International Conference on Image Processing, Hong Kong, 26-29 September 2010

CLASSIFICATION AND FEATURE SELECTION WITH HUMAN PERFORMANCE DATA

Christina Pavlopoulou and Stella X. Yu

Boston College
Computer Science Department
Chestnut Hill, MA

ABSTRACT

We investigate the utility of a novel form of prior, namely the accuracies with which humans categorize briefly displayed images. Such information reflects the complexity of an image for the visual system and carries information about the features important for categorization. We incorporate the prior in an SVM framework, by biasing the decision boundary towards examples difficult for humans, and by learning a suitable kernel. We focus on the task indoors vs. outdoors using a variety of histogram and interest point features. We observe improvement in classification especially for the indoor class when gist features are used.

Index Terms— Image classification, Image recognition, Feature extraction

1. INTRODUCTION

Many problems in computer vision are often cast as classification problems. A set of training images along with their human-provided labelings is used to predict the category of unknown images. This is particularly challenging as the unknown images can be very different from the ones in the training set. Computer vision approaches addressing this problem can be divided into those relying on very large training sets and simple computational machinery, and those relying on complex computations and few training examples.

The first type of approaches ([1, 2, 3]) relies on the premise that a large training set may contain representatives of all possible types of images. In its simplest form, such an approach categorizes a query image by comparing it, using simple descriptors, to the vast number of training images. The creation of large datasets required by such approaches was made possible by the ubiquity of internet and emergence of frameworks like the Amazon Mechanical Turk.

Alternative approaches use priors on the classes or training exemplars in conjunction with more elaborate computational techniques. Their goal is to extract as much information as possible from the few training images and transfer it to the unseen ones. Example approaches include [4, 5].

Our approach follows the second line of work: we improve the classification performance by imposing priors on the training exemplars. What differentiates our method from existing ones is the type of priors we use. In most cases, priors are derived after the features have been decided. For example, if color features are used for a task, knowing a-priori that ’blue’ is predictive of a category helps estimating the range of color features characterizing this category, without the need of many training exemplars. Our priors, on the other hand, refer to how accurately an image is categorized by humans; the feature space in which this categorization takes place is unknown. Incorporating such priors is challenging because the perceptual feature space has to be approximated as well.

To address this problem, we assume that the available machine features are a transformed superset of the perceptual features. Given a large set of machine features, our goal becomes to infer the transformation between the two spaces, and select the relevant features. These requirements can be simply modeled in a Support Vector Machine (SVM) framework. The subjects’ accuracies guide the learning of the kernel and an estimate of the transformation is provided. Additionally, they bias the decision boundary towards images difficult for humans and features likely to be used by humans are selected.

We focus on categorizing indoor vs. outdoor scenes using grayscale images. These categories are challenging because they cannot be discriminated based on any obvious salient feature, for example ‘blue sky’. We use four types of features (gist, tiny images, sparse SIFT, textons) [6], and observe improvements especially for the indoor category when gist and tiny image features are used. The utility of human performance data was first explored in [7], in conjunction with linear SVM’s and gist features. This paper provides a more natural model, includes non-linear classification, kernel learning, and a larger variety of features.

We describe the nature of our priors in Section 2, the formulation in Section 3, and our experimental validation and conclusions in Section 4.

2. A HUMAN-PERFORMANCE PRIOR

Our prior consists of the accuracies with which subjects categorize images very briefly presented. These data are obtained through a psychophysics experiment described in [8]. The stimuli employed consist of 50 indoor and 50 urban outdoor

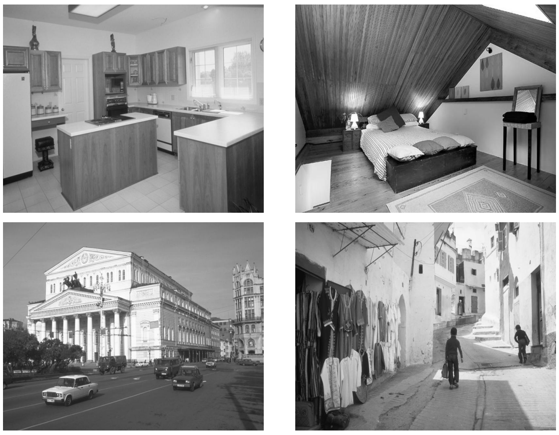

Fig. 1. Example indoor and outdoor stimuli used in the rapid categorization experiment. The images in the first column were categorized 100.0%, whereas those in the second column 75.0%. Images not categorized very accurately tend to be rich in texture and lighting.

images collected from the internet so that the spatial layout of the scene is clearly visible. During an experimental trial, a stimulus is displayed for only 16ms and is followed by a perceptual mask prompting the subject to provide his/her decision. Even for such small exposure duration, the subjects are able to categorize the scenes correctly 90.2%. In more detail, 44 scenes are categorized more than 95.0% correctly, and only 10 scenes less than 10.0%. Images not categorized very accurately are usually complex scenes, with complex lighting, rich in texture and clutter (Fig. 1).

We interpret accuracies as a form of distance from an ideal decision boundary - the decision boundary humans use. An image categorized very accurately (poorly) by humans should be far from (close to) the decision boundary. Such prior can reduce the required amount of training data. The optimal decision boundary is usually midway the points of the two classes. If the number of exemplars is small, then the computed boundary might deviate from the optimal one. Knowledge of distances may rectify this deviation.

3. MAX-MARGIN FORMULATION

We employ the subjects' accuracies in a Support Vector Machine (SVM) framework. We first explain the linear classifier formulation, proceed to the non-linear case, and follow by the learning of the kernel.

Linear Classification

SVM classification [9] is formulated as a constrained satisfaction program. For the linear case, the category of an unseen example with feature vector $\mathbf{x}$ is given by the sign of $f(\mathbf{x}) = \sum_i w_i x_i + b$. The goal is to find the function $f(\cdot)$ that minimizes the number of misclassifications. If $\mathbf{x}_i$ is the feature vector for the i-th image and $y_i \in \{1, -1\}$ its categorical label, then the separating hyperplane $(\mathbf{w}, b)$ is given by:

$$\begin{aligned} \text{min} \quad & ||\mathbf{w}||_p + C \sum_i \xi_i \\ \text{s.t.} \quad & (\mathbf{x}_i \cdot \mathbf{w} + b) \cdot y_i \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \ldots n \end{aligned} \tag{1}$$

The slack variables ξ_i relax the constraints when the data are not linearly separable. Since we are interested in feature selection as well, we use the L_1 norm, which is known to enforce additional sparsity ([10]).

To incorporate the human performance data, we substitute the 1's in the constraints with the accuracies α_i. This way we don't penalize only the number of misclassifications but also the scores that are lower than the subjects' ones.

$$\begin{aligned} \text{min} \quad & ||\mathbf{w}||_1 + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i \ (\mathbf{x}_i \cdot \mathbf{w} + b) \geq \alpha_i - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \ldots n \end{aligned} \tag{2}$$

Non-linear Classification

We include non-linearity in our formulation in a manner similar to Generalized Support Vector Machines [11]. In this case, the decision boundary has the form $f(\mathbf{x}) = \sum_i y_i k(\mathbf{x}, \mathbf{x}_i) + b$, where $k(\mathbf{x}, \mathbf{x}_i)$ is the kernel product between the unseen datapoint $\mathbf{x}$ and the training exemplar $\mathbf{x}_i$. When training, we seek to optimize:

$$\begin{aligned} \text{min} \quad & ||\mathbf{w}||_1 + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i \ (\sum_{j \neq i} y_j \ k(\mathbf{x}_i, \mathbf{x}_j) \ w_j + b) \geq \alpha_i - \xi_i \\ & \xi_i \geq 0, \quad i = 1, \ldots n \end{aligned} \tag{3}$$

Kernel Learning

The subjects' accuracies refer to a perceptual feature space which is not known. They are useful as priors when the machine features are correlated to the perceptual features. To increase the correlation, we learn the kernel $k(\cdot, \cdot)$ according to a correlation measure dictated by the subjects' accuracies.

For images with perceptual features $\mathbf{z}_i$ and $\mathbf{z}_j$, We define the perceptual correlation as:

$$r(\mathbf{z}_i, \mathbf{z}_j) = \begin{cases} 1 - |A(i) - A(j)|, & \text{if } y_i = y_j \\ |A(i) - A(j)|, & \text{if } y_i \neq y_j \end{cases} \tag{4}$$

We denote the subjects' accuracy on the i-th image with $A(i)$. When two images have the same (different) category and are categorized with similar accuracies, then their perceptual features should be correlated (uncorrelated). The above measure behaves like correlation: its values range between 0 (uncorrelated) and 1 (correlated).

	SVM (%)		D-SVM (%)	
feature type	indoors	outdoors	indoors	outdoors
gist	56.5	78.7	62.1	79.6
tiny images	56.0	67.1	60.2	66.5
sparse SIFT	74.9	62.8	75.4	63.1
textons	69.3	86.2	68.3	87.1

Fig. 2. Accuracies obtained with non-linear SVM and cosine kernel and our D-SVM with polynomial cosine kernel and polynomial transformation. The degrees of the polynomials used with D-SVM (Eqs. 5,6) were: for gist $d = 2$, $m = 5$, for tiny images $d = 2$, $m = 1$, for sparse SIFT and textons $d = 1$, $m = 1$. D-SVM outperforms SVM for gist and tiny images features.

The correlation $h(\mathbf{x}_i, \mathbf{x}_j)$ between machine feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed as the cosine between the vectors. In our formulation, we use a polynomial transformation:

$$h(\mathbf{x}_i, \mathbf{x}_j) = (\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|\mathbf{x}_i||\mathbf{x}_j|} + 1)^d \quad (5)$$

$\mathbf{x}_i \cdot \mathbf{x}_j$ denotes the inner product and $|\mathbf{x}_i|$ the Euclidean length.

To learn the actual kernel $k(\mathbf{x}_i, \mathbf{x}_j)$ used in classification we assume it is a polynomial with respect to $h(\mathbf{x}_i, \mathbf{x}_j)$. The coefficients of the polynomial are estimated with linear least-squares regression.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=0}^{m} \beta_i h(\mathbf{x}_i, \mathbf{x}_j)^i \quad (6)$$

Previous Work

Prior knowledge has been incorporated into SVM's in a variety of ways [12]. Our formulation is similar in form but different in goal from support vector regression (SVR) [13], soft-SVM [14], and weighted margin support vector machines (WMSVM) [15]. SVR does not take into account the discrete category of a datapoint and is thus unsuitable for our purpose. Soft-SVM assumes uncertainty in the class label, whereas in our case there is no such ambiguity. Finally, WMSVM makes assumptions regarding the frequency of the exemplars, whereas our prior assumptions concern the categorization difficulty of the exemplars.

4. EXPERIMENTAL EVALUATION

We compare the performance of our method using subjects' accuracies and kernel learning (D-SVM) against that of the standard non-linear SVM formulation.

We train the classifiers on our spatial layout dataset of 50 indoor and 50 outdoor images. Our testing set consists of 1000 indoor and 1000 outdoor images collected from datasets available online [2, 16, 17]. The images are at least of size 256x256. The spatial layout of the scene is clearly visible and no single object is the main focus of the image.

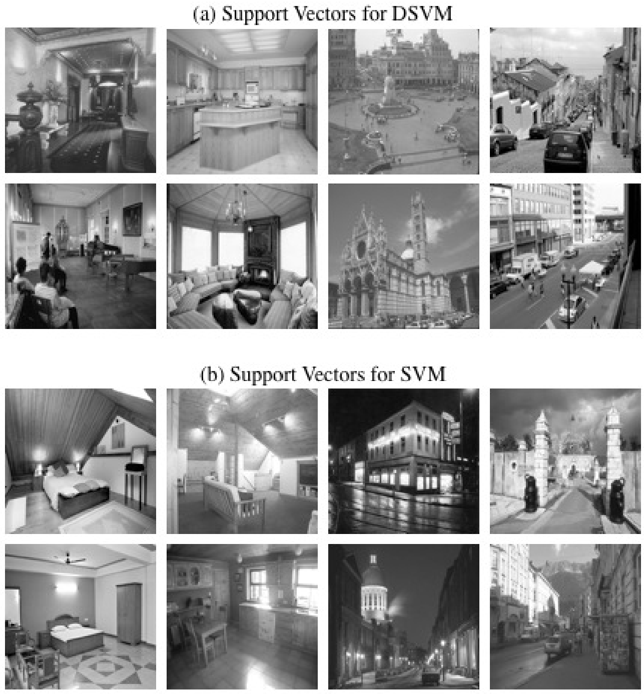

Fig. 3. **(a)** Support vectors selected by D-SVM only. **(b)** Support vectors selected by SVM only. The prior biases the selection of support vectors towards scenes of varying layout structure as opposed to varying appearance.

We compute 4 types of features available from [6].

Gist: The image is convolved with Gabor filters at 4 scales and 8 orientations. The filter responses are averaged in each of 4x4 divisions of the image [18] (512 features).

Tiny images: The image is matched to a dataset of 80 million images reduced in size [3], and the quality of the match is considered as a descriptor (3072 features).

Sparse SIFT: The SIFT descriptor is computed at Hessian-affine and MSER interest points. The SIFTs are clustered into 1000 clusters using k-means and each image is represented with two histograms of 1000 bins, where each SIFT is soft-assigned to its nearest cluster (2000 features).

Textons: A texton dictionary of 512 entries is built by clustering responses to a bank of filters [6]. For each image, a histogram is computed by assigning each pixel's set of filter responses to the nearest dictionary entry (10752 features).

Fig. 2 shows the classification results for SVM and our method (D-SVM). The cosine kernel is used for SVM; it performs better than polynomial or cosine polynomial kernels. The polynomial cosine kernel (Eq. 5) is used for D-SVM and is polynomially transformed to the perceptual correlations r (Eq. 6). Different degrees of polynomials are suitable for different features: for gist we set $d = 2$, $m = 5$, for tiny images $d = 2$, $m = 1$, and for the remaining features $d = 1$, $m = 1$. Improvements are observed for the gist and tiny images features. It is possible that sparse SIFT and textons do not benefit

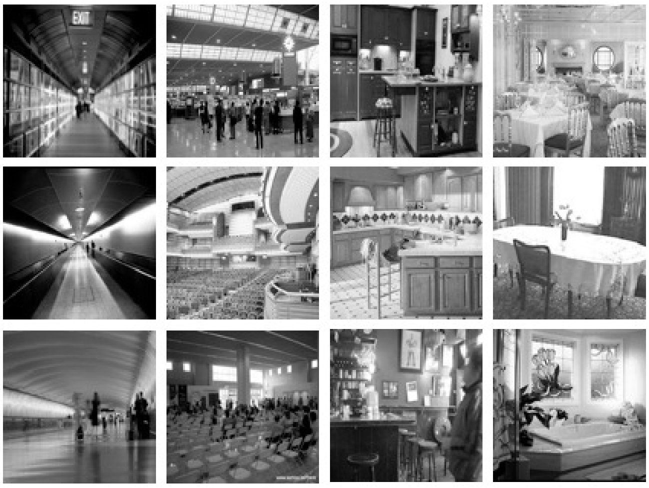

Fig. 4. Indoor scenes categorized accurately by D-SVM only: they vary in structure and layout.

from our formulation because they rely on universal dictionaries computed without knowledge of the prior. The biggest improvement observed is 5% for the indoor category in conjunction with gist features.

This improvement can be attributed to the wider variety of layout structure exhibited by the D-SVM support vectors. Fig. 3 shows support vectors selected by SVM only or D-SVM only. In total SVM selects 29 support vectors and D-SVM 27. From these, 16 are selected by SVM only and 14 by D-SVM only. SVM maximizes the margin for difficult to discriminate examples. For the gist features, it selects images rich in texture or complex lighting conditions; for example, the night scene has similar light blobs as the indoor scenes. Such images tend to have varying appearance and not necessarily varying layout structure. In contrast, D-SVM takes into account the prior and the support vectors selected are diverse in terms of layout structure.

Our method is more effective for the indoor class and some examples are shown in Fig. 4. In total, 79 images are categorized by D-SVM only and 23 by SVM only. The indoor images D-SVM categorized correctly consist of corridors, very open indoor spaces, and a variety of interiors. The features D-SVM favors characterize the geometry and the structure of a scene. This is expected, since gist features may account for the excellent performance of the human vision system when it comes to scene recognition [18].

In summary, we employ the accuracies with which humans categorize rapidly presented scenes as priors in an SVM classifier. The priors guide both the training of the classifier and the learning of the kernel. Increased performance is observed for gist and tiny images features especially for the indoor class. The gist features selected by our method characterize more effectively the layout structure of a scene as opposed to its appearance.

Acknowledgments

This research is funded by NSF CAREER IIS-0644204 and a Clare Boothe Luce Professorship to Stella X. Yu.

5. REFERENCES

[1] James Hays and Alexei A. Efros, "Scene completion using millions of photographs," in *SIGGRAPH*, 2007.

[2] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: A database and web-based tool for image annotation," *Int'l J. of Comp. Vision*, vol. 77, no. 1, 2008.

[3] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE TPAMI*, vol. 30, 2008.

[4] E. Miller, N. Matsakis, and P. Viola, "Learning from one example through shared densities on transforms," in *CVPR*, 2000.

[5] Li Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE TPAMI*, vol. 28, no. 4, 2006.

[6] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN Database: Large Scale Scene Recognition from Abbey to Zoo," in *CVPR*, 2010.

[7] C. Pavlopoulou and S.X. Yu, "Indoor-Outdoor Classification with Human Accuracies: Image or Edge Gist?," in *Workshop on Advancing Computers with Humans in the Loop*, 2010.

[8] M. Woods, C. Pavlopoulou, and S. X. Yu, "Rapid Categorization of Spatial Layout in Real Images and Line Drawings," In preparation.

[9] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[10] P.S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. of the Int'l Conf. on Machine Learing*, 1998.

[11] O. Mangasarian, "Generalized Support Vector Machines," in *Advances in Large Margin Classifiers*. MIT Press, 2000.

[12] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector machines for classification: A review," *Neurocomputing*, vol. 71, no. 7-9, 2008.

[13] A. J. Smola and B. Schlkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, 2004.

[14] Yi Liu and Y. F. Zheng, "Soft svm and its application in video-object extraction," *IEEE Trans. on Signal Processing*, vol. 55, no. 7, 2007.

[15] X. Wu and R. Srihari, "Incorporating Prior Knowledge with Weighted Margin Support Vector Machines," in *Proc. KDD*, 2004, pp. 326–333.

[16] S. X. Yu, H. Zhang, and J. Malik, "Inferring spatial layout from a single image via depth-ordered grouping," in *IEEE Workshop on Perceptual Organization in Computer Vision*, 2008.

[17] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*, 2009.

[18] A. Oliva and A. Torralba, "Modeling the shape of a scene: a holistic representation of the spatial envelope," *Int'l J. of Comp. Vision*, vol. 42, pp. 145–75, 2001.