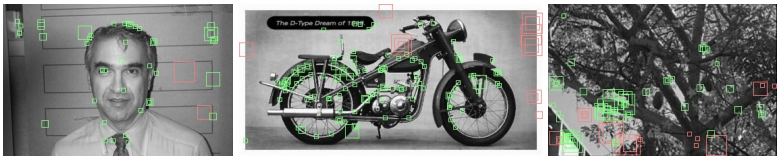

Figure 6: A correctly classified face, a correctly classified motorbike, and a background image classified as motorbike.

model that is refined using a discriminative training criterion, we present an effective way of combining generative and discriminative ideas. The final model is a generative model with all its merits but achieves the classification performance of discriminative models. It is extended to incorporate absolute and relative positions of the parts. The use of spatial information leads to a significant improvement of the results.

By combining these methods, i.e. incorporating spatial information and discriminative training of the mixture weights, we obtained very good results on several databases. The very promising results on the four tasks of the PASCAL database prove that the proposed method is able to successfully classify images under challenging conditions.

Two aspects of this work require further research: first, the incorporation of relative patch positions needs to be investigated further, as theoretically it should lead to better results than absolute patch positions, but in practice is outperformed. Second, it is also possible to refine the means and the variances of the mixture densities and our first informal experiments have shown that this approach is promising.

References

[1] T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition using Image Patches. In *CVPR 05*, vol. 2, San Diego, CA, pp. 157–162, June 2005.

[2] T. Deselaers, D. Keysers, and H. Ney. Improving a Discriminative Approach to Object Recognition using Image Patches. In *DAGM 2005, Pattern Recognition*, LNCS 3663, Vienna, Austria, pp. 326–333, Aug. 2005.

[3] G. Dorko and C. Schmid. Selection of Scale-Invariant Parts for Object Class Recognition. In *ICCV 2003*, vol. 1, Nice, France, pp. 634–640, Oct 2003.

[4] M. Everingham, et al. The 2005 PASCAL Visual Object Classes Challenge. In *Selected Proceedings of the first PASCAL Challenges Workshop*, LNAI, Southampton, UK, in press, 2006.

[5] R. Fergus, P. Perona, and A. Zissermann. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *CVPR 03*, Blacksburg, VG, pp. 264–271, June 2003.

[6] R. Fergus, P. Perona, and A. Zisserman. A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition. In *CVPR 05*, San Diego, CA, pp. 380–389, June 2005.

[7] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. In *ICCV 2005*, vol. 2, Beijing, China, pp. 1363 – 1370, Oct. 2005.

[8] A. Hegerath. Patch-based Object Recognition. Diploma thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany, March 2006.

[9] A. Holub and P. Perona. A Disriminative Framework for Modelling Object Classes. In *CVPR 05*, San Diego, CA, pp. 663–670, June 2005.

[10] Y. Linde, A. Buzo, and R. Gray. An Algorithm for Vector Quantization Design. In *IEEE Transactions on Communications*, vol. 28, pages 84–95, Jan 1980.

[11] E. Loupias, N. Sebe, S. Bres, and J. Jolion. Wavelet-based Salient Points for Image Retrieval. In *ICIP 2000*, vol. 2, Vancouver, Canada, pp. 518–521, Sep. 2000.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journ. Computer Vision*, 60(2):91–110, Feb. 2004.

[13] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic Object Recognition with Boosting. IEEE Trans. PAMI 28(3):416–431, March 2006.

[14] R. Schlüter, W. Macherey, B. Müller, and H. Ney. Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition. *Speech Communication*, 34:287–310, May 2001.

[15] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV 2005*, vol. 2, Beijing, China, pp. 1331 – 1338, Oct. 2005.

[16] I. Ulusoy and C. M. Bishop. Generative Versus Discriminative Methods for Object Recognition. In *CVPR 05*, San Diego, CA, pp. 258–265, June 2005.

[17] J. Winn, A. Criminisi, and T. Minka. Object Categorization by Learned Universal Visual Dictionary. In *ICCV 05*, vol. 2, Beijing, China, pp. 1800 – 1807, Oct. 2005.

[18] W. Zhang, B. Yu, G. J. Zelinsky, and D. Samaras. Object Class Recognition Using Multiple Layer Boosting with Heterogeneous Features. In *CVPR 05*, San Diego, CA, pp. 323–330, June 2005.