The Role of Features, Algorithms and Data in Visual Recognition

Devi Parikh
Toyota Technological Institute, Chicago (TTIC)
dparikh@ttic.edu

C. Lawrence Zitnick
Microsoft Research, Redmond
larryz@microsoft.com

Abstract

There are many computer vision algorithms developed for visual (scene and object) recognition. Some systems focus on involved learning algorithms, some leverage millions of training images, and some systems focus on modeling relevant information (features) with the goal of effective recognition. However, none of these systems come close to human capabilities. If we study human responses on similar problems we could gain insight into which of the three factors (1) learning algorithm (2) amount of training data and (3) features is critical to humans' superior performance.

In this work we take a small step towards this goal by performing a series of human studies and machine experiments. We find no evidence that human pattern matching algorithms are better than standard machine learning algorithms. Moreover, we find that humans don't leverage increased amounts of training data. Through statistical analysis on the machine experiments and supporting human studies, we find that the main factor impacting accuracies is the choice of features.

1. Introduction

Many computer vision approaches address the problem of visual recognition *i.e.* scene and object categorization. These approaches often focus on one of three different aspects of the recognition problem: the amount of training data, the learning algorithm or the feature representations.

Some recent approaches leverage large amounts of training data. For instance, Torralba *et al.* [1] collect a dataset of 80 million low resolution images and demonstrate the effectiveness of simple nearest neighbor classifiers for some recognition tasks such as pedestrian detection. Russel *et al.* [2] have an online-annotation tool LabelMe containing thousands of images annotated with a few thousand object categories. Among other things, this dataset has been used to transfer object and region labels to previously unseen images [3, 4]. Deng *et al.* [5] constructed a large image ontology, containing hundreds of exemplar images for every node in parts of the WordNet hierarchy.

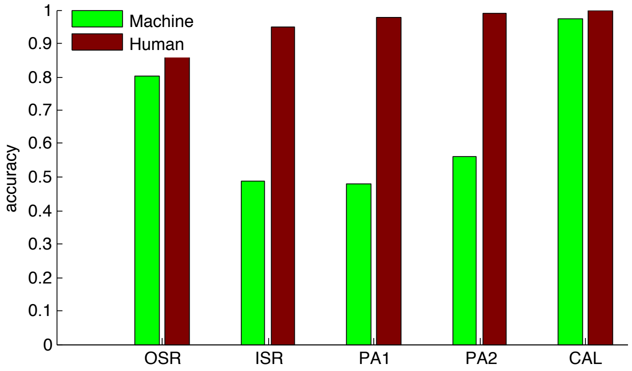


Figure 1: A comparison of human accuracies and machines accuracies on visual (scene and object) recognition tasks. We can see that for contrived datasets such as Caltech-6 (CAL) [34] and outdoor scene recognition (OSR) [18], machine performance is close to human performance. However, for more realistic and relatively newer datasets such as the indoor scene recognition (ISR) [10], and PASCAL (PA1 and PA2) [33] datasets, humans significantly outperform machines. [1]

Many approaches develop advanced learning algorithms. Fei-Fei *et al.* [6] learn object categories from small amounts of training data. Optimal combinations of features types are learned in Varma *et al.* [9]. Quattoni *et al.* [10] develop an exemplar based prototype learning approach to leverage contextual information between scenes and objects, while Li *et al.* [11] use graphical models. Relationships between objects are learned using a discriminative approach in Desai *et al.* [12], and Felzenszwalb *et al.* [13] use latent-SVMs to model relationships between the parts of objects. Other learning algorithms such as PLSA [14], boosting [15], decision trees [16], and boosted random fields [17] have also been applied to recognition and detection tasks.

Other approaches design effective feature representa-

[1] The machine accuracies are reported using the best combination among standard feature and classifiers. State-of-the-art accuracies on these datasets may be higher, but are still not in the ball-park of human performance.

tions. Gist [18] captures the global layout of the scene by describing the spatial distribution of textures for scene recognition. Bag-of-words [19] approaches use the frequency of occurrence of low-level patch appearances, while spatial pyramid matching [20] incorporates additional spatial information. Groups of contours have been used for shape matching [21], and histograms of oriented gradients [22] have been used for object detection [13, 22].

Inspite of the wide variety of approaches proposed for visual (scene and object) recognition, none of these existing approaches to visual recognition perform as well as humans do on difficult tasks. For example see Figure 1. Humans significantly outperform machines on the hardest and most realistic datasets.

What makes humans so much better at these tasks than today's machines? Is it that humans have better pattern matching or learning capabilities than machines? Or that humans have had access to significantly more training data? Or that humans are much better at extracting meaningful features that make the classification task straightforward? If we could "reverse engineer humans" and narrow-down the factors that make humans better at these tasks than machines, this could provide us insightful guidance as to which aspects we should focus our attention on to advance the field of visual recognition.

In this paper we take a small step towards this goal. We perform a variety of human studies and machine experiments to examine whether human learning and pattern matching algorithms are significantly better than some of today's popular classification strategies. The amount of training data and feature types are varied, and their affects on accuracies studied. While many pattern machining algorithms perform both feature extraction and classification, we study these two aspects separately since feature extraction is largely concerned with dimensionality reduction and classification with labeling. In human experiments, the same experiments are performed with the features abstracted to remove any bias due to a human's prior knowledge.

In our experiments we find no evidence that the human learning algorithm is better than standard machine learning algorithms popular today. Moreover we find that humans don't aggressively leverage more training data. As a result, we hypothesize feature representation as the factor that gives humans an advantage over machines. In fact, through multi-way analysis of variance (ANOVA), we find that the choice of features impacts the machine accuracy the most as compared to the other factors.

The rest of this paper is organized as follows. We discuss related work in Section 2. In Sections 3.1 and 3.2 we present the machine experiments and human studies we performed. In Sections 4, 5 and 6 we present results on human and machine experiments varying algorithms, amount of training data and features respectively. We bring up some points of discussion in Section 7 and conclude the paper in Section 8.

2. Related work

Many previous works have studied humans in the hope of gaining insight into the recognition problem. David Marr's book [23] is an early example of studying humans to design computational models with similar behavior. Liu *et al.* [24] conducted human studies and developed a Bayesian model to demonstrate that the high human performance in 3D object discrimination can only be explained if humans are using 3D information. Tarr *et al.* [25] and Hinton *et al.* [26] study whether humans use mental rotation for recognition and determining if shapes have the same handiness[2] . Humans can also learn and recognize novel objects even if camouflaged [27]. A comparison of human and machine algorithms for selecting regions-of-interest in images was conducted by Privitera *et al.* [28]. Fei-Fei *et al.* [29] show that humans can recognize natural scenes rapidly while being distracted by another demanding task. They also demonstrated that human subjects can provide a large amount of detailed information about the scene and objects present in the image after viewing it for a very brief period of time [30]. Bacham *et al.* [31] show that humans can reliably recognize faces in images as small as 16×16 pixels, and Oliva *et al.* [32] present similar results for scene recognition. Torralba *et al.* [1] and Parikh *et al.* [7] show that humans can detect objects in 32×32 images with significantly higher performance than state-of-the-art machine algorithms using high resolution images. Apart from visual recognition, Wolf *et al.* [8] demonstrate through human studies that a simple bag-of-words model without syntactic information is adequate for humans to classify text documents. Our goal in this work is to explore specifically which factors are critical to superior human performance in visual recognition.

3. Experimental Setup

We present identical learning tasks *i.e.* the same feature representation and the same training data, to machines and humans, thus allowing us to draw comparisons between the two. We first describe our machine experiments, and then the corresponding human studies.

3.1. Machine Experiments

We perform a wide variety of machine experiments using various standard classification algorithms for several scene and object recognition tasks using different datasets and feature types. We vary the underlying learning problem by varying the number of training instances, the dimensionality

[2] Updated from original version of paper in proceedings

of the feature vectors, and the proportion of noisy features added to the data. All combinations of the following settings are used in our machine experiments unless otherwise stated.

Algorithms: For all our experiments, we used the following 10 different classifiers. NN: nearest neighbor, NCM: nearest class-mean, LSVM: linear SVM, QSVM: SVM with a quadratic polynomial kernel, CSVM: SVM with a cubic polynomial kernel, RBFSVM: SVM with an Radial Basis Function (RBF) kernel, DT: decision tree, NET: a multi-layer perceptron neural network with 1 hidden layer and 20 hidden layer nodes, BOOST: boosting with linear SVM on individual features as the simple learners, LDASVM: Principal Component Analysis (PCA) then Linear Discriminant Analysis (LDA) followed by a linear SVM classifier.

Datasets: We experimented with the following 5 datasets: OSR: eight categories (coast, forest, highway, inside-city, mountain, open-country, street, tall-building) of outdoor scene recognition dataset [18], ISR: eight categories (bathroom, bedroom, dining room, gym, kitchen, living room, movie theater and stairs) from the indoor scene recognition dataset [10], PA1: eight categories (bird, bottle, cat, dog, horse, person, pottedplant, sheep) from the PASCAL object recognition dataset [33], PA2: eight other categories (aeroplane, bicycle, boat, chair, car, diningtable, motorbike, sofa) from the same PASCAL object recognition dataset [33] and CAL: six categories (aeroplane, car-rear, face, ketch, motorbike, watch) from the Caltech-101 object categories dataset [34]. For PA1 and PA2, since we are interested in categorization and not detection, we worked with the isolated bounding boxes for each object, and not the entire image. In each experiment, 12 instances from each class (96 instances for the eight-class OSR, ISR, PA1 and PS2 datasets, and 72 instances for the six-class CAL dataset) were used for testing. Example images from each of the datasets are shown in Figure 2.

Feature types: We experimented with the following 3 features for all datasets. CH: color histogram computed by assigning all pixels in an image to a pre-computed universal color dictionary computed using k-means, TH: texture histogram computed over a discretization of multi-scale edge orientations in the image, and GIST: gist descriptor using code provided by [18]. In addition to these, we also used BOW: bag-of-words feature descriptor for the CAL dataset computed by assigning the SIFT [35] descriptors of detected interest points to a pre-computed dictionary of SIFT codewords. Moreover, for PA1 and PA2, we used ATT: binary attributes of Farhadi *et al*. [36] which indicate whether the objects have certain higher-level attributes such as being round, or furry, or having a head, etc. It should be noted that unlike other features, these are not machine generated features, but are provided by humans, and are especially interesting in the context of this paper.

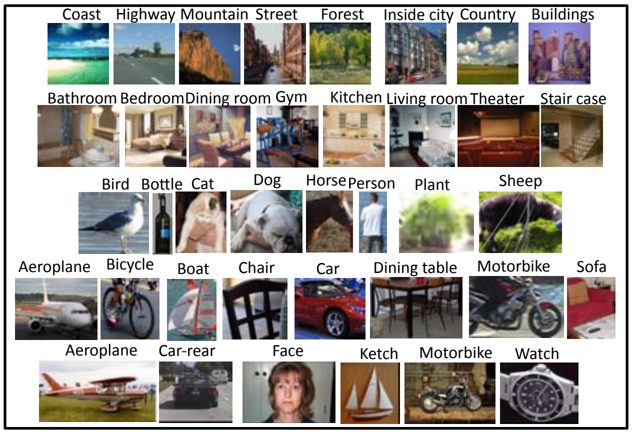
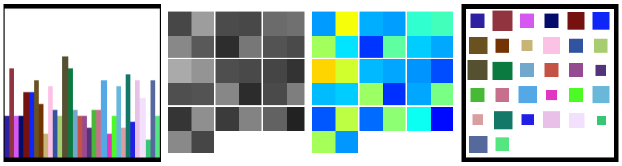


Figure 2: Example images from all datasets used. Top to bottom: OSR, ISR, PA1, PA2, CAL

Dimensionality: We vary the dimensionality of the feature vectors in the range $\{4, 8, 16, 32, 64, 128, 256\}$. For CH, this is done by varying the number of colors in the color-dictionary. For GIST, we can vary the number of edge-orientations (o_{s_i}) at each of the three scales ($[s_1, s_2, s_3]$), as well as the number of spatial blocks ($n \times n$) the image is divided into. To achieve the specified dimensionality, we use the following settings, specified as $[n, o_{s_1}, o_{s_2}, o_{s_3}]$, {[1,4,0,0],[2,2,0,0],[2,4,0,0],[2,4,2,2],[4,4,0,0],[4,4,2,2],[4,8,4,4]}. The dimensionality of TH is varied similarly, except since no spatial information is captured, the image is considered to be a single block. We vary the number of orientation bins across the three scales to obtain descriptors of different dimensionality. Using the same notation as above, the settings used are {[1,2,2,0],[1,4,2,2],[1,8,6,2],[1,16,12,4],[1,32,24,8],[1,64,48,16],[1,128,96,32]}. For BOW the dimensionality was kept fixed at 200 by using a dictionary with 200 SIFT codewords. For ATT, the dimensionality was also kept fixed at 64 for PA2, while PA1 used a 32 bit version in addition to the 64 bit one by dropping the attributes that were almost always set to zero across the dataset.

Proportion of noisy features: We vary the proportion of noisy features added to the original feature vector in the range $\{0\%, 25\%, 50\%, 100\%, 200\%\}$, where 200% indicates that twice the number of original features are added as noisy features. Each entry in the noisy feature is generated randomly using a Gaussian distribution with the same mean and standard deviation as the original clean features.

Number of training instances: For each task, we vary the number of training instances used per category in the range $\{2, 4, 8, 16, 32, 64, 100 (88 \text{ for CAL})\}$.



(a) Colored-bars (b) Intensity Map (c) Heat Map (d) Colored-squares

Figure 3: Example visualizations of a 32 dimensional feature vector. The value of each of the 32 entries in the feature vector is converted to the height of the 32 bars in (a), the intensity of each of the 32 blocks in (b) and (c), and the area of each of the 32 squares in (d). Best viewed in color.

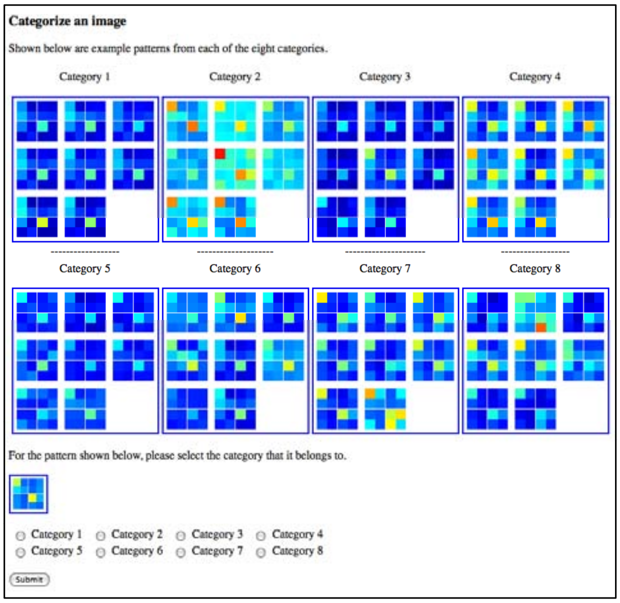
3.2. Human Studies

We perform the human studies on Amazon Mechanical Turk. The use of this service does not allow us to precisely control the experimental conditions of our experiments, such as properties of a subject or the subject's viewing device and environment. However, it does provide a means to carry out a significant number of experiments that would be prohibitively time consuming otherwise. These results should not be viewed as definitive results for average human performance, but as a guide to their abilities. To motivate the subjects to provide correct answers, they were told they would only be paid after their responses were verified.

To prevent the use of prior knowledge about images by the subjects, we do not display to them any direct image information such as texture patches or color. Instead, we use abstracted visual patterns as stimuli. We experimented with four different patterns as shown in Figure 3. We found that for a particular task, subjects performed at 34% for Figure 3(a), 47% for (b), 50% for (c) and 47% for (d). Hence the heat-map visualization was used for the rest of the experiments. We note that the 2D spatial layout of the visualization may introduce notions of "neighboring" features (e.g. adjacent blocks in the pattern) to the subjects, even though there is no such notion in the feature space. This may influence which patterns subjects deem to be similar. As shown in Figure 4, in all our experiments we strived to display the training patterns in a small area (800 × 600 pixels) to remove the need for scrolling.

In order to compare human performance to that of the machines, we mimic a subset of the scenarios used for machine experiments in our human studies. For each scenario, each of the same 96 test images used in the machine experiments were to be classified by 25 subjects.

Since each test subject is allowed to work on an arbitrary subset of the test images, we only consider subjects who labeled at least 20 images. Since the quality of human subjects on Amazon Mechanical Turk varies, the following calibration process was used to obtain accuracies comparable to those obtained by the authors on some pilot tests. The workers are sorted based on accuracy. Performance numbers are reported as the average of the top quartile accuracies on the entire test set.



Categorize an image

Shown below are example patterns from each of the eight categories.

For the pattern shown below, please select the category that it belongs to.

○ Category 1 ○ Category 2 ○ Category 3 ○ Category 4
○ Category 5 ○ Category 6 ○ Category 7 ○ Category 8

Figure 4: Example Amazon Mechanical Turk interface used for the human studies.

4. The Role of Algorithms

Our first set of experiments evaluates various machine learning algorithms with respect to human accuracies. To isolate the effect of changing the learning algorithm, we fix the set of input features and training data for each set of experiments. The humans are shown the same training and test features as machines, using an abstract visualization as discussed in Section 3.2, to remove any bias from prior knowledge.

As shown in Figure 5, machine accuracies vary based on the features and datasets used. However, the relative accuracies of the various techniques are fairly consistent with some exceptions. Linear SVMs outperform simple classifiers such as nearest neighbor and nearest class-mean. This is consistent with recent works that favor the use of SVMs. The complexity of the datasets is also apparent from these results. Even simple techniques such as NN and NCM perform well for the outdoor scene recognition and Caltech 6 datasets.

Surprisingly, the human accuracies on these same datasets are consistently worse than the best machine accuracies. Even simple classifiers such as nearest neighbor performs comparable to human accuracies. As a result, we conclude that the learning algorithm used by humans

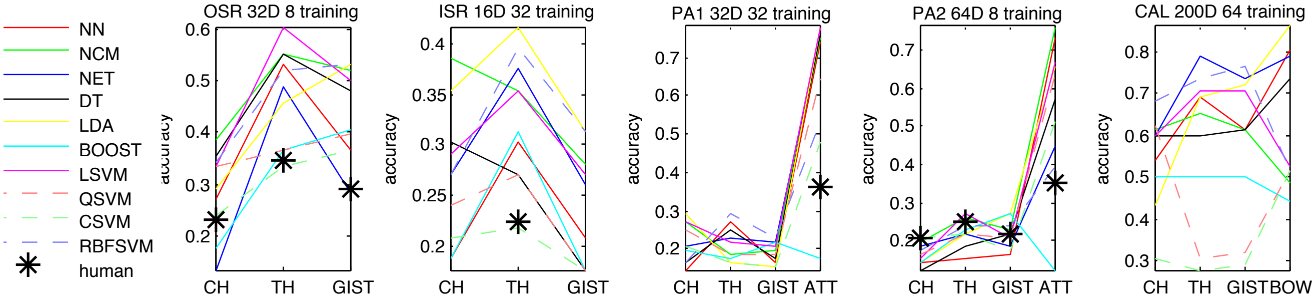


Figure 5: A comparison of the accuracies achieved by various machine algorithms, as compared to humans. We see that human accuracies are consistently lower than the best machine accuracies, and are comparable to some simple classifiers such as nearest neighbor (NN). In all cases, no noise was added to features. Human accuracies are shown for the settings for which human studies were conducted. Best viewed in color.

is not superior to state-of-the-art techniques on these types of problems.

5. The Role of Data

Our next set of experiments isolates the role of training data on accuracies. For these experiments we varied the amount of training data, the noise present in the data as well as the dimensionality of the descriptors as described in Section 3.1.

As seen in Figure 6, the machine experiments show consistent improvement as the number of training instances increase. This supports the works of [2, 5] that addresses the need for larger and more detailed training data to improve accuracies. However, even after extrapolating the results to larger datasets, it is unlikely that machine accuracies will match those of humans when given the original images, Figure 1. Certain algorithms take better advantage of more training data such as SVM classifiers.

Human experiments show a remarkably different trend. The accuracies of humans quickly levels off after only four to sixteen training examples. This indicates that humans may not be as capable at leveraging large amounts of training data for pattern matching. Conversely, it has been noted that humans are very capable of generalizing from a small number of training examples, even when the training data is ambiguous [27]. Since it might take longer for humans to learn a training dataset, we also analyzed the human accuracies as the experiments progressed. We found accuracies did not vary as more questions were answered.

Data noise also plays a role in learning. For machine experiments significantly more training examples are needed to achieve similar levels of accuracy, as shown in the 2D plot of Figure 7, especially for the nearest neighbor classifier. Linear SVMs and NCM are less sensitive to noise as seen in the plot in Figure 7. Humans are also susceptible to data noise.

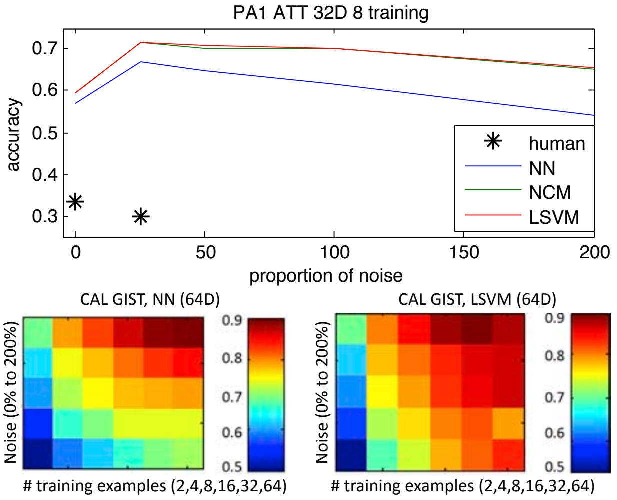


Figure 7: Human and machine accuracies with increasing noise in the data. Human accuracies are shown for the settings for which human studies were conducted. Best viewed in color.

6. The Role of Features

Finally, we address the role of features in visual recognition. Figure 8 shows the accuracies for various features types using different algorithms and datasets along with corresponding human results. Interestingly, edge and gradient based features typically out perform color based features across the various datasets. This supports the large body of works using gradient histograms for recognition [13, 18, 22]. Humans are also known to be very sensitive to edge or contour information. They can recognize objects just from line drawings with very high accuracy [38].

Figure 8 has two sets of experiments involving humans. First, we perform human studies on recognition using the same features and training sets as the machine experiments. The results show similar treads as the machine experiments.

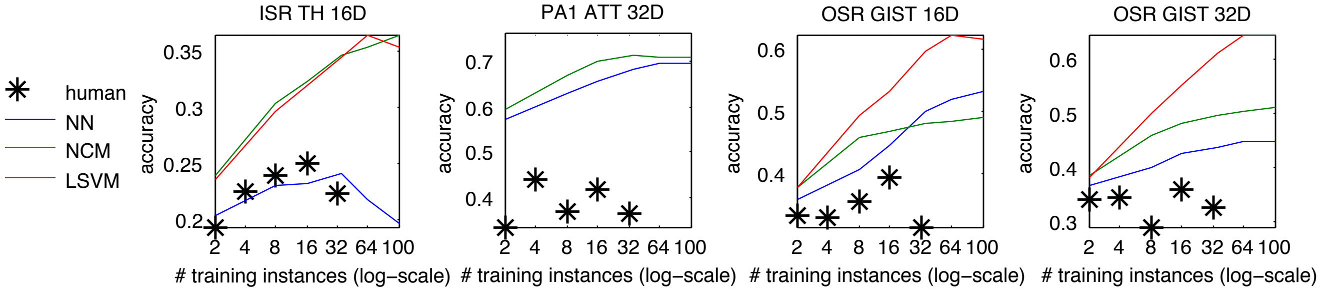


Figure 6: Human accuracies as the number of training examples increase. We can see that humans do not leverage more training data aggressively. Human accuracies are shown for the settings for which human studies were conducted. Best viewed in color

Certain features types, such as Gist and texture histograms, significantly out perform color based features. The second set of experiments use humans to generate the feature set as provide by [36]. This feature set called attributes, as described in Section 3.1 is a set of binary labels describing an object indicating properties such as being round, furry, or having a head. As we can see, these features are much more informative for both machines and humans for recognition on the challenging PASCAL datasets (PA1 and PA2). This illustrates how critical the feature set can be for recognition.

Given the importance of features, we provide a few additional experiments investigating the features used by humans. In these experiments, humans are shown natural images from the outdoor scene recognition dataset [18] under different transformations and asked to select a category name from a list. Since we want humans to use their own pre-built models and internal features we don't provide subjects with any training data.

Expanding on the work of Vogel *et al.* [39], we experiment with two transformations: (1) Block test, where the image is divided into non-overlapping blocks, and the pixels in each block are randomly shuffled. This maintains the global layout of the scene, but the local statistics are lost. (2) Puzzle-test, where the image is again divided into non-overlapping blocks, but the blocks are randomly shuffled in the image while maintaining the pixels' relative locations in the block. In this case, local regions of the image are preserved while the global layout is not. Both these transformations were applied to low resolution (equivalent of) 32×32 images, and high resolution 256×256 images. We see in Figure 9, that in both high and low resolution images, human recognition is robust to a significant loss of local statistics. This indicates that humans rely on the global layout of the scene for scene recognition. This is also supported by tests conducted by Torralba *et al.* [1] and Parikh *et al.* [7] on low resolution images. However, in high resolution images, human recognition rates are also very robust even when the global layout of the scene is drastically altered, which indicates that humans can also rely on local regions of images for scene recognition.

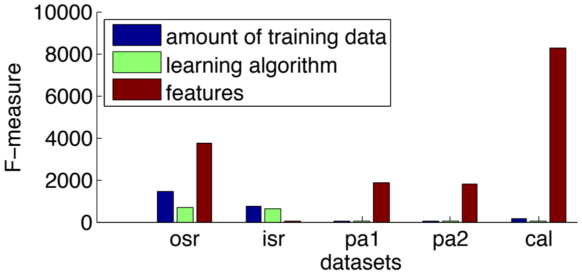


Figure 10: The F-measure computed using analysis of variance (ANOVA). We can see that for all datasets where machines can achieve reasonable recognition accuracy, the features have the highest influence on the performance. For ISR (indoor scene recognition), machines have low recognition rates in all settings.

These tests indicate that humans do not rely on a fixed set of features. Depending on the information available to them, humans can adaptively rely on different sets of features during testing. This is true even if similar instances have never been seen before. This ability to adapt during testing is not seen in standard machine learning algorithms.

7. Discussion

In addition to our qualitative results, we performed a multi-way analysis of variance (ANOVA) on our machine experiments. In Figure 10, we plot the F-measure for each of the three factors. We find that the choice of features impacts the recognition accuracy the most, further supporting our hypothesis. This is especially pronounced for the PASCAL dataset in which the human generated feature attributes are studied.

For lack of space, we did not include details of an additional experiment we conducted to test if human subjects were essentially using a nearest-neighbor classification strategy. In one test, we included some copies of the

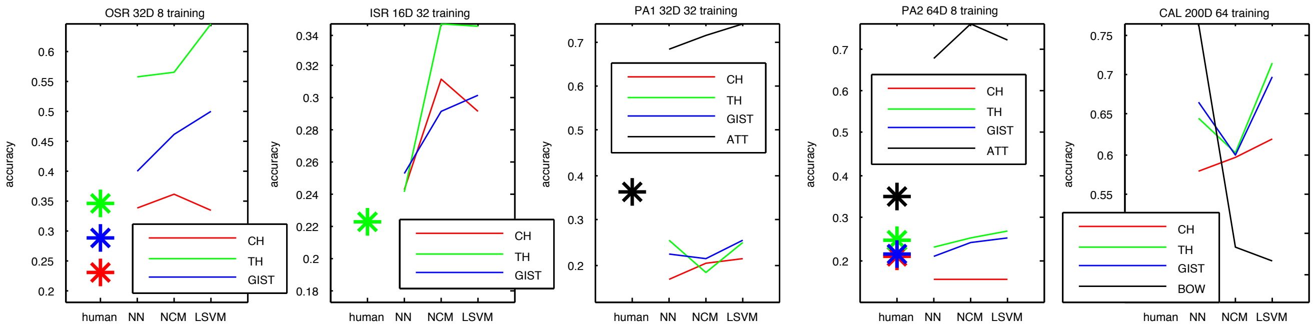


Figure 8: A comparison of machine algorithms and human accuracies with varying feature choices across datasets. Human accuracies are shown for the settings for which human studies were conducted. Best viewed in color.

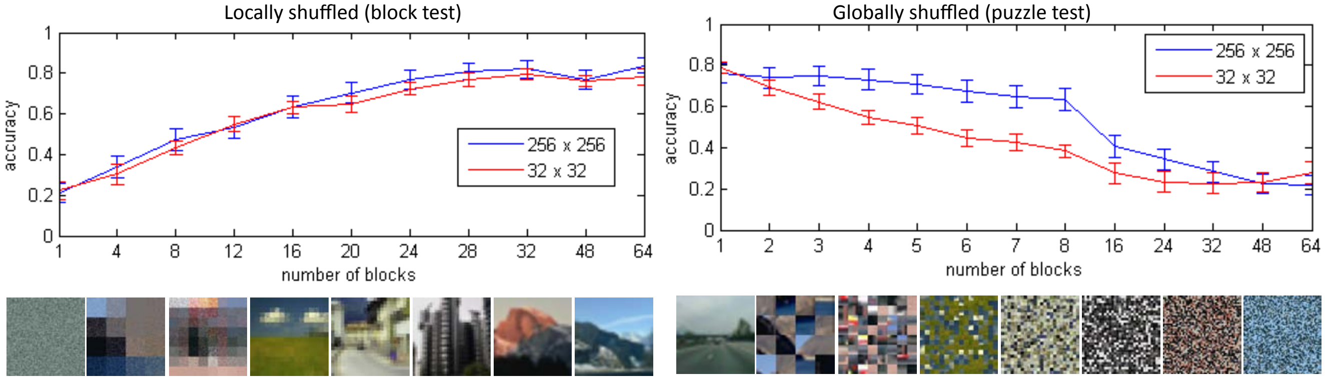


Figure 9: Tests to evaluate human recognition rates on natural images under various transformations. We can see that humans adaptively rely on global layout or local regions of an image for scene recognition. Human performance drops significantly only when both global and local information is impoverished.

training instances in testing. Surprisingly, we found that the accuracy of the human subjects on these repeated instances was not higher than the novel test instances (48%).

The notion of features goes beyond the choice between colors, texture, the need for spatial information, etc. It includes the concepts of incorporating semantic attributes, such as the ones proposed by Farhadi *et al*. [36] and Lampert *et al*. [37] that are shared across categories. Perhaps what makes the human feature representation so powerful is that these feature representations are tuned for high performance at a variety of tasks. Humans solve more visual tasks apart from identifying scene and object categories, such as guessing the functionality of unknown objects, or anticipating the feel of an object before touching it. A word of caution against the human labeled semantic attributes [36, 37] is that their correlation with category labels can be artificially high. For instance, a human subject may label an object as furry only after the subject recognizes it as a cat.

It is important to note that in addition to visual features, humans leverage prior knowledge from several non-visual higher-level and semantic features about how the world we live in functions. In order to thoroughly analyze how much we rely on visual information as compared to this prior information, we would have to design new experiments. An abstract visualization of natural images would be needed that contains the same visual features as natural images, without allowing the human subjects to relate the abstract images to their prior knowledge. Unfortunately, because of the tight coupling of this prior with the visual stimulus, it is exceedingly difficult to design this abstract visualization. If we attempted to abstract real images, how would one display an image of a highway scene? Similar edge gradients, color histograms, etc. would need to be available for extraction, without the image being interpreted as a highway scene. However if this can be done, it would provide us with a plausible upper-bound on machine visual recognition performance when prior or semantic information from non-visual sources is not provided.

8. Conclusion

In this paper we study human responses on visual recognition problems as posed to machines, to gain insight into which of the three factors (1) learning algorithm (2) amount of training data and (3) features is critical to humans' superior performance. We find no evidence that human pattern matching algorithms are better than standard machine learning algorithms. Moreover, we find that humans don't leverage increased amounts of training data. We thus hypothesize with the aid of ANOVA analysis that features are the main factor contributing to superior human performance. Future work involves extensive studies to identify which visual features humans rely on to aid in the development of novel machine recognition algorithms.

References

[1] A. Torralba, R. Fergus and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *PAMI*, 2008.

[2] B. C. Russell, A. Torralba, K. P. Murphy and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 2008.

[3] C. Liu, J. Yuen, A. Torralba, J. Sivic and W. T. Freeman. SIFT flow: dense correspondence across difference scenes. *ECCV*, 2008.

[4] B. C. Russell, A. Torralba, C. Liu, R. Fergus and W. T. Freeman. Object Recognition by Scene Alignment. *NIPS*, 2007.

[5] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR*, 2009.

[6] L. Fei-Fei, R. Fergus and P. Perona. One-Shot learning of object categories. *PAMI*, 2006.

[7] D. Parikh, C. Zitnick and T. Chen. From Appearance to Context-Based Recognition: Dense Labeling in Small Images. *CVPR*, 2008.

[8] F. Wolf, T. Poggio and P. Sinha. Human Document Classification Using Bags of Words. *Tech report*, 2006.

[9] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. *ICCV*, 2007.

[10] A. Quattoni and A. Torralba. Recognizing Indoor Scenes. *CVPR*, 2009.

[11] L. J. Li and L. Fei-Fei. What, where and who? Classifying event by scene and object recognition. *ICCV*, 2007.

[12] C. Desai, D. Ramanan and C. Fowlkes. Discriminative models for multi-class object layout. *ICCV*, 2009.

[13] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models *PAMI* (to appear)

[14] J. Sivic, B. Russell, A. Efros , A. Zisserman, and W. Freeman. Discovering Objects and their Location in Images. *ICCV*, 2005.

[15] J.Shotton, J.Winn, C.Rother and A.Criminisi. TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, 2006.

[16] J. Shotton, M. Johnson and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. *CVPR*, 2008.

[17] A. Torralba, K. Murphy and W. Freeman. Contextual models for object detection using boosted random fields. *NIPS*, 2005.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001

[19] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. *CVPR*, 2005.

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *CVPR*, 2006.

[21] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of Adjacent Contour Segments for Object Detection. *PAMI*, 2008.

[22] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005.

[23] D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman. 1982.

[24] Z. Liu and D. Kersten. 2D observers for human 3D object recognition? *Vision Research*, 1998.

[25] M. J. Tarr and S. Pinker. When Does Human Object Recognition Use a Viewer-Centered Reference Frame? *Psychological Science*, 1990.

[26] G. E. Hinton and L. A. Parsons. Frames of reference and mental imagery. In *J. Long and A. Baddeley, editors, Attention and Performance IX*, 1981.

[27] M. J. Brady and D. Kersten. Bootstrapped learning of novel objects. *Journal of Vision*, 2003.

[28] C. M. Privitera and L. W. Stark. Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *PAMI*, 2000.

[29] L. FeiFei, R. VanRullen, C. Koch and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proc. Nat. Acad. of Sciences*, 2002.

[30] L. FeiFei, A. Iyer, C. Koch and P. Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 2007.

[31] T. Bachmann. Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *Europ. Jour, of Cognitive Psychology*, 1991.

[32] A. Oliva and P. G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 1976.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results.

[34] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR, Workshop on Generative-Model Based Vision*, 2004.

[35] D. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 60(2):91110, 2004.

[36] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by their Attributes. *CVPR*, 2009.

[37] C. H. Lampert, H. Nickisch and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. *CVPR*, 2009.

[38] I. Biederman. Human image understanding: recent research and a theory. *CVGIP*, 1985.

[39] J. Vogel, A. Schwaninger, C. Wallraven and H. Bülthoff. Categorization of Natural Scenes: Local versus Global Information and the Role of Color. *TAP*, 2007.