Comparison of dataset bias in object recognition benchmarks

Ian Model[1], Lior Shamir[1]
[1]Department of Computer Science, Lawrence Technological University, 21000 W Ten Mile Rd., Southfield, MI 48075 USA

Current research in the area of automatic visual object recognition heavily relies on testing the performance of new algorithms by using benchmark datasets. Such datasets can be based on standardized datasets collected systematically in a controlled environment (e.g., COIL-20), as well benchmarks compiled by collecting images from various sources, normally via the World Wide Web (e.g., Caltech 101). Here we test bias in benchmark datasets by separating a small area from each image such that the area is seemingly blank, and too small to allow manual recognition of the object. The method can be used to detect the existence of dataset bias in a single object recognition dataset, and compare the bias to other datasets. The results show that all tested datasets allowed classification accuracy higher than mere chance by using the small images, although the sub-images did not contain any visually interpretable information. That shows that the consistency of the images within the different classes of object recognition datasets can allow classifying the images even by algorithms that do not recognize objects. Among the tested datasets PASCAL is the dataset with the lowest observed bias, while datasets acquired in a controlled environment such as COIL-20, COIL-100, and NEC Animals are more vulnerable to bias, and can be classified by the sub-images with accuracy far higher than mere chance.

Index Terms—**Object recognition, performance evaluation, benchmarks, validation, computer vision, pattern recognition.**

I. Introduction

As the field of computer vision has been growing consistently, automatic visual object recognition has been becoming an increasingly important topic of study. Numerous methods for automatic object recognition have been proposed and implemented in the past few decades, and efforts are still being continued. To assess new object recognition algorithms and compare their performance to previously proposed methods, it is common to use pre-defined publicly available object recognition benchmarks, allowing a standard comparison of the efficacy of different algorithms [1]. Examples of some commonly used visual object recognition benchmarks include Caltech 101 [2], Caltech 256 [3], COIL-20 [4], COIL-100 [5], Tiny [6], LabelMe [7], Fifteen Scenes [8], SUN2012 [9], ImageNet [10], PASCAL [11], and more. Since benchmarks have become critical in the development of object recognition methods, it is important to assess their ability to reflect real-world object recognition.

Ideally, an algorithm that demonstrates accuracy in classifying images contained in object recognition benchmark datasets should also be able to be equally efficient in recognizing real-world objects. However, recent studies have shown that some of the commonly used datasets are biased [12], [13], [14], and do not necessarily provide a reliable tool to measure the ability of algorithms to automatically recognize objects in a real-world environment. That can be evident from the observation that when using training and test data from two different datasets the performance is significantly lower than the performance achieved by separating the same dataset into training and test sets [12]. That observation is counterintuitive to most machine learning systems, in which the classification accuracy is expected to increase as the number of training samples gets higher [15].

Another set of experiments showed dataset bias when the training and test samples are taken from the same dataset, compared to the recognition accuracy when the training images are taken from one dataset and the test images are taken from another dataset [13]. Such experiments are limited to image classes that are common across several datasets such as *car* and *cow*. Combining several datasets in a single experiment can be done in a leave-one-dataset-out fashion, leaving one dataset for testing, to avoid dataset bias when using heterogeneous datasets [14].

The use of natural images for the purpose of object recognition has been challenged by evidence showing that these datasets are biased, and might not be fully reliable as tools for testing the real-world accuracy of object recognition methods [16]. For instance, high-level features such as the size of the object can lead to false detections, and the false detection rate of objects that are too small or too large varies between different object classes [16]. On the other hand, it has been shown that state-of-the-art object recognition methods applied to benchmarks such as Caltech 101 [2] can be outperformed by using merely global low-level features that do not aim at identifying objects [17]. These results demonstrate the existence of dominant low-level features leading to classification accuracy higher than the accuracy achieved in a real-world setting [17].

Low-level features have been identified as a source of bias that can have substantial impact on the validity of quantitative results produced using benchmark datasets. For instance, it has been shown that many face recognition benchmarks can be identified with accuracy far higher than random, sometimes even close to 100%, even if the part of the image being analyzed has no face or hair features in it [18]. That is, by analyzing a seemingly blank background area of the face image, the image can be identified and correctly matched with the gallery face images of the same person [18]. These experiments show that high classification accuracy of face recognition datasets can be achieved without recognizing faces [18].

Another example is the field of bioimage informatics, where

Corresponding author: L. Shamir (email: lshamir@mtu.edu).

removing cell areas from the microscopy images used in machine learning experiments provided results very similar to experiments done with the original images [19], demonstrating that the performance of computer vision methods reported in the literature could be biased, and driven by artifacts of the image acquisition process rather than identifiable biological features of the cells [19]. In the field of automatic speech recognition, separating one second of silence from the beginning of each speech sample provided very high accuracy of accent classification, although no accent information was contained in the samples [20].

In this paper we examine some common object recognition datasets and test their ability to provide a reliable assessment of the performance of visual object recognition algorithms. The experiments are based on separating a small seemingly uninformative area from each image such that the image content in that area is insufficient for human recognition of the object. Ability of the algorithm to recognize the objects using the small blank sub-images in accuracy higher than mere chance is an indication that the computer analysis outperforms human identification, which is not expected in automatic object recognition and indicates that the computer uses other visual aspects of the image for the classification rather than the visual information of the objects. These low level features can be difficult to sense by the unaided human eye, but their presence can be used by machine vision algorithms to distinguish between the object classes, and therefore the object recognition accuracy of these datasets might not reflect the actual ability of the object recognition algorithm to detect objects in a real-world environment.

II. Data

We used image data taken from nine commonly used visual object recognition benchmarks: COIL-20 [4], COIL-100 [5], NEC Animals [21], Caltech 101 [2], Microsoft Research Cambridge Object Recognition Image Database (MSRCORID) [22], SUN2012 [9], PASCAL [11], [23], ImageNet [10], [24] Large Scale Visual Recognition Challenge (ILSVRC2011), and Fifteen Scenes [8]. From each image we separated a small 20×20 pixel area from the center of the image.

The reason for using 20×20 pixel areas is that these small sub-images do not contain information that is sufficient to identify the object visually. Effective machine vision for object recognition is expected to provide the same output as the human vision given the same input data. Using the 20×20 images, the human vision is not able to recognize the imaged objects, and therefore differences between the performance of machine and human vision can indicate that the performance of the machine vision is not based on the recognition of the object, but instead uses other information contained in the image. If the human vision outperforms machine vision in object recognition we can reasonably conclude that the machine vision algorithm does not match the performance of the brain in that highly complex cognitive task. However, if computer vision can recognize objects that the human vision cannot, it is required to investigate whether the computer uses some patterns in the data that assist it to classify the images, rather than actually identifying the objects in them.

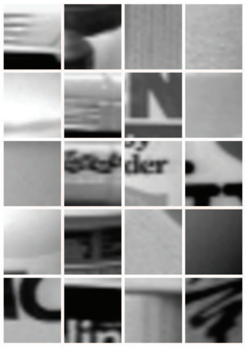


Fig. 1. Example 20×20 images separated from COIL-20 images.

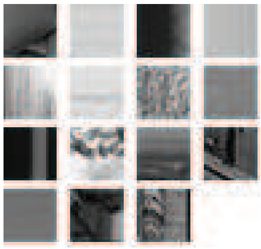


Fig. 3. Example 20×20 images separated from images of the Fifteen Scenes dataset.

Figure 1 shows examples of small 20×20 areas separated from images of each of the 20 classes of COIL-20 displayed in Figure 2. The position of each sub-image displayed in Figure 1 corresponds to the position of the full image in Figure 2. As can be seen in the figure, it is not possible to identify the object from these small sub-images, and in some cases they seem blank and identical to the small areas separated from images of other object classes. For instance, the top right image is an image of a small bowl, but in the 20×20 sub-image no information that implies that the imaged object is a bowl is present, and visual inspection of the image does not provide information related to a bowl. The bottom right image is taken from an image of a cup, but the sub-image does not contain visual information sufficient to identify a cup.

Figure 3 shows 20×20 windows separated from one image in each class of the Fifteen Scene dataset. Like with COIL-20, the small images do not contain visual information that can identify the scene. For instance, the top left image is an image of a bedroom, the image next to it is an image of a suburb, and the third image in that row is an image of industry, separated from the original images displayed in Figure 4. Some of the sub images are seemingly blank, and contain no interpretable information.

The samples of each dataset were separated to training and test sets as specified in Table I, and each classification experiment was repeated 20 times such that in each run different images were randomly allocated to training and test sets. For the SUN2012 dataset only classes with 64 or more samples were used, which reduced the number of categories to 39 classes that had 64 or more images.

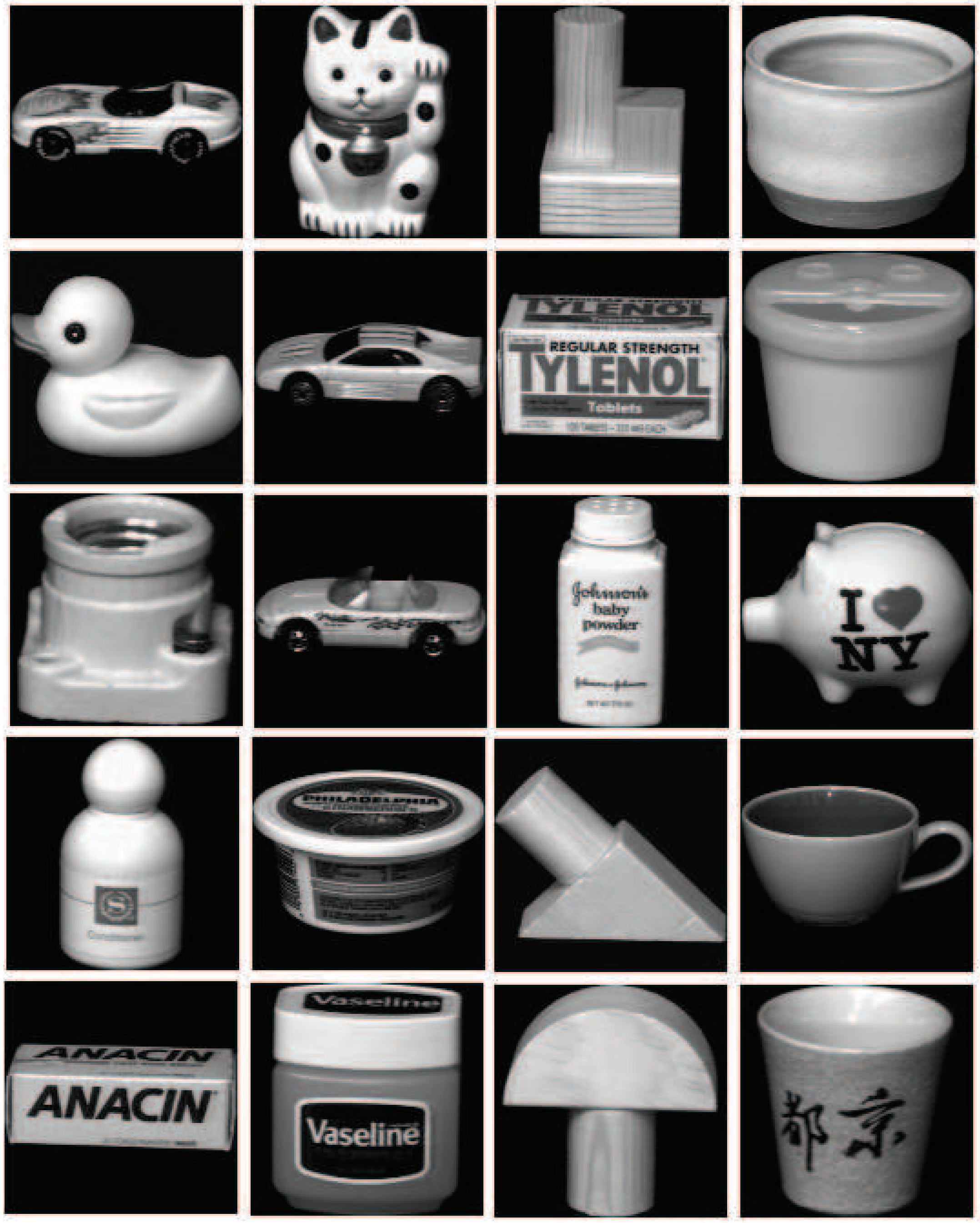


Fig. 2. Example images taken from COIL-20, from which the example sub images of Figure 1 were separated.

III. Image analysis method

The image analysis method used in the experiments was *Wndchrm* [25], [18], [26], [27], [28]. *Wndchrm* utilizes several different image analysis algorithms to extract from each image a comprehensive set of numerical image content descriptors.

In summary, *Wndchrm* computes a total of 2885 numerical image content descriptors from each image which include textures [29], [30], edges [31], shapes [26], statistical distribution of the pixel intensities [32], polynomial decomposition of the image [33], and fractal features [34].

These features are computed from the raw pixels, but also from the image transforms and multi-order transforms. The image transforms include the Fourier, Chebyshev, Wavelet, and the edge-magnitude transform, as well as combinations of multi-order transforms [25], [28], [18], [26]. The combinations of transforms include the Fourier transform of the Chebyshev transform, the wavelet transform of the Chebyshev transform, the Fourier transform of the wavelet transform, the wavelet transform of the Fourier transform, the Chebyshev transform of the Fourier transform, and the Fourier and Chebyshev transforms of the edge magnitude transform [25], [28], [18], [26].

After the numerical image content descriptors are computed,

Fig. 4. Examples of full images taken from the the Fifteen Scenes dataset from which the 20×20 example sub images displayed in Figure 3 were separated.

TABLE I
THE NUMBER OF TRAINING AND TEST SAMPLES PER CLASS USED IN THE DIFFERENT OBJECT RECOGNITION DATASETS. EACH EXPERIMENT WAS REPEATED 20 TIMES SUCH THAT IN EACH RUN DIFFERENT IMAGES WERE RANDOMLY ALLOCATED TO TRAINING AND TEST SETS

Dataset	# classes	# Training samples per class	# Test samples per class
Fifteen Scenes	15	60	12
Caltech 101	101	26	4
COIL-20	20	60	12
COIL-100	100	60	12
MSRCORID	20	60	8
ImageNet	689	60	12
NEC Animals	60	60	5
SUN2012	42	60	4
PASCAL	19	60	12

each numerical image content descriptor is assigned with its Fisher discriminant score [35], computed using the training samples. The numerical image content descriptors are then ranked based on their Fisher discriminant scores, and the 85% of the least informative features with the lowest Fisher scores are rejected from the analysis [25], [27], [18], [28]. The images are then classified using the Weighted Nearest Neighbor scheme such that the Fisher discriminant scores are used as weights.

The comprehensive set of numerical image content descriptors makes the algorithm informative for different types of visual content, and it has been tested and demonstrated efficacy in several different domains [36], [37], [38], [39], [40], [41]. A detailed description of the method and comprehensive performance analysis is available in [25], [18], [26], [42]. The source code of the method is publicly available, as well as binaries for MS-Windows [43].

TABLE II
CLASSIFICATION ACCURACY ACHIEVED WHEN USING 20×20 SUB-IMAGES SEPARATED FROM THE ORIGINAL IMAGES

Dataset	Mere chance accuracy (%)	Sub-image accuracy (%)	Improvement over mere chance (%)
Fifteen Scenes	6.7	27.5	310
Caltech 101	1	4.5	350
MSRCORID	5	21.6	332
ImageNet	0.15	0.69	360
SUN2012	2.6	7.2	177
PASCAL	5.3	7.6	43
COIL-20	5	81.1	1522
COIL-100	1	70.6	6960
NEC Animals	1.7	33.7	1882

IV. RESULTS

The method described in Section III was applied to the modified object recognition benchmark datasets described in Section II. Each experiment was run 20 times such that in each run the images were randomly allocated to training and test sets based on the numbers specified in Table I. The classification accuracy achieved with the modified object recognition datasets are specified in Table II.

As the table shows, in all cases the classification accuracy was higher than the accuracy achieved by mere chance, showing that the objects could be identified by images in which no information that can visually identify the object is present. The dataset that has the lowest classification accuracy of the sub-images compared to random accuracy is PASCAL, with classification accuracy of ∼7.6% compared to 5.3% of random classification accuracy.

The experimental results also show that the standardized datasets COIL-20, COIL-100, and NEC Animals show much higher classification accuracy compared to the datasets compiled from many different sources, and contain images that were not necessarily taken for the purpose of developing object recognition systems. The most extreme example is COIL-100, where random guessing accuracy of 1% was increased by almost 7000% to ∼70.6% when using just a 20×20 pixel area of each original image. These results show that image features that are irrelevant for object recognition but can differentiate between the image classes are more dominant in object recognition benchmarks acquired in a controlled fashion, and are less dominant when the images are collected from various sources.

The results also show that automatic classification using the sub-images of Fifteen Scenes, Caltech 101, ImageNet and MSRCORID has similar improvement over mere chance. COIL-20 and NEC Animals also have a similar improvement of the classification accuracy compared to mere chance accuracy, and that improvement is much higher compared to the datasets compiled from the internet. PASCAL and SUN2012 have the mildest improvement when classified with the sub-images, showing that the information in the sub-images is less consistent within the classes of these datasets, and therefore does not allow achieving high classification accuracy by using merely small parts of the original images.

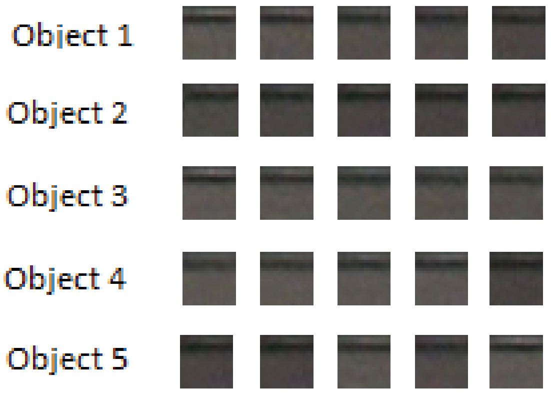


Fig. 5. Example 20×20 images separated from the bottom right corner of the first five objects of NEC Animals dataset

In COIL-20 and COIL-100 the background has been removed, but in NEC Animals the standardized background is present in the images. That allows repeating the experiment such that the small 20×20 sub-image is taken from the background, and does not contain any foreground pixels. Figure 5 shows the background images of the first five images of the first five objects, such that the 20×20 sub images are taken from the bottom right corner of each original image.

As the figure shows, each object is represented in the dataset by a sub-image that does not contain any information about the object, and since the background has been standardized all images look similar to the unaided human eye. However, applying the image classifier described in Section III to that dataset provided classification accuracy of ∼33.7%, clearly higher than the ∼1.7% of mere chance accuracy. When using 200×200 areas taken from the bottom right corner of each image the images in the new dataset were also similar to each other and did not contain foreground pixels, but the classification accuracy was elevated to 45.5%. These results show that the images in the dataset contain artifacts that allow the recognition of the object regardless of the apparent visual content of the images. While contextual information can be used to recognize objects [44], the information that allowed the recognition is clearly not contextual, as no contextual information was included in these sub images.

In another experiment we separated a 20×20 pixel area from each image such that the 20×20 window was selected randomly. In the case of COIL-20 and COIL-100, the area was selected such that it did not contain just black pixels with intensity value set to 0. The results of the experiment are displayed in Table III.

As the table shows, there was no substantial change when the 20×20 regions were selected randomly. Some higher difference is observed in the case of COIL-20 and COIL-100, although the classification accuracy is still far higher than mere chance.

When the size of the sub-image increases, it is reasonable to assume that the classification accuracy increases as more information is contained in the sub-images. Figure 6 shows the

TABLE III
CLASSIFICATION ACCURACY ACHIEVED WHEN USING 20×20 SUB-IMAGES SEPARATED FROM RANDOM LOCATIONS IN THE ORIGINAL IMAGES

Dataset	Accuracy (%)
Fifteen Scenes	27.7
Caltech 101	4.8
MSRCORID	21.3
ImageNet	0.69
SUN2012	7.1
PASCAL	7.9
COIL-20	75.6
COIL-100	63.5
NEC Animals	31.5

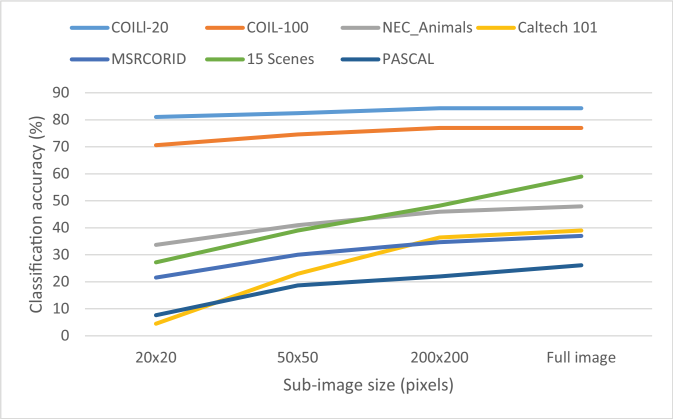


Fig. 6. The classification accuracy of the different datasets when using different sizes of sub-images separated from each image

changes in the classification accuracy of the different datasets when the sub-images are of different sizes. As the figure shows, the classification accuracy indeed increases when the sub-image gets larger, but the increase is weaker in controlled datasets such as COIL-20 and COIL-100, compared to datasets of images collected from the internet such as Caltech 101 and Fifteen Scenes. A possible explanation is that the controlled datasets can be classified by artifacts of the image acquisition as discusses in Section VI, and therefore increasing the size of the sub-image does not necessarily contributes substantial information that was not included in the smaller window. The classification of images collected from the internet, on the other hand, can be improved by adding more information that can assist in recognizing the object.

Computer vision does not necessarily follow the human vision, and the small areas selected from the images in some cases might not include intelligible information, but the computer can still use that information to identify the object. To test whether the sub images contain information that can be used to identify the object, we separated the training and the test datasets [12], such that the classifier was trained with 20×20 sub-images separated from one dataset and tested with the 20×20 sub-images separated from the other dataset. If the sub images contained information that can identify the object, we would expect some classification accuracy higher than mere chance.

TABLE IV
CLASSIFICATION ACCURACY WHEN USING 20×20 SUB-IMAGES OF CARS AND ANIMALS, AND TRAINING THE CLASSIFIER WITH ONE DATASET WHILE TESTING WITH ANOTHER DATASET

Dataset	Caltech 101	MSRCORID	SUN2012	PASCAL
Caltech 101	72.6%	51.5%	51.8%	50.4%
MSRCORID	51.4%	70.2%	52.1%	48.4%
SUN2012	50.2%	50.8%	68.5%	49.2%
PASCAL	48.9%	49.2%	50.5%	62.7%

For the experiment we used classes that are common in two or more datasets. Following [12] we used the classes *cars* and *person*, available in PASCAL and Caltech 101. The classification accuracy between these two classes of 20×20 sub-images is ∼64% and ∼71% for PASCAL and Caltech 101, respectively. However, when training with the sub-images of Caltech 101 and testing with the sub-images of PASCAL the classification accuracy was ∼52.0%, and was ∼49.3% when training with PASCAL and testing with the sub-images of Caltech 101.

To test with more datasets, we performed another experiment such that the two classes were *cars* and *animals*, two classes that are common to four of the tested benchmark datasets. Table IV shows the results of training the classifier with one dataset (the rows) and testing with another dataset (the columns). In the experiments the numbers of training and test sub-images per class were the numbers specified in Table I, and each experiment was repeated 20 times such that in each run the images were randomly selected to training and test sets.

The diagonal of the table, which shows the classification accuracy when the same dataset was used for both training and testing, shows that when separating the same dataset into training and test images the classification accuracy is higher than random when using the sub-images. However, when training with sub-images of one dataset and testing with sub-images of another dataset, non of the experiments provided classification accuracy beyond mere chance. These results show that the sub-images do not contain information that the classifier use to identify the object.

A. Computational complexity

Although Wndchrm is computationally intensive, its computational complexity is a function of the size of the image. An Intel Core-i7 processor can compute a single 20×20 image within less than eight seconds, so that a single core can compute the 1950 images of "Fifteen Scenes" in about four hours and 20 minutes, and the 7200 images of COIL-100 can be computed within ∼16 hours using a single core. The largest dataset used in this study, ImageNet, required computing 49,608 images, and can be completed by a single core in ∼4.6 days. However, to reduce the response time Wndchrm can be easily parallelized with negligible overhead [25]. In this study a Linux machine with 64 cores was used, so that COIL-100 was computed in about 15 minutes, and ImageNet in less than two hours.

V. Conclusion

While the field of computer vision has been expanding rapidly and new problems are introduced [45], [46], [47], [48], [49], automatic object recognition has been one of the primary and most studied machine vision tasks. Like some many other tasks in computer vision, it is modeled as a labeling problem [50], and studied using benchmark datasets. These object recognition benchmark datasets are highly useful tools for the development and evaluation of automatic object recognition methods. Here we tested possible bias in object recognition benchmark datasets by classifying small sub-images separated from the original images such that the sub images do not have sufficient information that allows the recognition of the object.

The difference between the accuracy achieved by the method and the expected mere chance classification accuracy provides an indication of the magnitude of the dataset bias. For instance, the experiments showed that PASCAL has the smallest difference between the accuracy achieved with the method and mere chance accuracy. On the other hand, in COIL-100 mere chance accuracy was 1%, while the accuracy of the method was over 70%. The observation that 20×20 sub-images that contain no interpretable visual information can be classified in accuracy higher than mere chance shows that the datasets can be classified by algorithms that do not necessarily identify objects, but can take advantage of the image selection and the consistency of the images within the classes to accurately classify the images.

Some object recognition methods apply a first step of object segmentation to separate the background from the object in the image. However, the fact that the seemingly blank background of the image contains information that can differentiate between the classes makes it reasonable to assume that information that can differentiate between the images also exists in the foreground, so separating the foreground by automatic segmentation does not guarantee unbiased classification. Because the foreground area of the objects also contains the differences between the objects, using just the foreground makes it difficult to know whether the classification accuracy can be attributed to dataset bias or to the actual differences between the objects. Because the foreground and background are part of the same image, the foreground areas can not be safely considered unbiased.

The proposed method can be used to identify the existence of bias in new benchmark datasets before they are released, or to measure the magnitude of such bias in comparison to other object recognition benchmark datasets. Source code of the image classification algorithm is publicly available [43], and can be used by researchers to detect possible bias of object classification algorithms.

As the results showed, some commonly used object recognition datasets can be classified with accuracy higher than random even when using small parts of the images that do not contain sufficient information to identify the object manually. However, the effect is much more dominant in the object datasets that are acquired in a controlled environment such as COIL-20, COIL-100, and Nec Animal, compared to datasets compiled from natural images downloaded from the internet such as Caltech 101. The fact that these datasets were classified with accuracy much higher than random by using just seemingly identical background areas shows that these artifacts exist in the images, and it is reasonable to assume that they can exist also in the foreground object areas, allowing machine vision algorithm classify the images even without identification of the imaged object.

Also, datasets compiled from natural images collected from the internet are also vulnerable to dataset bias, as the collection of the images is also dependent on the human perception of the person selecting the images and constructing the dataset. The results show that PASCAL and SUN2012 are less biased compared to Fifteen Scenes, Caltech 101, ImageNet, and MSRCORID. That means that information that is not related to the imaged objects in the classes of these datasets is less consistent within the classes, and therefore does not allow accurate recognition of the images.

VI. Discussion

A. Possible dataset bias factors

One of the possible reasons for the stronger bias identified in datasets collected in controlled environment is that all images are taken in the same session [18], [19]. Image datasets compiled such that images of each class are acquired in the same session can have common characteristics that are driven by the image acquisition session rather than by the object being imaged. That is, the images are classified by the session in which they were acquired rather than by the object in them. The same can happen when all images of a certain object are acquired in a single video, and then the frames are separated to create a dataset of single images.

While image acquisition is done in attempt to normalize the images, it is difficult to control all aspects to provide a fully standardized dataset. One cause for dataset bias can be the object being imaged itself, as evident by dataset bias in face recognition caused by different facial expressions, orientation of the face, clothes, background, etc [51], [52]. Image quality, resolution, and number of images can also lead to differences in the measured performance of face recognition algorithms [51].

However, bias can exist also within a single dataset even when all images in the dataset is acquired in the same fashion. For instance, acquiring all or some of the images of a certain object in the same session can lead to a dataset that supervised machine learning algorithms can classify by the image acquisition session rather than by the object. Effects such as subtle changes in the lighting conditions or the temperature of the camera CCD at the time of imaging can result in artifacts that are difficult to sense visually, but might lead to differences that can be detected by computer algorithms. For instance, if the environment is slightly darker when imaging a certain object, simple global features such as pixel intensity statistics can be used to identify that class, while the human eye is not sensitive enough to sense such subtle differences.

The effect of acquiring a class of images in one session has shown substantial bias in face recognition [18] and microscopy image classification [19]. For instance, if during the preparation of the dataset all images of object 1 are acquired, and then all images of object 2 are acquired, the two classes of images can be differentiated by the different conditions or status of the camera when each set of images was taken.

Another reason for such differences can be different photographers imaging different classes of objects. For instance, if photographer A acquires all images of object 1, and photographer B acquires all images of object 2, the two classes can be differentiated by the person taking the images rather than by the imaged objects. Different photographers can hold the camera differently or use the focus in a different manner, leading to systematic differences between the sets of images that allows computer classification between them regardless of the imaged objects.

B. Reducing dataset bias

The problem of differences between images that are not driven by the imaged objects can be addressed by acquiring the images such that the images within the different classes are acquired in random order rather than all images of a certain class at a time, so that each session has a random order of images of different classes [19]. Taking one image from each class at a time requires more efforts, but in that case the images cannot be distinguishable by the session in which they were acquired, as the images of each class are not taken consequently, and each session collects just one image.

The image acquisition process should also be normalized by the photographer such that all images are acquired by the same person, or images are imaged randomly by different photographers, but the imaging process should not by based on a strategy according which imaging duties are shared such that different classes of objects are imaged by different photographers. If the camera is placed on a tripod, the tripod or camera should not be touched until the image acquisition process ends.

A different solution can also be separating the acquisition of the training and the test sets. In automatic object recognition it is common to first acquire all images of all classes, and before the experiment separate each class into training and test images. If systematic artifacts in a certain class exist, the training stage can identify the patterns of these artifacts, and use them to classify the test samples. However, if the training and test sets are acquired in two completely separate sessions, these artifacts are not expected to exist in both the training and test sets, and the classification accuracy achieved should be unbiased [19].

Datasets compiled by collecting natural images from the Internet can also be biased by the preferences of the person collecting the images. Therefore, also when compiling benchmark datasets from the world wide web it can be safer to avoid sharing image collection duties such that each class is collected by a different person, and use as many data collectors as possible to increase the diversity in the image selection and avoid a bias driven by the perspective of the person who happen to collect the images, and might consciously or subconsciously prefer one type of images over another.

Computer vision does not necessarily aim at mimicking the way the human eye or brain work, but at providing the same output given the visual input. That allows machines to identify visual objects by analyzing visual elements of the image that are not necessarily the same elements used by the brain. Therefore, machine vision algorithms may validly use contextual information that can assist in recognizing objects. However, in this study we do not apply low-level features computed from the entire image [17], but instead we separate very small areas of the image that are either completely blank, or do not have information that allows visual identification of the object. Computers may operate differently from the human brain, but machine vision is measured by comparing the input and output of the computer to the input and output of human manually processing the visual data, even if the information processing itself is done in a completely different fashion. In the experiments above, the input does not allow human identification of the object by either the object itself or by contextual information, and therefore the fact that the images can still be classified by the computer shows that the performance achieved in classifying these benchmarks might be biased as a tool to evaluate real-world object recognition performance. Experiments using the sub-images of one dataset for training and another for testing did not result in classification accuracy higher than mere chance, also showing that the classifier did not identify the objects with the information in the sub-images.

Automatic object recognition is an increasingly important task in computer vision and image analysis, allowing better interaction of machines with the real world. To satisfy that important need it is important to review the existing assessment methods that are used to advance the field, and correct these methods to make them as reliable as possible for the purpose of assessing the performance of automatic object recognition in real-world settings.

Acknowledgment

The study was supported in part by NSF grant IIS-1546079. The computing facility used to analyze the data was funded by NSF grant CNS-1157162.

We would also like to thank the anonymous reviewers for the insightful comments that helped to improve the manuscript.

References

[1] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba *et al.*, "Dataset issues in object recognition," in *Toward category-level object recognition*. Springer, 2006, pp. 29–48.

[2] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[3] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Technical Report 7694, California Institute of Technology, Tech. Rep., 2007.

[4] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," Technical Report CUCS-005-96, Tech. Rep., 1996.

[5] S. K. Nayar, S. A. Nene, and H. Murase, "Columbia object image library (coil 100)," *Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96*, 1996.

[6] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[7] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “Labelme: a database and web-based tool for image annotation,” *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[8] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 2169–2178.

[9] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[11] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2014.

[12] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1521–1528.

[13] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A deeper look at dataset bias,” *arXiv preprint arXiv:1505.01257*, 2015.

[14] D. Stamos, S. Martelli, M. Nabi, A. McDonald, V. Murino, and M. Pontil, “Learning with dataset bias in latent subcategory models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3650–3658.

[15] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, “Undoing the damage of dataset bias,” in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 158–171.

[16] D. Hoiem, Y. Chodpathumwan, and Q. Dai, “Diagnosing error in object detectors,” in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 340–353.

[17] N. Pinto, D. D. Cox, and J. J. DiCarlo, “Why is real-world visual object recognition hard?” *PLoS Computational Biology*, vol. 4, no. 1, p. e27, 2008.

[18] L. Shamir, “Evaluation of face datasets as tools for assessing the performance of face recognition methods,” *International Journal of Computer Vision*, vol. 79, no. 3, pp. 225–230, 2008.

[19] ——, “Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis,” *Journal of Microscopy*, vol. 243, no. 3, pp. 284–292, 2011.

[20] L. Bock, Benjamin andShamir, “Assessing the efficacy of benchmarks for automatic speech accent recognition,” in *8th International Conference on Mobile Multimedia Communications*.

[21] H. Mobahi, R. Collobert, and J. Weston, “Deep learning from temporal coherence in video,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 737–744.

[22] A. Criminisi, “Microsoft research cambridge object recognition image database,” 2004.

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, pp. 1–42, 2014.

[25] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. G. Goldberg, “Wndchrm – an open source utility for biological image analysis,” *Source Code for Biology and Medicine*, vol. 3, p. 13, 2008.

[26] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, “Wnd-charm: Multi-purpose image classification using compound image transforms,” *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[27] L. Shamir, S. M. Ling, W. W. Scott, A. Bos, N. Orlov, T. J. Macura, D. M. Eckley, L. Ferrucci, and I. G. Goldberg, “Knee x-ray image analysis method for automated detection of osteoarthritis,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 407–415, 2009.

[28] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, “Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art,” *ACM Transactions on Applied Perception*, vol. 7, no. 2, p. 8, 2010.

[29] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man and Cybernetics*, no. 6, pp. 610–621, 1973.

[30] H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[31] J. M. Prewitt, “Object enhancement and extraction,” *Picture Processing and Psychopictorics*, vol. 10, no. 1, pp. 15–19, 1970.

[32] E. Hadjidemetriou, M. D. Grossberg, and S. K. Nayar, “Spatial information in multiresolution histograms,” in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2001, pp. I–702.

[33] M. R. Teague, “Image analysis via the general theory of moments,” *Journal of the Optical Society of America*, vol. 70, no. 8, pp. 920–930, 1980.

[34] C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, “Texture features for classification of ultrasonic liver images,” *IEEE Transactions on Medical Imaging*, vol. 11, no. 2, pp. 141–152, 1992.

[35] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 1.

[36] L. Shamir, S. M. Ling, W. Scott, M. Hochberg, L. Ferrucci, and I. G. Goldberg, “Early detection of radiographic knee osteoarthritis using computer-aided analysis,” *Osteoarthritis and Cartilage*, vol. 17, no. 10, pp. 1307–1312, 2009.

[37] E. Schwartz and L. Shamir, “Correlation between brain mri and continuous physiological and environmental traits using 2d global descriptors and multi-order image transforms,” *Journal of Medical Imaging and Health Informatics*, vol. 3, no. 1, pp. 12–16, 2013.

[38] L. Shamir and J. A. Tarakhovsky, “Computer analysis of art,” *ACM Journal on Computing and Cultural Heritage*, vol. 5, no. 2, p. 7, 2012.

[39] L. Shamir, “Computer analysis reveals similarities between the artistic styles of van gogh and pollock,” *Leonardo*, vol. 45, no. 2, pp. 149–154, 2012.

[40] L. Shamir, A. Holincheck, and J. Wallin, “Automatic quantitative morphological analysis of interacting galaxies,” *Astronomy and Computing*, vol. 2, pp. 67–73, 2013.

[41] E. Kuminski, J. George, J. Wallin, and L. Shamir, “Combining human and machine learning for morphological analysis of galaxy images,” *Publications of the Astronomical Society of the Pacific*, vol. 126, no. 944, pp. 959–967, 2014.

[42] L. Shamir, N. Orlov, D. M. Eckley, T. J. Macura, and I. G. Goldberg, “Iicbu 2008: a proposed benchmark suite for biological image analysis,” *Medical & Biological Engineering & Computing*, vol. 46, no. 9, pp. 943–947, 2008.

[43] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. Goldberg, “Wnd-charm: Multi-purpose image classifier,” *Astrophysics Source Code Library*, vol. 12, p. 002, 2013.

[44] J. Zhu, J. Yu, C. Wang, and F.-Z. Li, “Object recognition via contextual color attention,” *Journal of Visual Communication and Image Representation*, vol. 27, pp. 44–56, 2015.

[45] J. L. Sanz, *Advances in machine vision*. Springer Science & Business Media, 2012.

[46] S. Cubero, N. Aleixos, E. Moltó, J. Gómez-Sanchis, and J. Blasco, “Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables,” *Food and Bioprocess Technology*, vol. 4, no. 4, pp. 487–504, 2011.

[47] W. Ji, D. Zhao, F. Cheng, B. Xu, Y. Zhang, and J. Wang, “Automatic recognition vision system guided for apple harvesting robot,” *Computers & Electrical Engineering*, vol. 38, no. 5, pp. 1186–1195, 2012.

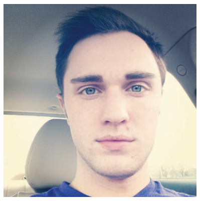
[48] R. Carloni, V. Lippiello, M. D’Auria, M. Fumagalli, A. Y. Mersha, S. Stramigioli, and B. Siciliano, “Robot vision: obstacle-avoidance techniques for unmanned aerial vehicles,” *Robotics & Automation Magazine, IEEE*, vol. 20, no. 4, pp. 22–31, 2013.

[49] M. A. Fischler and O. Firschein, *Readings in Computer Vision: Issues, Problem, Principles, and Paradigms*. Morgan Kaufmann, 2014.

[50] K. Alahari, D. Batra, S. Ramalingam, N. Paragios, and R. Zemel, “Guest editors introduction: Special section on higher order graphical models in computer vision,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 1321–1322, 2015.

[51] P. Forczmański and M. Furman, “Comparative analysis of benchmark datasets for face recognition algorithms verification,” in *Computer Vision and Graphics*. Springer, 2012, pp. 354–362.

[52] M. De Marsico and M. Nappi, “Face recognition in adverse conditions: A look at achieved advancements,” *Face Recognition in Adverse Conditions*, p. 388, 2014.



Ian Model is an undergraduate student of computer science at Lawrence Technological University.

Lior Shamir is an associate professor of computer science at Lawrence Technological University.