Recognizing Jumbled Images:
The Role of Local and Global Information in Image Classification

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

Abstract

The performance of current state-of-the-art computer vision algorithms at image classification falls significantly short as compared to human abilities. To reduce this gap, it is important for the community to know what problems to solve, and not just how to solve them. Towards this goal, via the use of jumbled images, we strip apart two widely investigated aspects: local and global information in images, and identify the performance bottleneck.

Interestingly, humans have been shown to reliably recognize jumbled images. The goal of our paper is to determine a functional *model that mimics how humans recognize jumbled images i.e. exploit local information alone, and further evaluate if existing implementations of this computational model suffice to match human performance. Surprisingly, in our series of human studies and machine experiments, we find that a simple bag-of-words based majority-vote-like strategy is an accurate functional model of how humans recognize jumbled images. Moreover, a straightforward machine implementation of this model achieves accuracies similar to human subjects at classifying jumbled images. This indicates that perhaps existing machine vision techniques already leverage local information from images effectively, and future research efforts should be focused on more advanced modeling of global information.*

1. Introduction

Recognizing scenes and objects depicted in an image are central tasks towards the goal of automatic image understanding. Consequently, a vast amount of effort in computer vision and machine learning has been focused on improving machine accuracy at these tasks. Unfortunately, machine visual recognition performance still falls significantly short when compared to human abilities.

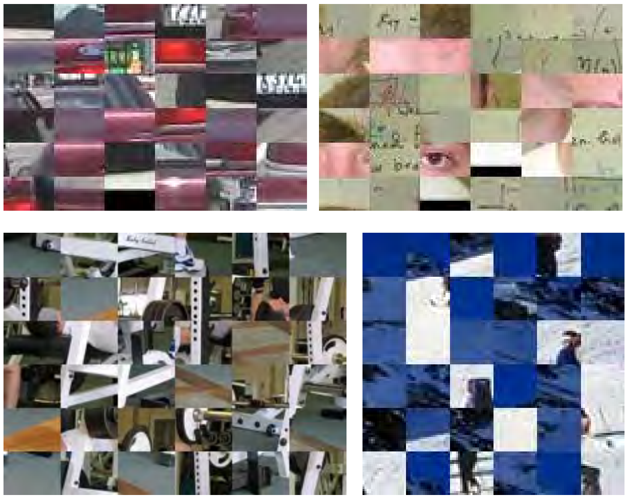


Figure 1: Jumbled images of a car, face, gym and mountain scene. We study human and machine classification of jumbled images, which contain only local information and no global information, to determine which of these two factors needs most improvement in machine vision.

The goal of boosting machine performance at these tasks, and thus reducing this gap between human and machine abilities, can be pursued in one of two ways. The first is to focus on *how* to solve the problems, set the human aside, and optimize different stages of the machine visual recognition pipeline, be it the models, data, features or learning algorithms. This is a valid line of research, after all machines are wired differently than the finely evolved human brains; both with their own sets of strengths, weaknesses and arguably, even goals. This approach has led to significant progress in the field, and in fact most existing works in literature fall in this category.

The other approach, which this paper follows, is to focus on *what* problems to solve, by evaluating if machines have 'mastered' certain aspects of the problem, but are weaker at others. We focus our attention on humans as a working system of what we hope to achieve (*i.e.* mastery at vision), and explicitly incorporate them to constraint the search space of potential research pursuits for optimizing machines. Specifically, by isolating different factors relevant to the recognition problem, we can 'reverse engineer' humans and study how they leverage each of these factors individually. Moreover, by comparing human and machine abilities along each of these factors, we could gain insights into which aspects of the machine recognition pipeline are falling short and need further advancements, and which aspects are effective enough already.

In this paper, we focus on image classification, where an image is to be classified into a scene or object category depicted in the image. We wish to disentangle two widely investigated sources of information in an image: local and global information, and analyze which one of these aspects is in most need for advancements. While different researchers may have an opinion about this, with or without consensus across the board, we believe it is important to give this aspect explicit and formal treatment, towards the goal of scientifically determining which factors are crucial for future progress in visual recognition.

Jumbled images, as shown in Figure 1, provide an appropriate regime for our study, where local information is preserved, but the global layout of the scene is lost. It has been shown that humans can reliably classify jumbled images [1, 2]. Our goal is to identify a functional model that describes *how* humans leverage local information that allows for such reliable classification of jumbled images. When a subject identifies the scene category depicted in a jumbled image, what functional model[1] best mimics the process (s)he employs? Is it a representation that encodes the object category of each block, and describes the distribution of object categories in the image? A computational model mimicking this has been proposed by Vogel *et al.* [1]. Alternatively, are the objects bypassed, each block directly indicates a scene category, and these cues are accumulated via a majority-vote to deduce the scene category of the image, similar to a bag-of-words model [3]? Or is it a complex inference mechanism such as a Markov Random Field (MRF) to perhaps re-assemble the image [4] in an attempt to restore the missing global information? Further, if we were to identify the functional model, are existing machine implementations of the corresponding computational model effective enough to replicate human performance at recognizing jumbled images?

To this end, we conduct a series of human studies. Subjects are asked to classify jumbled images from multiple scene and object recognition datasets, as well as individual blocks in isolation. To our surprise, we find that the simple majority-vote accumulation strategy over the human classification of the individual blocks, predicts the human classification of the entire jumbled image well. We also perform machine experiments with a naive implementation of such a local-block-based majority-vote scheme, and find its predictions to be well correlated with the human responses. This indicates that not only does a simple functional model explain how humans leverage local information, but existing tools in literature can match human performance at classifying images when global information is absent. Hence, the large gap in machine and human abilities at classifying intact images, must stem from an inability of existing computational models to effectively capture global information. We hope this insight better guides future research endeavors in the community.

2. Related Work

Many previous works have studied humans in the hope of gaining insight into the recognition problem. David Marr's book [5] is one of the earliest attempts at studying humans to design computational models with similar behavior. Closest in philosophy to our approach of isolating different factors of the recognition problem is the work of Parikh *et al.* [2, 6], who evaluate the roles of features, algorithms and data for image classification [2], and the relative importance of part detection, spatial modeling and contextual reasoning for object detection [6].

Global: It is well accepted that the global layout of a scene plays a key role in how humans recognize scenes [7], especially when subjects see the images for a short duration of time. Fei-Fei *et al.* [8] show that humans can recognize scenes rapidly even while being distracted, and can provide detailed information about the scene after viewing it briefly [9]. Just as jumbled images contain only local (no global) information, low resolution images contain only global (no local) information. Bachamann *et al.* [10] show that humans can reliably recognize faces in 16×16 images, and Oliva *et al.* [11] present similar results for scene recognition. Torralba *et al.* [12] show that humans can reliably detect objects in 32×32 images. Computational models that effectively capture the global layout of the scene such as the gist [16] and spatial pyramid [17] representations have also been proposed.

Local: Interestingly, it has been demonstrated that humans also leverage local information effectively, allowing them to recognize scene categories even from jumbled images [1, 2] as shown in Figure 1. A large variety of computational models have been proposed, such as the popular bag-of-words model [3] and many variations exploring numerous local features, appearance models, interest-point-detectors, spatial sampling strategies, dictionary formation approaches, classifiers, learning algorithms, etc. – a complete survey of which is beyond the scope of this paper.

While significant research effort has been invested in improving local as well as global models for images, it is not clear which one of these aspects is where existing machine techniques lag as compared to humans, and have the most

[1] We note that our goal is to identify a *functional* model that mimics human *responses*, and not determine a biologically plausible model to explain human behavior (although the former may lead to insights for the latter, and vice versa).

potential for improvement as compared to humans. This forms the focus of our work.

Global vs. Local: The tradeoff between local and global information has been explored for face [13] and object recognition [14]. Parikh *et al.* [15] study the role of appearance (local) and contextual (global) information for image labeling in low and high resolution images. Closest to our work is the work of Vogel *et al.* [1], which explores the role of global vs. local information for scene recognition, by conducting human studies with blurred and jumbled images respectively. They find that both local and global sources of information are exploited. To the best of our knowledge, the question of *how* humans utilize the local information has not been studied thus far, which is a goal of this paper. Vogel *et al.* [1] propose a computation model that first maps each block to an object category, and the distribution of these object categories is used to identify the scene category. One of the goals of this work is to understand if such a complex model is required to predict human responses, or if a simpler model suffices.

Jumbled Images: Jumbled images (Figure 1) have been used before to better understand different aspects of the human recognition system. Biederman studied the impact of jumbling images on human object recognition accuracy given short exposure times [18]. The effect of jumbling on subject response times at the task of object detection is explored in [19]. The impact of varying levels of confusion among the labels on human image classification accuracy, and of varying exposure times on object detection accuracy is studied in intact and jumbled images to understand if humans use holistic information or a sampling of objects [20]. Both are conjectured to be useful. Loftus *et al.* [21] study different types of tasks, where either holistic information or specific features are crucial. They find that at very short exposure times performance based on holistic information was superior, whereas the reverse was true when sufficient study time was available. Giraudet *et al.* [22] use jumbled images to alter the contextual information in the image, and blurring to alter the image quality, to understand the cooperation between the top-down and bottom-up information. Top-down contextual information is important when the bottom-up appearance information is degraded. However, this importance is reduced as the subjects become more familiar with the images. Tjan *et al.* [23] use tile-jumbled and pixel-jumbled images to understand the role of color and local information in priming rapid scene recognition (early scene perception). Yokosawa *et al.* [24] explore how human change-detection performance is affected by various image transformations such as jumbled images, images with missing blocks, etc. In this work we use jumbled images to determine a functional model that mimics how human recognize images using local information alone, and can predict human responses to jumbled images. Since our goal is not to determine a biologically plausible model, aspects like response time of subjects to jumbled images, etc. explored by some of the above works are less relevant to this study.

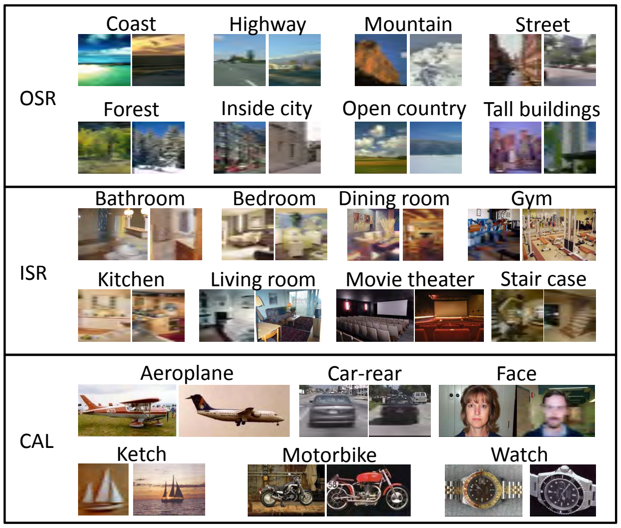


Figure 2: Example images from datasets we experiment with.

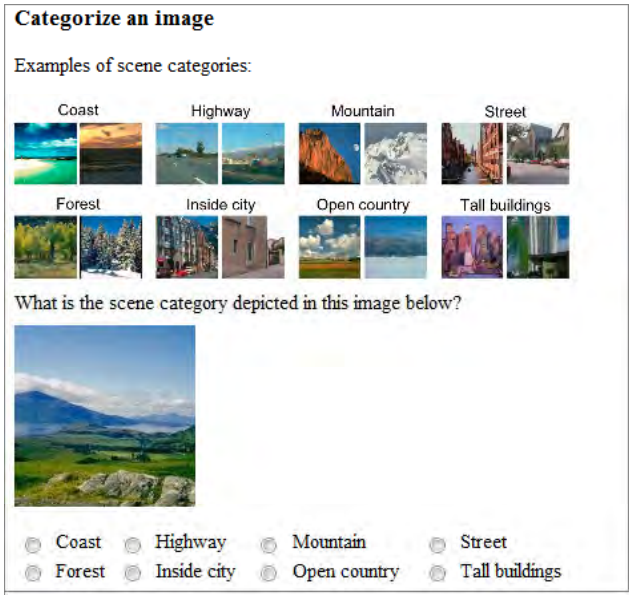
3. Method

To reiterate, we are interested in determining if humans can classify jumbled images because they resort to a complex model incorporating mid-level representations or interactions among the blocks, or if a simple bag-of-words based majority-vote model suffices. The underlying intuition behind how we determine this is as follows: if subjects can recognize the individual blocks even in isolation from the rest of the image reliably enough, such that a majority-vote on these local decisions predicts the label that subjects assign to the jumbled image globally, we can conclude that the simple model is sufficient. Otherwise it is not, and something more complex must be at work.

We perform experiments with the following three datasets: OSR: Outdoor Scene Recognition [16] containing images from coast, forest, highway, inside-city, mountain, open-country, street and tall-building categories; ISR: Indoor Scene Recognition [25] containing images from bathroom, bedroom, dining room, gym, kitchen, living room, theater and staircase categories; and CAL: CALtech object recognition [26] containing images from airplane, car-rear, face, ketch, motorbike and watch categories. We select 50 random images from each category to form our datasets. Example images from these datasets can be seen in Figure 2.

3.1. Human Studies

All human studies were performed on Amazon's Mechanical Turk, an online crowd-sourcing service. In all our studies, each test instance was assigned to 10 subjects. Sub-


Categorize an image

Examples of scene categories:

What is the scene category depicted in this image below?

○ Coast ○ Highway ○ Mountain ○ Street
○ Forest ○ Inside city ○ Open country ○ Tall buildings

Figure 3: Amazon Mechanical Turk interface snapshot.

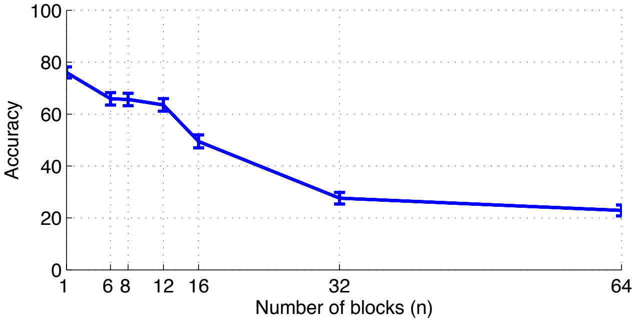


Figure 4: Human recognition accuracy on the OSR dataset on jumbled images with varying number of blocks.

jects were paid half a cent for labeling each test instance. The image was displayed to the subjects along with a list of radio buttons indicating the categories to select from, as seen in Figure 3. There were no constraints on subjects' response times. The images were resized such that the largest dimension was 256 pixels.

3.1.1 Jumbled Images

Jumbled images were created by dividing an image into $n \times n$ non-overlapping blocks, and then randomly shuffling them (see Figure 1). The accuracy of human subjects at classifying jumbled images for varying values of n for the OSR dataset can be seen in Figure 4[2]. It can be seen that beyond $n = 12$ there is a significant drop in accuracy, so we choose $n = 12$ for further experiments. A similar analysis for ISR and CAL led to $n = 6$ and $n = 12$ respectively. We note that the rate at which the curve in Figure 4 falls depends on the classification task at hand. With few and very distinct categories (*e.g.* forrest vs. sunset scenes), human recognition of jumbled images may be high even for very large values of n, while for more fine-grained tasks (*e.g.* classifying different makes and models of cars), the value of n would need to be low. By adaptively picking the value of n where human accuracy transitions from good to bad, we are effectively normalizing for the inherent difficulty of the recognition task, which depends on external factors such as the choice and number of categories in a dataset.

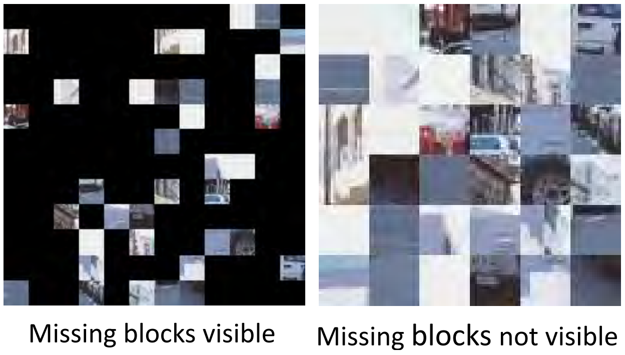


Figure 5: Two different visualizations for retaining only 36 of the 144 blocks in jumbled images.

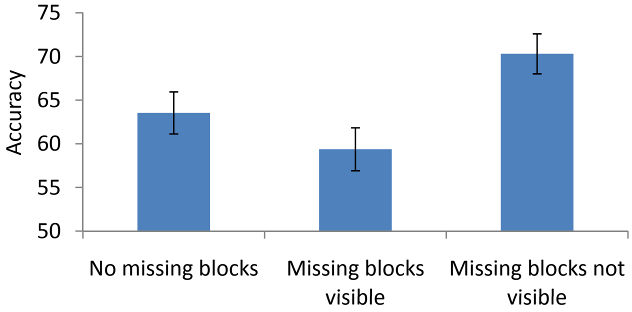


Figure 6: Human recognition accuracy on jumbled images in the OSR dataset based on different visualizations.

Since the eventual study will consist of having 10 subjects classify every block from every image in isolation, it is important for the number of blocks to be classified to not be very large. With $n = 6$ for ISR, each image contained only 36 blocks. But OSR and CAL have 144 blocks per image, which was prohibitively large for our budget. We consider dropping 75% of the blocks from the OSR and CAL images. We do not expect this to affect the human classification accuracy on jumbled images, because most local structures in images are repetitive.

To verify this, we compare human accuracy when using all 144 blocks in an image, to that of using a random subset of 36 out of the 144 blocks. This can be achieved in two ways. The blocks to be removed can be replaced with black pixels (Figure 5 (left)), or the blocks can be dropped, and the retained blocks can be re-arranged into a 6×6 grid (Figure 5 (right)). A comparison of human accuracies on

[2] For this test, all images in the original OSR dataset containing over 2600 images were randomly split across the different values of n.

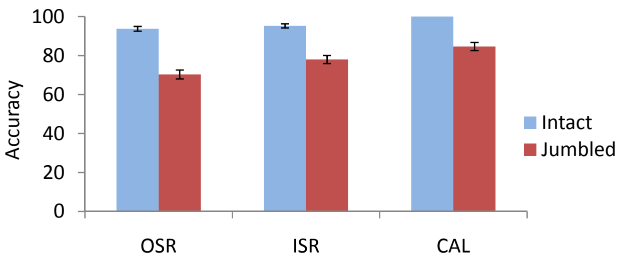


Figure 7: Human recognition accuracy on intact and jumbled images (with missing blocks when applicable).

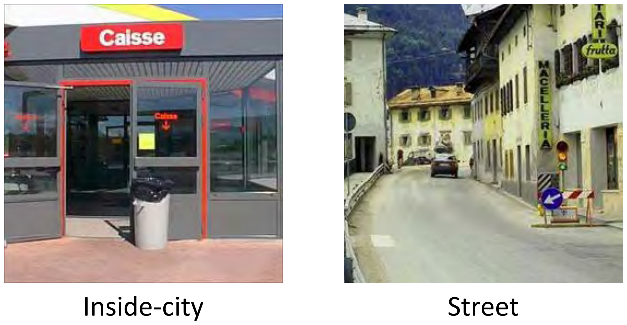
OSR in these different scenarios can be seen in Figure 6. First, we see that when we remove 75% of the blocks, the recognition accuracy is lower, but not with statistical significance. Secondly, we see that not displaying the holes in the image is more effective. Interestingly, the accuracies are higher when using only 25% of the blocks (without visible holes) than when using all the blocks. We suspect that this may be because even with fewer blocks, the images were displayed on the screen at the same size, effectively making each block larger. The robustness of the human recognition system to a large proportion of missing blocks in jumbled images indicates that the processing does not involve reassembling the image back together, or any such complex inference process that requires consistency in the appearance of the blocks.

To summarize, for further tests, for OSR and CAL we use only 36 of the 144 blocks in each image without displaying the holes corresponding to the missing blocks. We note that this is not the same as using $n = 6$ when jumbling the images. For ISR (where $n = 6$) all 36 blocks are retained. Subjects' recognition accuracies on these jumbled images with missing blocks (where applicable) on the three datasets are shown in Figure 7. For comparison, we also show the accuracy of human subjects at recognizing intact (unjumbled) images (where of course, no blocks are missing). We can see that although the accuracies on the jumbled images are significantly higher than chance, they are lower than the accuracies on intact images with statistical significance. This difference in accuracies can be attributed to the necessity of global information for image classification.

3.1.2 Filtering the Datasets

In order to work with only those images that allow for meaningful conclusions to be drawn, we filter our dataset and conduct further experiments only with a subset of the images. In a later experiment, we also show results using the entire unfiltered dataset.

Removing semantic inconsistencies: To remove the effects of unrelated factors that influence the underlying classification task, we further experiment with only those images that are classified correctly and consistently across subjects when they are intact (not jumbled). This is to ensure that none of the images included in the remaining studies are confusing because of semantic or linguistic properties of the labels. For example, "inside city" and "street" scenes can be confusing for certain images in the OSR dataset, as seen in Figure 8. As our first filtering step, all intact images that were incorrectly classified by even one out of the 10 subjects were discarded.



Inside-city

Street

Figure 8: Semantic ambiguities in the OSR dataset.

Removing human inconsistencies: In order to understand which functional model predicts human categorization of jumbled images well, it is important to work only with those images that are consistently classified across subjects, so that "the human response" is well defined. As the second filtering step, only those jumbled images that are consistently classified across subjects are considered for further tests. A jumbled image is assigned to the class picked by a majority of the subjects. It is considered to be classified consistently if twice the number of subjects picked the most frequent class, as compared to the number of subjects that pick the second most frequent class. This ensures that the image is classified consistently, but not necessarily correctly (since the class with the most votes is not guaranteed to be the correct class). For each dataset, we retained at most 7 consistently and correctly classified and at most 5 consistently but incorrectly classified images from each category. Using this procedure, we retained 43 images from OSR (35 classified correctly, and 8 incorrectly), 70 images from ISR (55 correctly and 15 incorrectly), and 47 images from CAL (42 correctly and 5 incorrectly) for further investigation. Overall, of the retained jumbled images, 18% were incorrectly classified.

3.1.3 Individual Blocks

The final test is to have subjects classify each block from the above selected images in isolation. Each block is displayed individually, and subjects are asked to assign it to one of the categories. OSR contains 43 images with 36 blocks per image, which were classified by 10 subjects each, resulting in a total of ∼16k responses. Similarly, we obtained ∼25k re-

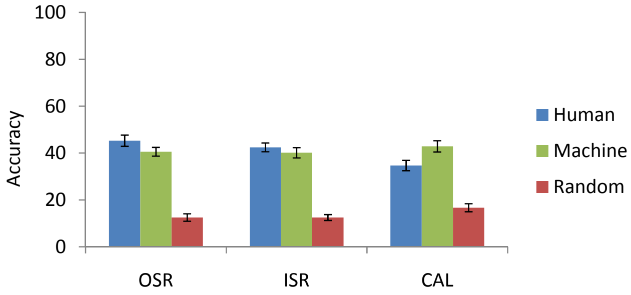


Figure 9: Human and machine accuracy at assigning each individual block in isolation to a scene/object category.

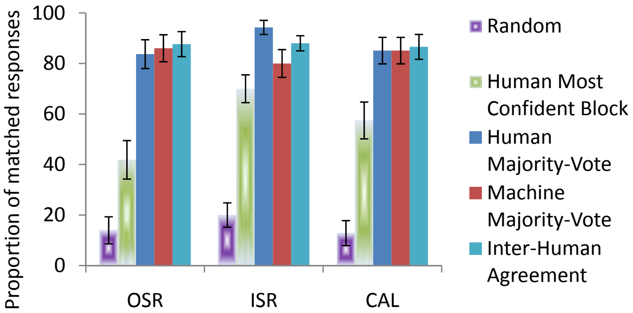


Figure 10: Percentage of images on which the majority-vote models' predictions match subjects' classification of entire jumbled images.

sponses for ISR and ∼17k responses for CAL. Subjects' accuracy at assigning the individual blocks to the corresponding classes is shown in Figure 9. It can be seen that while the classification of the individual blocks is higher than random indicating that the responses have some signal and not just noise, it is significantly lower than the classification of intact or jumbled images. We note that the accuracy for CAL on the individual blocks is lower than OSR and ISR (even though accuracies on intact and jumbled images for CAL are higher in Figure 7). This is because the clear background in many CAL images leads to empty blocks which subjects can not classify reliably. This test also indicates that the number of blocks n that we choose to divide the images into (as described in Section 3.1.1) does not result in large enough blocks that would allow for reliable scene recognition using the individual blocks alone. This demonstrates that the information available is truly local in nature.

We accumulate these noisy local classifications of the individual blocks across subjects. Each of the 36 blocks in an image is associated with a weighted vote for each class, where the weight corresponds to the proportion of the 10 subjects that selected that class. We determine the class that receives the largest weight of votes across the 36 blocks in the image. This class label can be compared to the label assigned by the subjects when viewing the entire jumbled images. If a large percentage of the majority-vote predictions match the global responses, we conclude that the majority-vote is a good functional model to describe how humans leverage local information to classify jumbled images. Before we present results in Section 4, we describe our machine experiments.

3.2. Machine Implementation

In addition to the human studies, we perform machine experiments implementing a bag-of-words-based majority-vote strategy to further test the model's ability to predict human classification of jumbled images. Computational models that aggregate local block-based decisions has been explored by Szummer *et al.* [27] for indoor-outdoor image classification. Our implementation is simpler than most of their models.

We describe each block in a jumbled image with a spatial distribution of color and texture features. For color, we use the RGB (3 features) and HSV (3 features) channels. For each of these channels, the average value within each cell of a 2 × 2 grid on the block is recorded. For texture, the block is filtered with a bank of 12 filters at 6 orientations and 2 scales [28]. For each filter channel, the average and maximum response within the cells of the 2 × 2 grid is recorded. Each block is thus represented with a 120 dimensional descriptor.

These descriptors are collected from blocks from a separate set of training images (55k blocks for OSR, 14k for ISR and 43k for CAL), and clustered into a dictionary of 500 codewords for each dataset using k-means clustering. During training, for each codeword, we estimate the posterior distribution of a scene/object category given the codeword. During testing, each block in the (jumbled) test image is assigned to a codeword, and the posterior distribution corresponding to the codeword is used as votes for the different class categories. These weighted votes are accumulated across all the blocks (36) in the image (similar to the weighted-majority-vote accumulation on human responses to individual blocks), and the class with the largest weight of votes is used as the predicted class for the image. Again, these predictions can be compared to the labels assigned by human subjects to the entire jumbled images, to assess how closely a machine implementation of a majority-vote model describes how humans classify jumbled images. For reference, the accuracy of classifying each individual block in a jumbled image by assigning it to the category with the highest posterior distribution given the codeword the block is closest to is shown in Figure 9. We see that the machine accuracy is very similar to the human accuracy, providing a strong indication that existing machine approaches capture local information as effectively as humans.

4. Results

We compare how closely the majority-vote model mimics human classification of jumbled images. Figure 10 shows the percentage of human responses to entire jum-

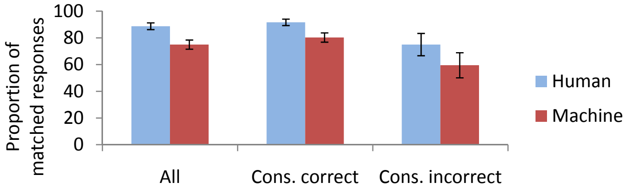


Figure 11: Percentage of matches between the majority-vote model applied to the human/machine classification of individual blocks, and the human responses to the consistently correctly and consistently incorrectly classified entire jumbled images (across all three datasets).

bled images that match the predictions of the majority-vote model, applied to the human responses to local individual blocks, as well as our bag-of-words based machine implementation. We can see that in both cases, the majority-vote model correctly predicts a large percentage of the human responses to the entire jumbled images. To place this matching rate in perspective, we also report the inter-human matching rate in Figure 10. As described in Section 3.1, each jumbled image was classified by 10 subjects. Although we selected those images that had high consistency in the responses obtained (as described in Section 3.1.2), the responses were not 100% consistent. We compute the inter-human matching rate by averaging across images the percentage of responses that matched the response most subjects gave the image. As seen in Figure 10, the matching rate of the majority-vote model on the human responses to individual blocks, as well as the machine implementation, are statistically the same as the inter-human agreement. This indicates that majority-vote is indeed an accurate functional model that mimics how humans leverage local information alone to classify jumbled images. For sake of comparison, in Figure 10 we also show the matching rate of a baseline model where an image is assigned to the class picked by the single block most consistently classified by humans (i.e. most confident block) in the image. We see that this model, although significantly better than random, is too simplistic to predict human responses to jumbled images reliably.

As described in Section 3.1.2, we are only working with jumbled images that were consistently (correctly or incorrectly) classified by the subject. We now look at the images that were correctly classified (132 images across datasets) and incorrectly classified (28 images across datasets) separately. The results are shown in Figure 11. We see that the majority-vote model, on both humans and machines responses to individual blocks, can accurately predict the correct human responses to jumbled images, *as well as their mistakes*, further stressing the aptness of the majority vote model as a functional model for how humans utilize local information to classify jumbled images.

Lastly, to ensure that our findings are not biased by only considering filtered datasets (Section 3.1.2), we apply our naive machine implementation of the bag-of-words based majority-vote scheme to all 400 (300 for ISR) jumbled images in our pre-filtered datasets. The accuracies obtained are shown in Figure 12. For sake of reference, we also show the corresponding human accuracies on the jumbled images. We see that the accuracies are very comparable, with no statistical difference for OSR and CAL. This demonstrates that existing implementations of a simple computational model suffice to match human accuracies at classifying images using local information alone, and future research endeavors should focus on modeling global information in images. The ISR dataset has a wider variety in local visual appearances, and perhaps a larger dictionary size in the machine implementation could further improve the performance.

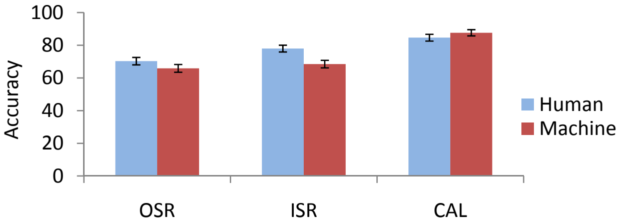


Figure 12: Accuracy of machine implementation of a majority-vote strategy for recognizing jumbled images (50 images per category in each dataset) as compared to accuracy of human subjects.

5. Discussion

Our analysis so far was focused on a particular choice of block sizes. A study for the future would be to understand if as the block size changes (particularly, decreases), do subjects employ more complex models than majority-vote to make up for the degrading local information, or is the drastic drop in human accuracy of recognizing jumbled images with smaller blocks (Figure 4) precisely due to the majority-vote model's inability to deduce the correct category from local responses based on these small blocks. What about increasing number of scene and object categories? This paper focuses only on upright color images, and whether the findings carry over to gray scale images or randomly oriented images/blocks is a subject of future research.

We note that to keep the experimental set up consistent between the machine experiments and human studies, we did not optimize the block size in our machine implementation. The choice of features or codeword dictionary size were not optimized either. Moreover, our implementation is even simpler than the typical bag-of-words implementation, where a classifier is often trained on the codeword-distribution in the image (in our case, ≈500D). In our implementation, each codeword (1D) votes for the scene category, and the votes are accumulated via a straightforward majority-vote instead of a discriminatively trained measure.

The simplicity of the functional model found to accurately mimic how humans leverage local information in absence of global information, and the ability of a straightforward machine implementation to match human accuracies at recognizing jumbled images is a strong indication that the research community should focus on modeling global information in images. This global information could be the spatial layout of scenes (especially for indoor and other complex scenes where gist is found to be less effective), or spatial interactions among the local entities (be it patches, or segments, or parts, or objects) or contextual reasoning. To focus on modeling global information, in similar spirits to this paper, perhaps it would be useful to study scenarios where the local information is impoverished, but global information is intact, such as in low-resolution images, where existing computational models are known to fail, while human abilities are surprisingly effective [12, 15]. Attempting to solve machine vision in such low-resolution images would truly require us to explore global information in ways not explored thus far, potentially leading to unprecedented advancements.

Finally, we comment on the recent findings of Parikh *et al*. [6, 15] in light of this work. Consistent with this paper, Parikh *et al*. show that human contextual reasoning (global information) for semantic image segmentation [15] and object detection [6] is far superior to machines. However, seemingly contrary to the conclusions in this paper, Parikh *et al*. [6] show that humans are significantly superior to machines at recognizing parts from image patches (local information). Indeed, it is conceivable that the information in a local patch may be sufficient for humans to recognize parts (local concepts) but not scenes (global concepts). Systematically evaluating the role of local and global information for a spectrum of tasks to unify these findings is part of future work.

6. Conclusion

Machine performance at the task of image classification falls significantly short when compared to the corresponding human abilities. Understanding which aspects contribute to this gap the most could help the community to focus our research efforts better. In this paper, we isolate the local information in an image from it's global information via jumbled images. We study how humans leverage local information and test if existing computational tools are effective enough to match human recognition abilities in the presence of local information alone. To our surprise, through a series of human studies and machine experiments, we find that a simple bag-of-words based majority-vote model is an accurate functional model to mimic how humans recognize jumbled images. Moreover, a simple unoptimized machine implementation of this majority-vote strategy achieves comparable recognition accuracy to humans at classifying jumbled images. We believe this to be a strong indication that the community should focus future research endeavors towards modeling global information in images, perhaps by attempting machine visual recognition in low-resolution images.

Acknowledgements: This research was supported in part by the National Science Foundation under IIS-1115719.

References

[1] J. Vogel, A. Schwaninger, C. Wallraven and H. H. Blthoff. Categorization of Natural Scenes: Local vs. Global Information. *Symposium on Applied Perception in Graphics and Visualization (APGV)*, 2006.

[2] D. Parikh and C. L. Zitnick. The Role of Features, Algorithms and Data in Visual Recognition. *CVPR*, 2010.

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

[4] T. S. Cho, S. Avidan and W. T. Freeman. A Probabilistic Image Jigsaw Puzzle Solver. *CVPR*, 2010.

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

[6] D. Parikh and C. L. Zitnick. Finding the Weakest Link in Person Detectors. *CVPR*, 2011.

[7] A. Oliva and A. Torralba. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in Brain Research*, 2006.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] A. Schwaninger, C. Carbon and H. Leder. Expert Face Processing: Specialization and Constraints. *In Development of Face Processing*, 2003.

[14] W. Hayward. After the Viewpoint Debate: Where Next in Object Recognition? *Trends in Cognitive Sciences*, 2003.

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

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

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

[18] I. Biederman. Perceiving Real-World Scenes. *Science*, 1972.

[19] I. Biederman, A. Glass and E. W. Stacy. Searching for Objects in Real-World Scenes. *Journal of Experimental Psychology*, 1973.

[20] I. Biederman, J. Rabinowitz, A. Glass and E. W. Stacy. On the Information Extracted From a Glance at a Scene. *Journal of Experimental Psychology*, 1974.

[21] G. R. Loftus, W. W. Nelson and H. J. Kallman. Differential Acquisition Rates for Different Types of Information from Pictures. *Quarterly Journal of Experimental Psychology*, 1983.

[22] G. Giraudet and C. Roumes. Independence from Context Information Provided by Spatial Signature Learning in a Natural Object Localization Task. *Notes in Artificial Intelligence: Modelling and Using Context*, 1999.

[23] B. S. Tjan, A. I. Ruppertsberg and H. H. Blthoff. Early use of color but not Local Structure in Rapid Scene Perception. *Investigative Ophthalmology and Visual Science*, 1999.

[24] K. Yokosawa and H. Mitsumatsu. Does Disruption of a Scene Impair Change Detection? *Journal of Vision*, 2003.

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

[26] 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.

[27] M. Szummer and R. W. Picard. Indoor-Outdoor Image Classification. *IEEE International Workshop on Content-based Access of Image and Video Databases*, 1998

[28] T. Leung and J. Malik. Representing and Recognizing the Visual Appearance of Materials Using Three-dimensional Textons. *IJCV*, 2001.