Incorporating the Boltzmann Prior in Object Detection Using SVM

Margarita Osadchy and Daniel Keren
Computer Science Department
University of Haifa
Mount Carmel, Haifa 31905, Israel
(rita,dkeren)@cs.haifa.ac.il

Abstract

In this paper we discuss object detection when only a small number of training examples are given. Specifically, we show how to incorporate a simple prior on the distribution of natural images into support vector machines. SVMs are known to be robust to overfitting; however, a few training examples usually do not represent well the structure of the class. Thus the resulting detectors are not robust and highly depend on the choice of the training examples. We incorporate the prior on natural images by requiring that the separating hyperplane will not only yield a wide margin, but also that the corresponding positive half space will have a low probability to contain natural images (the background). Our experiments on real data sets show that the resulting detector is more robust to the choice of training examples, and substantially improves both linear and kernel SVM when trained on 10 positive and 10 negative examples.

1 Introduction

In this paper we focus on appearance-based object detection against general background. A great deal of work has been done in this direction ([21, 31, 29, 10, 26, 23, 24, 8, 18, 14, 1] and many others). However the problem is still challenging, because most of the existing methods require thousands of training images of the object. These images have to be collected, and sometimes manually segmented or aligned, which is tedious and expensive task. Another conundrum is the modeling of the background class. While the patterns of the target object usually have a definitive (to some degree) structure, the background is a much richer class, and learning it properly requires collecting huge data sets with large variability for the initial training, followed by several iteration of retraining on false positives. In this paper we address these issues and propose a method for object detection from a small number of both positive and negative examples.

The problem of learning from a few examples has been recently addressed in [30, 15, 16, 7, 32]. The techniques proposed in these papers are very different, but all of them stress the importance of the problem.

We base our method on Support Vector Machine (SVM) [28]. SVM separates two classes by a hyperplane with maximal margin between the two sets of training examples. The maximal margin constraint can be considered as a regularizer that improves the generalization of the classifier, thus allowing to train from relatively small datasets. However, in the case of object detection the dimensionality of images viewed as vectors is very high and even SVM cannot learn the separation properly from a small number of examples. Few training examples do not represent the distribution of the classes and can fall far from the separation boundary. In this case, demanding a maximal margin will not help to choose the correct separation hyperplane. Fortunately, images are not spread equally in all dimensions. The statistics of natural images shows that most of the energy in them resides in low frequencies. This means that the distribution of natural images is directional and this property can be efficiently used as a prior on background images. We incorporate this prior by requiring that the separating hyperplane will not only separate with a wide margin, but also that the corresponding positive half space (i.e. the set of images that are accepted) will have a low probability to contain natural images (the background). We show experimentally that the performance of a maximal margin classifier trained on 10 positive and 10 negative examples is highly dependent on the choice of the training samples. When samples are "good", it performs well, but if samples fall far from the boundary the performance drops significantly. Our method that combines a wide margin constraint with the prior on natural images is more robust to a choice of training samples and substantially improves over both linear and kernel SVMs.

2 Related Work

Object recognition models usually have a large number of parameters. Learning these parameters requires many training examples. The penalty for using small training sets is overfitting. This means that the classifier shows very good performance on the training samples, but fails to classify unseen samples. Such a classifier has poor generalization. There are two common (and somewhat related) ways to reduce overfitting: 1) regularization, 2) using priors.

There are several regularization techniques. One is limiting the number of free parameters in the model, for example LDA [6], RDA[3], or the more recent QDDA [4]. Another regularization technique is to minimize the empirical error subject to constraints on the learned functions, for example maximal margin in SVM [28]. Another method is *noise injection* [5], where the training dataset is enriched by multiple copies of each training data point with zero-mean, low-variance Gaussian noise. By keeping the labels of all the copies the same, these techniques force the classification function to be smoother and reduce overfitting. Another variation of noise injection is using rotated ([25]) or corrupted [30] copies of the training samples.

The second method to reduce overfitting is by using priors. Since priors can be of different nature, there is no one single way of incorporating them into a model. In [16] a Bayesian approach is employed, in which the prior information about object categories is incorporated as a probability density function of the parameters of the generative model. This density function is learned from previously seen models or unrelated categories. The prior is updated, when observations become available, into a "posterior" to be used for recognition. A very different way of using prior knowledge is introduced in [7]. This method learns to discriminate between two classes given a single example from each class. This is achieved by first learning from more examples of other related classes a metric over the instance space that guarantees that all within-class distances are smaller than all distances between classes. Then, the original examples are used in a nearest neighbor classifier that calculates distances using the class relevance metric. Fung et. al. [9] suggests incorporating a prior knowledge in the context of SVMs. They assume that there are polyhedral sets which are known in advance to belong to one of the categories (positive or negative).

Fisher kernel [12] allows incorporating prior knowledge about the data distribution into the classification process. Based on this feature, one can consider the Fisher kernel to be a possible choice for incorporating the knowledge on the distribution of natural images into SVMs. We discuss the Fisher kernel in Section 3.2.

In this paper we consider the problem of object detection against a general background, and the prior we use is on the distribution of the background class. We don't use any prior on the object class. Such use of prior is asymmetric, but the problem is asymmetric as well. The background class is significantly richer than the object class. Since we employ a discriminative framework, incorporating the prior on a much bigger class is a sensible method to reduce overfitting.

The prior on the distribution of the natural images has been successfully used in the *antiface* method [13] to model the negative set instead of learning it from examples.

Learning from a single example was also addressed in event recognition in video [32]. The prior we use for images also holds for video [20], and this suggests that our method can be extended to event detection.

3 Background

Our model combines linear SVM with soft classification with a prior on the distribution of natural images. In this section we give some background on both topics.

3.1 Linear SVM

The support vector machine (SVM) [28] model has proved to be extremely useful for the task of image recognition. Given positive and negative examples, SVM tries to find a hyperplane which separates them, and has the maximum margin among all separating hyperplanes.

Formally, consider a training set of m i.i.d. samples $(x_1, y_1), .., (x_m, y_m)$, where x_i, for $i = 1, .., m$ is a vector of length n and $y_i = \{+1, -1\}$ is the class label for data point x_i. In its simplest form, SVM searches for a vector of coefficients w and scalar b such that
(a) $y_i(w \cdot x_i + b) \geq 1, \quad \forall i$
(b) the margin between two classes, given by $2/\|w\|$, is maximal.
Such a hyperplane is found by solving the following minimization problem:

$$\min \frac{1}{2}\|w\|^2$$
$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1, \quad \forall i$$

This problem can be reformulated in a dual representation (in terms of support vectors) and solved using quadratic programming.

In practice, a separating hyperplane may not exist due to high noise level. To allow the possibility of samples violating condition (a), slack variable are introduces:

$$\xi_i \geq 0, \quad \forall i \tag{1}$$

in order to relax the constraints in (a) to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \forall i \tag{2}$$

Such a *soft margin* classifier is obtained by optimizing

$$\min \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i \tag{3}$$

subject to constraints 1 and 2, where the constant $C > 0$ determines the trade-off between margin maximization and training error minimization.

SVMs can be extended to non-linear models by using the famous "kernel trick" [28].

3.2 Modeling of natural images

SVM makes no assumptions on the distribution of classes; it learns the separation from the given training set. However, it is known that typical images are "smooth", that is, most of their energy is concentrated in the low frequencies. It was empirically demonstrated [13] that the distribution of natural images roughly follows the Boltzmann distribution, which has its origins in statistical mechanics. According to this distribution, the probability of a physical system to be in a certain state X equals $\frac{1}{Z}\exp(-\frac{E(X)}{kT})$, where $E(X)$ is the system's energy, T its temperature, k the Boltzmann constant, and Z a normalizing factor (the "partition function").

The idea to apply this type of probability was introduced in the seminal work [11]. Various modifications have been proposed. A simple choice for the energy function, which has proved helpful for detection [13], is to choose a global measure of image roughness as the image "energy". Thus, up to normalization, the probability of an image I is defined as e.g.

$$Pr(I) \propto \exp\left(-\lambda \iint (I_x^2 + I_y^2)dxdy\right) \tag{4}$$

where λ is a positive constant.

It is much easier to work in the Fourier or DCT domain, which diagonalizes the energy operator. For example, for a discrete $n \times n$ image I with DCT coefficients $\tilde{I}_{k,l}$, the probability density is defined as e.g.

$$Pr(I) \propto \exp\left(-\lambda \sum_{k,l}(k^2 + l^2)\tilde{I}_{k,l}^2\right) \tag{5}$$

Next we would like to use this simple prior to reduce overfitting in training of object detectors from a small number of examples.

The *Fisher kernel* [12] allows incorporating the prior knowledge on the marginal distribution $p(x)$ about the data into SVM. Thus it seems natural to use Boltzmann distribution as $p(x)$ in Fisher kernel formulation. Next we will show that this approach reduces to a Mahalanobis distance.

The Fisher kernel is defined as follows. For parameters θ^0 and observed data x, the derivative of the log likelihood is denoted $g(\theta^0, x) = \frac{\partial \log p_{\theta^0}(x)}{\partial \theta}$. Next, define $I = E[g(\theta^0, x)g(\theta^0, x)^T]$, and then the kernel is defined as $K(x, y) = g(\theta^0, x)I^{-1}g(\theta^0, y)^T$.

A straightforward calculation yields that for a diagonal Gaussian prior $p_0(x_1, ...x_n) \propto \exp\left(-\sum_i \frac{x_i^2}{2\sigma_i^2}\right)$, the Fisher kernel equals

$$K(x, y) = 1 - \sum_i \frac{x_i^2 + y_i^2}{2\sigma_i^2} + \sum_i \frac{x_i^2 y_i^2}{2\sigma_i^4}\ .$$

To see what this means about the distances induced in feature space, recall that a kernel $K(x, y)$ is associated with a map Φ into feature space, such that $K(x, y) = (\Phi(x), \Phi(y))$, hence $\|\Phi(x) - \Phi(y)\|^2 = K(x, x) - 2K(x, y) + K(y, y)$. A straightforward calculation yields that for the Fisher kernel $\|\Phi(x) - \Phi(y)\|^2 = \sum_i \frac{(x_i^2 - y_i^2)^2}{2\sigma_i^4}$. Thus, the Fisher kernel in this case yields a Mahalanobis-like distance, which is also noted in [2].

Note that the approach proposed in this paper doesn't measure the distances "correctly"; it proposes an optimization framework for reducing misclassification errors.

The model of natural images (Eq. 5) has been successfully used in the *antiface* method [13] to model the negative set instead of learning it from examples. In this approach, the positive set is tightly bounded by a number of slabs that are *"tilted away"* from the set of smooth – i.e. typical – images. This greatly reduces the number of false alarms. For further details, see [13].

A similar idea can be used in a discriminative framework. We will still require that the region of the space that a classifier predicts as the positive set will be chosen by minimizing the probability of natural images in this region.

4 Our Approach

When trained on an "unfortunate" choice of examples, the maximal margin separation in SVM is unfavorably affected by the "bad" training samples. We suggest to use the above approximation of the background class to choose the hyperplane (w, b) such that the positive half space

$$\mathcal{H} = \{(x_1...x_n) | \sum_{i=1}^n w_i x_i + b \geq 0\}$$

will contain as few natural images as possible.

In order to keep the notations used in Section 3.1, we will denote the DCT coefficients $\tilde{I}_{k,l}$ of an image I by x_i. Then the probability density in Eq. 5 will take the following

form:

$$Pr(I) \propto \exp\left(-\sum_{i=1}^{n} d_i x_i^2\right) \quad (6)$$

where $d_i = \lambda(k^2 + l^2)$.

We minimize the probability of natural images, described by Eq. 6, over the positive half space $\mathcal{H}$:

$$\min_{w,b} \int_{\mathcal{H}} \exp\left(-\sum_{i=1}^{n} d_i x_i^2\right) dx_1 ... dx_n \quad (7)$$

To compute the integral we make a change of variables $u_i = \sqrt{d_i} x_i$; this transforms the integrand to

$$\exp\left(-\sum_{i=1}^{n} u_i^2\right)$$

and the half-space to

$$\mathcal{H}' = \{(u_1 ... u_n) | \sum_{i=1}^{n} (\frac{w_i}{\sqrt{d_i}}) u_i + b \geq 0\}$$

Define a vector

$$w_i' = \frac{w_i}{\sqrt{d_i}}$$

(the Jacobian of the transformation need not be considered because it is fixed in the subsequent analysis, as are factors such as $(2\pi)^{n/2}$ which appear in the Gaussian integral). The integral simplifies to

$$\int_{\{u|(w',u)+b \geq 0\}} \exp\left(-\|u\|^2\right) du_1 ... du_n$$

Since the norm function is isotropic, one can rotate the axis without changing the integral's value; this is equivalent to rotating w'. Rotate the axis then so that w' transforms to $(\|w'\|, 0, 0, ...0)$. It is straightforward to see that, again up to a constant which is fixed throughout, the integral equals the one dimensional integral

$$\int_{\{x| \ \ \|w'\| x + b \geq 0} \exp(-x^2) dx$$

If we denote the above integral by Q then:

$$Q = \int_z^{\infty} \exp(-t^2) dt = \frac{\sqrt{\pi}}{2} \mathsf{erfc}(z)$$

where

$$z = \frac{-b}{\|w'\|}$$

Returning to the original variables we get

$$Q = \frac{\sqrt{\pi}}{2} \mathsf{erfc}\left(\frac{-b}{\sqrt{\sum_{i=1}^{n} \frac{w_i^2}{d_i}}}\right) \quad (8)$$

We empirically tested the correctness of this model. We randomly chose w, constraining its norm to be 1, and randomly chose b in the range [-0.5, 0,5]. For each choice of w, b we computed the integral in Eq. 8. Figure 1 shows the percentage of randomly chosen natural images that lie in the positive half space as a function of the integral value. It's very clear from this plot that the number of the background images residing in the positive half space is nearly proportional to the value of the integral. This demonstrates that by minimizing the value of the integral we minimize the probability of natural images in the positive half space. Based on these findings we propose the following algorithm that combines the wide margin requirement with the Boltzmann prior.

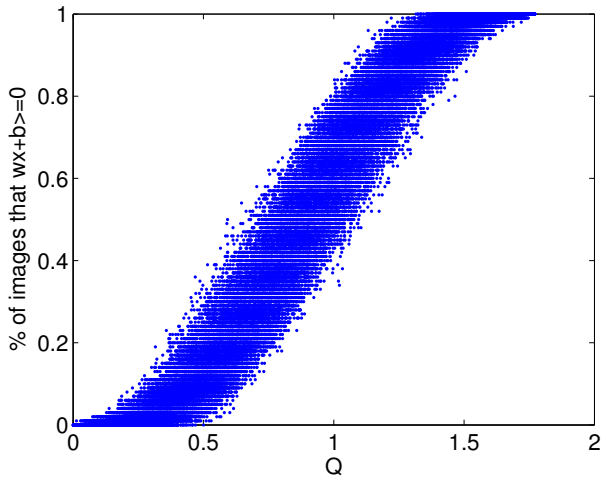

Figure 1. Relation between the number of natural random images in the positive half-space and the integral defined in Eq. 8.

4.1 Incorporating the prior

We showed that by minimizing the expression in Eq. 7 we reduce the probability of the positive half space to include background images. On the other hand we require this half space to include samples from the positive set. Note that we make no assumptions on the smoothness of the object class. The algorithm seeks an acceptance region which

contains as little smooth images as possible – except for the object class. Even if the object class samples are smooth, the acceptance region still contains a relatively small volume of the entire set of smooth images. Thus, the algorithm will perform well both for a smooth and non-smooth object class.

We also make use of negative examples, because if they are located close to the positive set, they provide information useful for discrimination. Thus we combine the SVM criteria with minimization of Eq. 7. Specifically if x_i are training samples and y_i are corresponding labels, $y_i \in \{-1, 1\}$, $i = 1..m$ (m is small), we first use standard SVM to obtain w_{MM} and b_{MM} that provide maximal margin on the training set transformed to frequency domain. Then we refine this hyperplane by solving the following minimization problem:

$$\min_{w,b} \frac{\sqrt{\pi}}{2} \mathsf{erfc}\left(\frac{-b}{\sqrt{\sum_{i=1}^{n} \frac{w_i^2}{d_i}}} \right) + C \sum_i \xi_i \qquad (9)$$

subject to

$$||w||^2 \leq \alpha ||w_{MM}||^2$$
$$y_i(wx_i + b) \geq 1 - \xi_i, \quad i = 1..m$$
$$\xi_i \geq 0, \quad i = 1..m$$

The above formulation of the problem sacrifices some of the margin's width in order to allow the prior to influence the hyperplane. The shrinkage of the margin is controlled by the parameter α, which determines the tradeoff between the data and the prior. In our experiments we used $\alpha = 2$.

The optimization in Eq. 9 is non-linear. We applied a standard routine for constrained non-linear optimization that uses a sequential quadratic programming (SQP) method [22]. In this method, the optimization solves a quadratic programming (QP) subproblem at each iteration. An estimate of the Hessian of the Lagrangian is updated at each iteration using the BFGS formula.

When the new w and b are found, the classification proceeds as with the ordinary linear classifier – an input x is classified according to $\mathsf{sgn}(w \cdot x + b)$.

5 Experiments

The goal of the experiments is to test the effectiveness of incorporating the prior in the case of rather small training sets. The results presented in this section are with training sets of 10 positive and 10 negative examples. Our experiments with less examples still show a significant advantage of our method. However, overall performance is not good enough to make an interesting case. The explanation for this is that too few positive examples with no prior on the object class are insufficient for training good classifiers.

5.1 Data sets

We have experimented with the UIUC car database[1] [1] and with the CBCL face set [2][27, 23].

In the car set we used the part that contains images of cars cropped around the bounding box. This collection contains 549 grey level images of cars viewed from the side. In order to speed up the testing we downscaled the original images from 100x40 pixels to 35x14 (which also makes the recognition task harder). We reserved the first 250 images for training and the rest were used for testing. This set contains also 500 background images. We reserved 250 background images for training and the rest were used for testing. Since the remaining part contains only 250 background images, we needed to add more background images from other source to make the ratio more realistic. A real image may contain just a few cars, but tens or even hundreds of background non-overlapping images. We added 9861 background images cropped from the Graz2 database[3][19] that contains general background images with no cars. Many of these images show roads and buildings which are natural surroundings of cars. Overall, our test set contained 299 car images and 10111 background images.

The face set contains 19x19 grayscale images of nearly frontal faces [27, 23]. The training set consists of 2,429 faces and 4,548 non-faces. The test set consists of 472 faces, 23,573 non-faces.

5.2 Experiments

We compared our method, outlined in Section 4.1, to linear SVM with soft margin and to kernel SVM with second degree polynomial kernel. The experiments reported here are with $C = 10$. However, varying C didn't influence the performance. All the tests were performed in MATLAB using OSU SVMs Toolbox based on LIBSVM [17] and MATLAB Optimization Toolbox for minimizing the expression in Eq. 9.

We ran 50 trials. In each trial we have randomly chosen 10 positive and 10 negative examples that were used to train linear SVM, kernel SVM and the SVM with prior on the background class. So overall we trained 50 classifiers for each method. Then we tested these classifiers on the test set. This experiment showed that our method generally performs better than linear and kernel SVMs that assume no prior knowledge, and it's more robust than other methods to

[1] http://l2r.cs.uiuc.edu/ cogcomp/Data/Car/

[2] http://cbcl.mit.edu/cbcl/software-datasets/FaceData2.html

[3] http://www.emt.tugraz.at/ pinz/data/

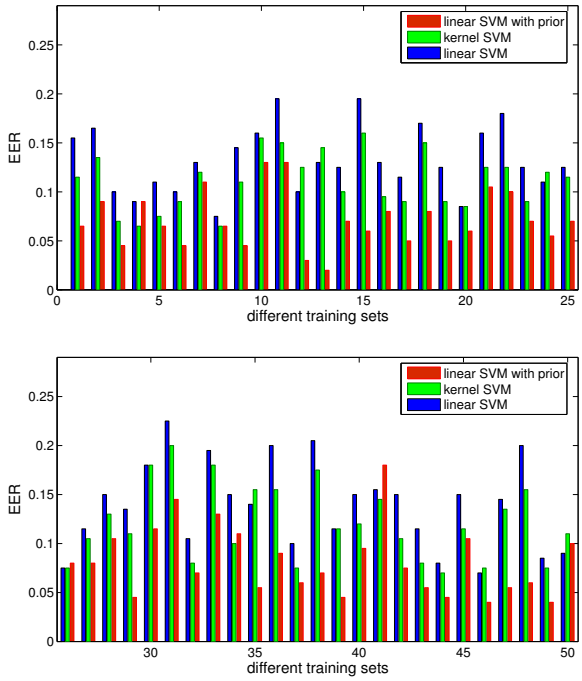

Figure 2. Results on the car set: Equal Error Rate (ERR) of three types of classifiers trained on different random training sets. The x-axis corresponds to trials, the number of samples was the same: 10 positives and 10 negatives in all trials.

the choice of training examples. This is not surprising, because it uses a prior on the larger class that helps to reduce overfitting. In order to quantify the results, we computed the equal error rate (EER) of each classifier in each trial. Figure 2 shows the results, and Table 1 summarizes means and variances of the detection performance in the car experiment (the point on the ROC curve that corresponds to EER) of the three method in 50 trials. Both Figure 2 and Table 1 clearly show the advantage of using the Boltzmann prior in SVM framework.

We ran the same protocol on the face database. Our method outperformed both linear and kernel SVM on this set too. However, the performance of all three methods was lesser than on the car set, probably due to higher variability in the set. Table 2 shows the average performance of the three methods in this experiment.

In summary, our method outperforms linear and kernel SVM. Although the gap in performance with kernel SVM is smaller, our method is significantly faster than the kernel approach. When trained on 20 examples, kernel SVM usually chooses all of them as support vectors. Thus, during runtime it must compute the kernel function for all the

	recognition rate mean(%)	recognition rate std(%)
Linear SVM	84	4.0
Kernel SVM	87	3.7
SVM with prior (ours)	93	3.1

Table 1. Results on car set: Mean and STD of recognition rate for classifiers of each type trained on 50 different random choices of 10 positive and 10 negative examples.

	recognition rate mean(%)	recognition rate std(%)
Linear SVM	69	5.9
Kernel SVM	71	6.2
SVM with prior (ours)	74	3.2

Table 2. Results on face set: Mean and STD of recognition rate for classifiers of each type trained on 50 different random choices of 10 positive and 10 negative examples.

support vectors (20 in this case), compared to a single inner product in our method.

6 Conclusions and future research

This paper suggests to incorporate the prior information on the background, in the form of a Boltzmann-like distribution, into the linear SVM paradigm. The suggested method outperforms both linear and kernel SVM; its detection rate is much higher than linear SVM, and significantly higher than kernel SVM. It is also much faster than kernel SVM.

We plan to further test and improve the method, e.g. by extending it to the kernel case, and refining the prior.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *In Proceedings of the Seventh European Conference on Computer Vision*, volume IV, pages 113–130, 2002.

[2] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, and N. Le Roux. Spectral clustering and kernel pca are learning eigenfunctions. *Technical Report 1239, Département d'informatique et recherche opérationnelle, Universite de Montréal*, 2003.

[3] H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:17431748, 1996.

[4] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. In *International Conference on Applied Stochastic Models and Data Analysis*, 2005.

[5] L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2nd edition, 2000.

[7] M. Fink. Object classification from a single example utilizing class relevance pseudo-metrics. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2004.

[8] R. Freund, F. Girosi, and E. Osuna. Training support vector machines: an application to face detection. In *CVPR*, pages 130–136, 1997.

[9] G. Fung, O. L. Mangasarian, and J. W. Shavlik. Knowledge-based support vector machine classifiers. In *Proc. Advances in Neural Information Processing Systems*, pages 521–528. MIT Press, 2002.

[10] C. Garcia and M. Delakis. A neural architecture for fast and robust face detection. *IEEE-IAPR Int. Conference on Pattern Recognition*, pages 40–43, 2002.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.

[12] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Proc. Advances in Neural Information Processing Systems*, pages 487 – 493. MIT Press, 1999.

[13] D. Keren, M. Osadchy, and C. Gotsman. Antifaces: A novel, fast method for image detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(7):747–761, 2001.

[14] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf. Face detection - efficient and rank deficient. In *Advances in Neural Information Processing Systems*, volume 17, pages 673–680, Cambridge, MA, 2005. MIT Press.

[15] K. Levi and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume II, pages 53–60, 2004.

[16] F.-F. Li, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of International Conference on Computer Vision*, pages 1134–1141, 2003.

[17] LIBSVM. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[18] M.Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

[19] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of 8th European Conference on Computer Vision*, volume II, pages 71–84, 2004.

[20] M. Osadchy, D. Keren, and Y. Gal. Anti-sequences: Event detection by frame stacking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 46–51, 2001.

[21] M. Osadchy, M. Miller, and Y. LeCun. Synergistic face detection and pose estimation with energy-based models. In *Proc. Advances in Neural Information Processing Systems (NIPS 2004)*. MIT Press.

[22] M. Powell. A fast algorithm for nonlinearly constrained optimization calculations. *Numerical Analysis, G.A.Watson ed., Lecture Notes in Mathematics*, 630, 1978.

[23] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20:22–38, 1998.

[24] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, 2000.

[25] B. Scholkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *In Artificial Neural Networks — ICANN'96*, volume 1112, pages 47–52. Springer Lecture Notes in Computer Science, 1996.

[26] K. Sung and T. Poggio. Example-based learning of view-based human face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20:39–51, 1998.

[27] K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

[28] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[29] P. Viola and M. Jones. Fast and robust classification using asymmetric Adaboost and a detector cascade. In *Proc. Advances in Neural Information Processing Systems (NIPS 2001)*, pages 1311–1318. MIT Press, 2001.

[30] L. Wolf and I. Martin. Robust boosting for learning from few examples. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 359 – 364, 2005.

[31] J. Wu, J. M. Rehg, and M. D. Mullin. Learning a rare event detection cascade by direct feature selection. In *Proc. Advances in Neural Information Processing Systems (NIPS 2003)*. MIT Press, 2003.

[32] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume II, pages 123–130, 2001.