IEEE CVPR Workshop on Advancing Computer Vision with Humans in the Loop, San Francisco, CA, June 14, 2010

Indoor-Outdoor Classification with Human Accuracies: Image or Edge Gist?

Christina Pavlopoulou
Computer Science Department
Boston College
Chestnut Hill, MA 02467
pavlo@cs.bc.edu

Stella X. Yu
Computer Science Department
Boston College
Chestnut Hill, MA 02467
syu@cs.bc.edu

Abstract

We investigate the utility of human performance data on indoor-outdoor scene categorization in improving the generalization performance of a machine indoor-outdoor classifier. On 50 indoor and 50 outdoor scenes, the human categorization accuracies are obtained for these stimuli rendered as either real images or line drawings. We study two types of features, image gist and edge gist, which are the scene gist features extracted from the original image and the edge map of the image respectively. Using human accuracies on real images and line drawings as constraints on these two sets of features in training a max-margin classifier, we observe 4% improvement in classifying never seen 10000 indoor and 10000 outdoor images. Our experiments also reveal that edge gist characterizes indoor scenes far better than image gist. Therefore, not only human labeling is necessary for machine classification, but how humans err on the labeling is instrumental for learning better generalizing features and machine classifiers.

1. Introduction

Machine recognition tasks are often cast as classification problems: given a set of training exemplars for which categorical labels have been provided by humans, the goal is to learn a classifier which not only performs well on the training data but on unseen data as well.

One of the biggest challenge in machine classification is the ability to generalize, that is, the ability to achieve good performance on test data which are significantly different from the training data. There are three main approaches: employing ever larger training datasets, incorporating more priors, and utilizing properties inherent in the test data.

The first type of approaches expands the training data set so that it has more potential to contain representatives for all possible types of images, allowing a classifier to interpolate rather than generalize well on new test data ([5, 12, 15]). The ubiquity of internet and emergence of frameworks such as the Amazon Mechanical Turk have made fast and cheap data collection and annotation possible.

The second type of approaches aims to use previously learned knowledge to improve performance on novel tasks, to learn properties of one object that can be used to make inferences about other objects, to acquire and organize information autonomously. They can be effective in complex classification tasks with few training examples ([8, 3]).

The third type of approaches focuses on identifying data properties that can help constructing good classifiers without requiring volumes of labeled data. Semi-supervised learning ([4]) aims to reduces the number of labeled samples required by taking into account the separability the unlabeled data might exhibit. Active learning ([13]) aims to select intelligently the most informative examples to label.

We propose a fourth approach, using human performance data on a visual task to construct a more generalizing machine classifier. Our idea is that, while the human labeling of the training data indicates the *outcome* of human classification, the human performance on a visual task with controlled stimulus presentation gives out clues to the *process* of human classification. Since the human visual system utilizes features and decisions that work for general images, its performance data provide additional constraints on what features have better generalization potentials.

In this paper, we focus on the usefulness of human accuracies obtained on an indoors vs. outdoors task in which the scenes are presented as grayscale images or detailed line drawings over an increasing exposure time. There are only 50 indoor and 50 outdoor scenes. These 100 images become our training dataset for machine classification. We incorporate human accuracy information in a max-margin framework in conjunction with two types of features: image gist and edge gist, which are the scene gist features extracted from the original image and the edge map of the image. We test the classifier on a large number of unseen images, 10000 indoors and 10000 outdoors. We observe that human performance can boost the performance of the

(a) 75% → 100% (b) 90% → 100% (c) 95% → 100% (d) 60% → 74%

(e) 80% → 100% (f) 100% → 100% (g) 95% → 100% (h) 80% → 89%

Figure 1. Example stimuli employed in our scene categorization experiment. The stimuli were displayed in two rendering modes: grayscale images and detailed line drawings. The percentages under the images indicate the categorization accuracy when the stimulus is displayed 16ms and 32ms. That is, $a\% \rightarrow b\%$ means that the stimulus was categorized $a\%$ and $b\%$ when displayed 16ms and 32ms respectively. Line drawings do not contain any shading cues and all the edges are of the same intensity.

classifier by 4% and edge gist characterizes indoor scenes better than image gist.

We will present our human vision experiment in Section 2, max-margin machine classifier in Section 3, features and parameters in Section 4, experimental evaluation in Section 5, and conclusions in Section 6. Our work shows that not only human labeling is necessary for machine classification, but how humans err on the labeling is instrumental for learning a better generalizing machine classifier.

2. Indoor-Outdoor Categorization by Humans

We conducted an ultra-rapid scene categorization experiment to investigate what features are employed by the human visual system in indoor-outdoor classification [16].

Our stimuli consisted of 50 indoor and 50 outdoor images, collected from the internet with the criterion that each image has a relative large field of view and typical scene complexity. Our artist created detailed line drawings from these images using a tablet in a tracing paper mode. The image and its corresponding line drawing are spatially in correspondence yet the details may be enhanced or omitted according to artistic choices. There were no calligraphic lines for depicting local shading (Fig. 1).

Each experimental trial began with a 2-second display of a blank screen. A fixation dot of radius 0.5° was subsequently shown at the center of the display for 1 second, prompting the subject to gaze at the center. The stimulus, extended 8° horizontally, was presented briefly for either 16 or 32 ms. A choice screen subsequently appeared with the words indoor on the left and outdoor on the right. The subject was required to respond as soon as possible by pressing a designated left of right key. A blank image indicated the start of the next trial. The grayscale images and line drawings were run in two separate blocks of 100 trials each, with one image per trial. The trials were completely randomized for each subject, and the block ordering was also randomized and balanced across subjects.

There were 31 participants in the experiment. The average accuracies over all subjects and stimuli were:

average accuracy	16 ms	32 ms
grayscale images	90.6%	97.8%
line drawings	89.5%	93.0%

These results suggest that highly discriminative features can be extracted from line drawings as well as grayscale images. Shading and texture cues present in real images and absent in line drawing counterparts have an advantage with more stimulus exposure.

More interestingly, different rendering modes lead to different performance for indoor and outdoor scenes. Fig. 2 shows the distribution of accuracies among 100 stimuli. While additional exposure time helps improve the accuracy of almost all grayscale images, it may not help some line drawings. Fig. 1 shows example stimuli and the corresponding categorization accuracies. Grayscale images are preferable for uncluttered scenes (Fig. 1b,d,f,h), whereas

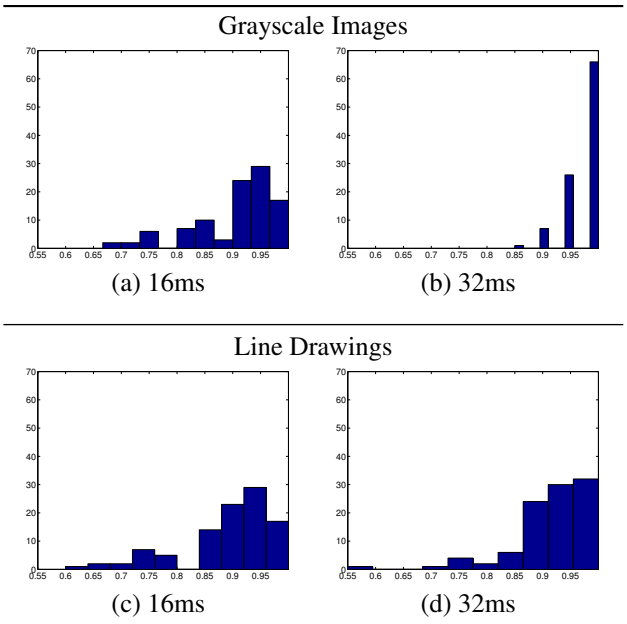


Figure 2. Histograms of per-stimulus average accuracies for grayscale and line drawing rendering modes. Additional exposure time is especially beneficial for grayscale images.

line drawings are preferable when illumination conditions create strong edges that confuse the subjects (Fig. 1a,c,e,g).

3. Max-margin Formulation

Support Vector Machines (SVM's, [2]) are some of the most effective classification methods. Their goal is to compute the decision boundary that minimizes the misclassifications and is maximally distanced from the training examples (Fig. 3(a)). It is formulated as a constraint satisfaction program. If 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:

$$\begin{array}{lrcl} \min & \sum_i \xi_i + \frac{\lambda}{2}||\mathbf{w}|| & & \\ \text{s.t.} & y_i\ (\mathbf{x_i} \cdot \mathbf{w} + b) & \geq & 1 - \xi_i \\ & \xi_i & \geq & 0 \end{array} \quad (1)$$

The slack variables ξ_i relax the constraints when the data are not linearly separable. The norm $||\mathbf{w}||$ is usually chosen as L_2 or L_1. In the latter case the above formulation results in a linear program and it is known to encourage sparsity of features [1, 9]. The decision boundary is determined by certain points in its vicinity, referred to as support vectors. Informally speaking, the support vectors define the area in the feature space with the hardest to categorize points.

Ultimately the success of a classifier is determined by how it performs on unseen data. The generalization performance can be compromised if the training set is small, if it contains non-representative examples, or if it is significantly different from the test set. Our goal is to employ the human accuracies so that the resulting classifier has improved generalization performance.

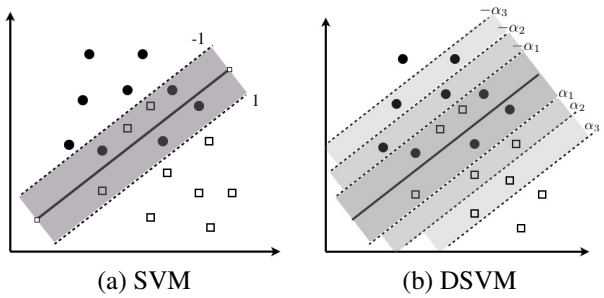


Figure 3. Solid circles and squares denote the training data from two classes. Solid line indicates the decision boundary computed. (a) SVM minimizes the number of misclassifications, when the points cannot be linearly separated. Misclassifications occurring far from the boundary are penalized more. (b) DSVM additionally penalizes points which do not fall behind pre-specified hyperplanes (dashed lines); it constraints the training set configuration.

The average accuracy a scene is categorized by subjects provides a measure of the complexity of a scene. In the ideal human feature space, scenes very accurately categorized should be far from the human decision boundary, whereas scenes not so accurately categorized close it. In other words, human accuracies can provide "prior" information regarding the support vectors of a human SVM classifier. Incorporation of such information can result in improved generalization performance.

We interpret human accuracies as a form of distance from an ideal decision boundary and develop the following so-called Distance SVM (DSVM):

$$\begin{array}{lrcl} \min & \sum_i \xi_i + \frac{\lambda}{2}||\mathbf{w}|| & & \\ \text{s.t.} & y_i\ (\mathbf{x_i} \cdot \mathbf{w} + b) & \geq & \alpha_i - \xi_i \\ & \xi_i & \geq & 0 \end{array} \quad (2)$$

Whereas SVM penalizes only misclassified points, DSVM additionally penalizes points whose algebraic distance from the boundary is more than the human-derived one. As such it imposes constraints on the configuration of the points belonging to the same class (Fig. 3(b)).

Our formulation is similar in form but different in goal to previously introduced SVM formulations. Support vector regression (SVR) [14] estimates the function that best fits the data α_i. However, it does not take into account the discrete category of a data point and may impair the classification accuracy. Soft-SVM [7] interprets α_i's as uncertainties regarding the class of data points. The objective function is modified accordingly in order to focus more on points with small uncertainty. In our case, there is no un-

certainty regarding the class of a data point and the human accuracies affect only the constraints of the formulation not the objective criterion.

4. Features and Parameters

For every training example, we compute the scene gist features [10] on the real image and its edge map. We call them image gist and edge gist features respectively. Whereas image gist has been routinely used in scene categorization and image retrieval, edge gist is our new addition to the family of holistic image representations. Our motivation for this type of features is the ability of humans to very accurately categorize a scene based on its line drawing. The gist features are defined using multiscale Gabor filters as in [10]. In total, there are 512 features computed across 4 scales, 8 orientations and 16 image sites. Edge gist features are computed in exactly the same routine but from the edge map of the image found with the Canny edge detector.

Important parameters of DSVM are the human-induced α_i's. Training exemplars with small α_i's will be encouraged to become support vectors. Such exemplars will in turn favor features for which the margin is maximized. The introduction of these parameters is advantageous if they express the distance from the optimal decision boundary *in a given feature space*. Although the feature space employed by humans is not known, the first features computed in the visual cortex are edge orientations [6] akin to those produced by Gabor filtering [10].

We will restrict ourselves to low-level features like gist [10] and we will define the α_i's based on the accuracy gain g_i, i.e. the absolute difference between the categorization accuracies at 16ms and 32ms. As the exposure time increases, the computation of intermediate level features becomes more likely. Scenes categorized very accurately at both 16ms and 32ms exposures are more likely to be accurately characterized by low-level features.

Because the differences among subject's accuracies are very small we exponentiate the accuracy gains g_i:

$$\alpha_i = e^{\gamma\, g_i} \tag{3}$$

We refer to γ as *gain enhancing parameter*. The highest the value of γ the more pronounced the effect of high accuracy gain examples on the decision boundary, that is, complex scenes difficult to be categorized by humans. For $\gamma = 0$ we obtain the standard SVM formulation (Eqn. 1).

5. Experimental Evaluation

Our goal is two-fold: establish whether human accuracy can improve the classification performance and investigate the roles of image gist and edge gist features for indoor and outdoor scene classification.

5.1. Training and Testing Data

The training set consisted of the 50 indoors and 50 outdoors scenes used in our human vision experiment. Our test set consisted of and exhaustive 10000 outdoor and 10000 indoor images. The outdoor images were selected from the LabelMe database [12] so that they depicted urban scenes with clearly visible layout. The indoor set was assembled as follows: 8660 from the indoors database available online [11], 340 from [17], and the remaining 1000 from Flickr. The indoor images depicted various spaces (airport, kitchen, bedroom, railway) for which the layout was clearly visibly and no single object was the main focus of the image. The sheer size of the datasets guarantees a large range of variation that would not be covered by our small training set.

5.2. Classification Results

We trained our DSVM method (Eqn. 2) using L_1 norm and γ values ranging from 0 to 7. The larger the values of γ the more influenced is the decision boundary by the human accuracies. For $\gamma = 0$ we obtain the original SVM formulation 1. Two scenarios were employed: image gist features in conjunction with the grayscale image accuracies and edge gist features in conjunction with the line drawing accuracies. The regularization parameter λ was set to 0.02 for all cases. Fig. 4 shows the classification results obtained with respect to the gain enhancing parameter γ.

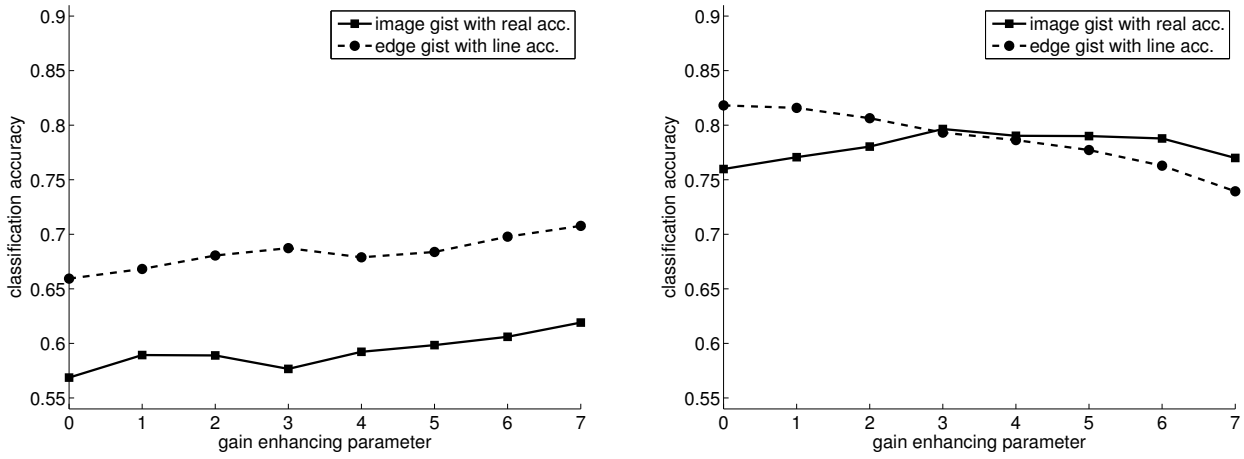
Edge gist features result in higher classification accuracy especially for the indoor scenes. In this case, the gap between edge gist and image gist classification remains the same, more than 10%, for all γ values. For the outdoor scenes, the classification rate is higher for edge gist than image gist features for $\gamma < 4$. For larger values of γ human accuracies in conjunction with image gist features give the best results.

Human performance data improve the classification rate for the case of image gist features and grayscale image accuracies. The classification accuracy increases monotonically with γ for both indoors and outdoors. The best performance in achieved for $\gamma = 6$, where indoor categorization improves by about 4% and outdoor by 3%.

For edge gist features and line drawing accuracies, the classification rate increases with γ for indoor images but decreases for the outdoor images. This could be because the line drawings employed in the human vision experiment contain different information than that in the edge maps obtained by the Canny edge detector.

5.3. Image Gist vs. Edge Gist

Edge gist features are particularly beneficial for the indoor class (Fig. 4). Indoor scenes consist of complex illumination phenomena: natural lighting coming through windows, artificial lighting from various sources, multiple sur-



(a) indoors (b) outdoors

Figure 4. Classification accuracies for indoor and outdoor scenes when when image gist features are used in conjunction with the accuracies obtained from the grayscale images and edge gist features used used in conjunction with line accuracies. The accuracies are plotted with respect to the gain enhancing parameter γ. Edge gist features are more effective for classification especially for the indoor class. Human accuracies from grayscale images improve the classification accuracy when image gist features are used; for the outdoor class they perform better than edge gist features.

Figure 5. Indoor and outdoor images correctly classified when using edge gist features and incorrectly when using image gist features. The indoor images contain complex illumination conditions and a lot of clutter. For the outdoor cases, edge detection may help enhance the structure of the scene.

faces with different reflection properties. Further, they can be very cluttered. Edge gist features may be more robust to such drastic changes in appearance than the image gist ones and result in improved classification rates.

5.4. Features Selected by SVM and DSVM

The improvement in classification when the image gist is used in conjunction with human accuracies stems from differences in the features selected by SVM and DSVM.

We visualize these features in Fig. 6 for $\gamma = 6$, where the increase in accuracy is maximum. The features characterizing the indoor (outdoor) scenes correspond to the positive (negative) decision boundary weights. The weights for both indoor and outdoor classes were averaged over all image sites and normalized to (0,1) (the output weights were converted to positive). The length of the arrows in Fig. 6 is proportional to the corresponding weight, the direction corresponds to the orientation of the Gabor filter and the thickness of the line to the scale. From left to right, the scale becomes more coarse.

Fig. 6 shows that while indoor scenes are categorized by similar features by both SVM and DSVM, outdoor scenes are categorized by finer features and more vertical orientations by DSVM. This limits the interference between the coarse scales employed for the indoor scenes and the finer ones employed for the outdoor scenes.

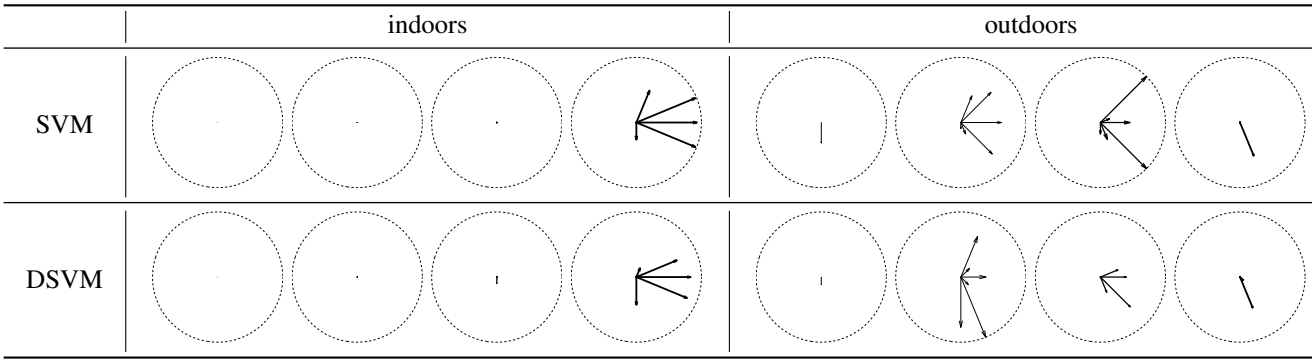


Figure 6. Features selected by SVM and DSVM for image gist features and $\gamma = 6$. The length of the arrows is proportional to the feature weight. Bolder lines correspond to features in coarser scales (left to right). Indoor scenes are similarly characterized by both SVM and DSVM. Outdoor scenes are characterized with finer scale features by DSVM than by SVM and more emphasis on the vertical directions.

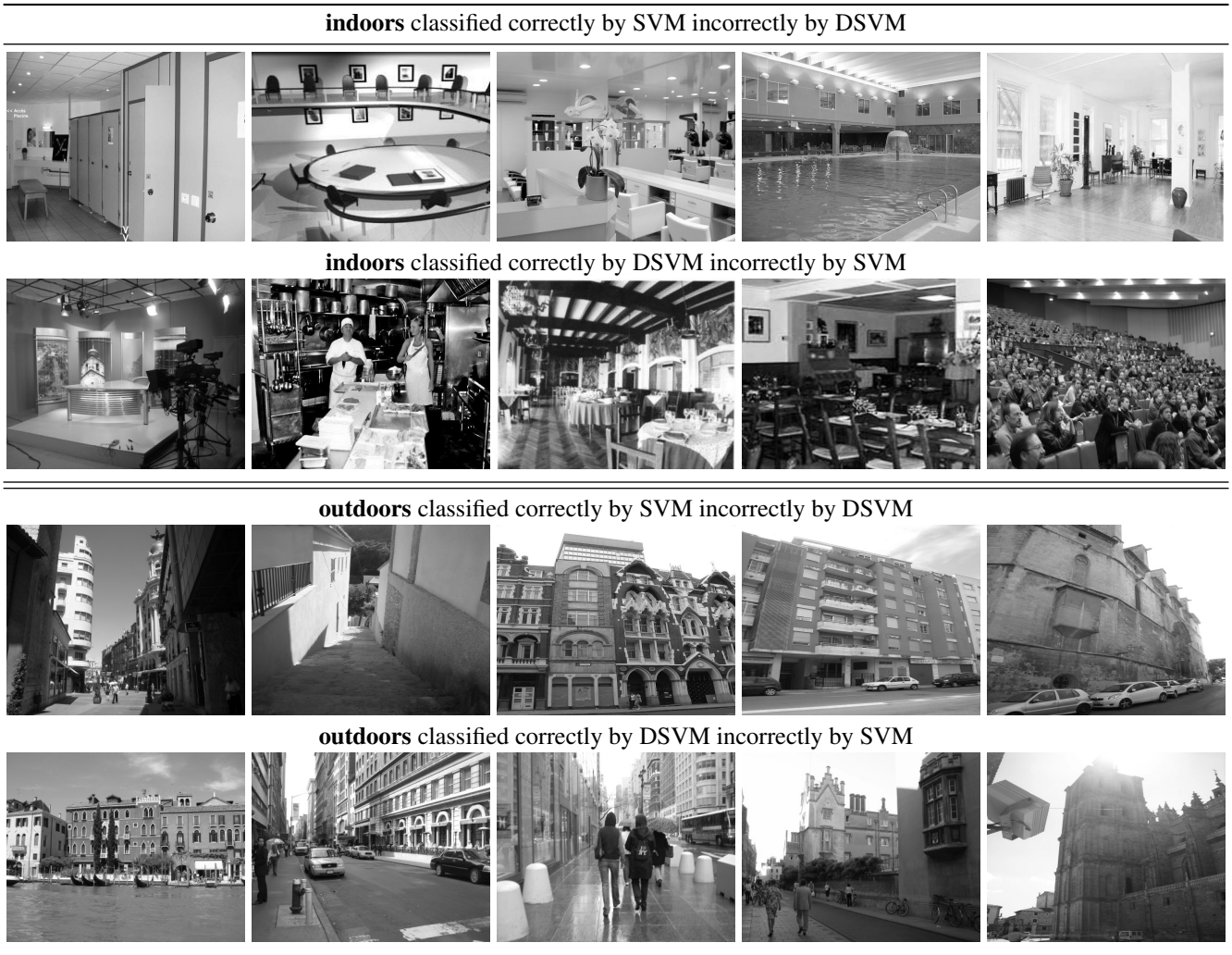


Figure 7. Sample indoor and outdoor test scenes where the SVM and DSVM classifiers produce different results. DSVM favors fine and vertical features for the outdoor class and as a result coarser features for the indoor class. Thus, DSVM can categorize more complex and cluttered indoor scenes than SVM, whereas SVM selects diagonal features and works on outdoor images with these types of distortions.

Consequently, as shown in Fig. 7, DSVM can categorize more cluttered indoor scenes than SVM, and it can accurately categorize outdoor scenes with prominent vertical directions.

6. Conclusions

We investigated the use of human performance data in improving indoor-outdoor scene classification with two types of features: image gist and edge gist.

Our experiments on very large datasets of indoor and outdoor scenes showed that human accuracies obtained from the categorization of grayscale images along with image gist features result in improved classification performance. Additionally, by a large margin, edge gist characterizes indoor scenes better than image gist.

Our work shows that not only human labeling is necessary for machine classification, but how humans err on the labeling is instrumental for learning better generalizing features and machine classifiers.

Acknowledgements

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

References

[1] P. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proc. of the Int'l Conf. on Machine Learing*, pages 82–90. Morgan Kaufmann, 1998. 3

[2] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995. 3

[3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006. 1

[4] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised Learning in Gigantic Image Collections. In *Advances in Neural Information Processing Systems*, 2009. 1

[5] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)*, 26(3), 2007. 1

[6] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J Physiol.*, 148:574–91, 1959. 4

[7] Y. Liu and Y. F. Zheng. Soft SVM and its application in video-object extraction. *Signal Processing, IEEE Transactions on*, 55(7):3272–3282, 2007. 3

[8] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. *Computer Vision and Pattern Recognition*, 2000. 1

[9] J. Neumann, C. Schnrr, and G. Steidl. Combined SVM-based Feature Selection and Classification. *Machine Learning*, 61(1-3), 2005. 3

[10] A. Oliva and A. Torralba. Modeling the shape of a scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–75, 2001. 4

[11] A. Quattoni and A. Torralba. Recognizing indoor scenes. *Computer Vision and Pattern Recognition*, 2009. 4

[12] 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*, 77(1):157–173, 2008. 1, 4

[13] V. S., J. P., and G. K. Far-Sighted Active Learning on a Budget for Image and Video Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA*, 2010. 1

[14] A. J. Smola and B. Schlkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14:199–222, 2004. 3

[15] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. on Patt. Anal. and Machine Intell.*, 30, 2008. 1

[16] M. Woods, C. Pavlopoulou, and S. X. Yu. Rapid categorization of spatial layout in real images and line drawings. In preparation, 2010. 2

[17] 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. 4