Detecting People in Cubist Art

Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik

University of California Berkeley

Abstract. Most evaluations of object detection methods focus on robustness to natural form deformations such as people's pose changes. However, the human visual system is surprisingly robust to artificial distortions as well. For example, Cubist paintings contain forms of man-made deformation to which human vision is tolerant such as perspective manipulation and, in particular, part-reorganization. We ask how would current object detection methods perform on these distorted images, and present an evaluation comparing human annotators to four state-of-the-art object detectors on Cubist paintings. Our results demonstrate that while human perception significantly outperforms current methods in object detection, human perception and part-based models exhibit a similarly graceful degradation in performance as objects become increasingly deformed, thus corroborating the theory of part-based object representation in the brain.

Keywords: recognition, deformations, paintings, cubism, perception

1 Introduction

In visual fine arts, such as painting and sculpture, artists often distort reality. In abstract art these distortions push the envelope of human perception while (usually) still containing figures that are recognizable by the viewer. Since these art forms characterize perception at its limit, we can use them to test computer vision methods for whether they achieve human-like performance and to design methods that mimic human vision. This stands in stark contrast to the common practice of evaluating computer vision models on datasets consisting only of camera snapshots. More abstract, artistic images often contain elements that the human vision system recognizes and interprets despite the lack of realism. This suggests that a corresponding computer vision method should have certain characteristics that enable it to mimic the recognition of these elements. As an example, we ask in this paper whether the characteristics of certain methods align well with the properties of human vision that enable the recognition of distortions found in Cubist art.

Cubist paintings depict objects as if they were seen from many viewpoints at once, breaking them into medium sized parts that appear out of their natural ordering and do not conform to the rules of perspective [1]. Because the deformed objects present in Cubist art are not normally seen in nature, the human visual

system must strain to recognize them. Since humans are usually still able to identify the depicted object, Cubism shows us that human perception does not rely on exact geometry and is tolerant to a rearrangement of mid-level object parts. However, findings from neuroscience show that this ability degrades as images become more scrambled or abstract [2][3]. If a method mimics human perceptual performance well, then it should behave similarly in the extreme conditions present in Cubist art. We aim to test whether there are computational systems that behave like human vision in the face of extreme form deformation.

We choose to focus on part-based methods that permit rearrangements of medium-complexity parts as they have been proven to do well at representing naturally occurring form deformations [4]. Here, we ask how they would perform in the face of abstract man-made distortions, and whether they would mimic human object detection performance better than object-level methods. We note that this is in itself a contribution, as there is little research into how current systems would react in novel circumstances. To this end, we compare human annotations of person figures in Picasso paintings to the detections of several object detection methods. Moreover, in order to chart the performance of the methods as human vision approaches its limit, we ask participants to divide the paintings into subsets according to the level of their abstractness and compare the performance of the humans and detection methods on each subset. Our results show that (1) existing part-based methods are relatively successful at detecting people even in distorted images, (2) that there is a natural correspondence between user ratings of image distortion and part-based method performance, and (3) that these properties are not nearly as evident in non-part-based methods. By demonstrating that part-based methods mimic human performance, we both show that these methods are valuable for object recognition in non-traditional settings, and corroborate the theory of part-based object representation in the brain.

2 Related Work

Since in most tasks the human visual system serves as an upper bound benchmark for computer vision, some studies focus on characterizing its capabilities at the limit. For instance, Sinha and Torralba examined face detection capabilities in low resolution, contrast negated and inverted images [5]. In other cases, computational models are used to test the validity of theories from neuroscience [6]. We take inspiration from these studies and evaluate human object detection in man-made art in order to provide a less restrictive benchmark of robustness to deformation than natural images. By using this benchmark we hope to discover parallels between the characteristics of human and algorithmic object detection.

From research in neuroscience, we know that the human visual system can detect and recognize objects even when they are deformed in various ways [7]. For instance, humans are able to recognize inverted objects, although their performance is degraded, especially when the objects are faces or words [8][5]. Similar results were obtained when comparing scrambled images to non-scrambled ones, leading to a theory of object-fragments rather than whole-object representations

in the brain [9][10]. This theory is strengthened by recordings from neurons in the macaque middle face patch that indicate both part-based and holistic face detection strategies [11]. Thus, although humans are capable of recognizing images distorted by scrambling, they are less adept at doing so. By analogy, we might expect methods trained on natural images to suffer a similar degradation in the face of the reorganization of object parts.

Object detection is one of the prominent unsolved problems in computer vision. Traditionally, object detection methods were holistic and template-based [12], but recent successful detection methods such as Poselets [13][14] and deformable part-based models [4] have focused on identifying mid-level parts in an appropriate configuration to indicate the presence of objects. Other part-based methods discover mid-level discriminative patches in an unsupervised way [15][16], use visual features of intermediate complexity for classification [6], or rely on distinctive sparse fragments for recognition [17]. Finally, a model inspired by Cubism itself that assembles intermediate-level parts even more loosely has shown success in detecting objects [18]. Another approach to detection that has recently shown remarkable detection results is based on convolutional neural networks [19][20]. We discuss the methods that we have chosen to benchmark in more detail in Section 4.

3 Cubist 'fragments of perception'

While there are many kinds of visual deformation to which human vision is robust, we choose to focus on the part-reorganization exhibited by Cubist paintings as it has an appealing correlation with the strengths of part-based detection methods. Cubism is an art movement that peaked in the early 20th century with the work of artists such as Picasso and Braque. Cubist painters moved away from the two-dimensional representation of perspective characteristic of realism [1]. Instead, they strove to capture the perceptual experience of viewing a three dimensional object from a close distance where in order to perceive an object, the viewer is forced to observe it part by part from different directions. Cubist painters collapsed these 'fragments of perception' onto one two-dimensional plane, distorting perspective so that the whole object can be viewed simultaneously. Despite the deformation of form, the original object is often readily detectable by the viewer as the parts chosen to represent it are naturalistic enough and discriminative enough to allow for the recognition of the object as a whole. However, this becomes harder with the degree of deformation and abstractness [3].

The fact that humans can detect distorted objects in Cubist paintings without prior training makes these paintings well-suited to benchmark robustness to deformation in detection methods trained on natural images. In order to provide intuition that part-based models will be able to perform well on this task, we provide some initial evidence that computer vision methods can successfully identify key parts in the Cubist depictions of objects. We train an unsupervised discovery method of mid-level discriminative patches on the PASCAL 2010 "person" class images [15][16][21], and compare the part-detector activations on natural


Fig. 1: Heat maps showing the discriminative patches activations on a natural training image (Left) and "Girl Before a Mirror 1932", a Picasso Cubist painting (Right). The color palette correlates with confidence score and ranges from blue (lowest) to red (highest). In both cases the most discriminative patches for class person are parts of faces and upper bodies, suggesting that computer vision methods are able to identify the key parts of human figures even when they are split into 'fragments of perception' in Cubist paintings.

training images and Cubist paintings by Picasso in Figures 1 and 2. Despite the difference in low-level image statistics, the detectors are able to discover the patches that discriminate people from non-people in both image domains. In the rest of the paper, we build on these results to test whether part-based object models can use the detected parts in order to recognize the depicted objects as a whole.

4 Object Detection Methods in Comparison

We compare four state-of-the-art person detectors on Cubist paintings, presented in the time ordering in which they were proposed: one holistic template-based method, one part-based model where the parts are learned automatically, one part-based model where the parts are learned from human annotations and the most recent deep learning method. Here we discuss the details of each one.

Dalal and Triggs: Object appearance in images can be characterized by histograms of orientations of local edge gradients binned over a dense image grid (HOG) [12]. The Dalal and Triggs (D&T) method trains an object-level HOG template for detection using bounding box annotations. Since the features are binned, the detector is robust only to image deformation within the bins.

Deformable Part Models: A holistic HOG template cannot recognize objects in the face of non-rigid deformations such as varied pose, which result in a rearrangement of the limbs versus the torso. Therefore, the deformable part-based models detection method (DPM) represents objects using collections of part models [4]. These can be arranged with respect to the root part, representing the full object, in deformable configurations that can be characterized by spring-like connections between certain pairs of parts. In practice, the model trained on natural images often learns sub-part models (such as half a face).

[1] For each detector, the 10 most confident activations are taken from the test set of paintings and presented by decreasing confidence, excluding lower confidence duplicates of 50% overlap or more. Detectors are sorted by the average score of their 20 top activations, excluding detectors where over 1/4 of activations are duplicates of activations from higher rated detectors.

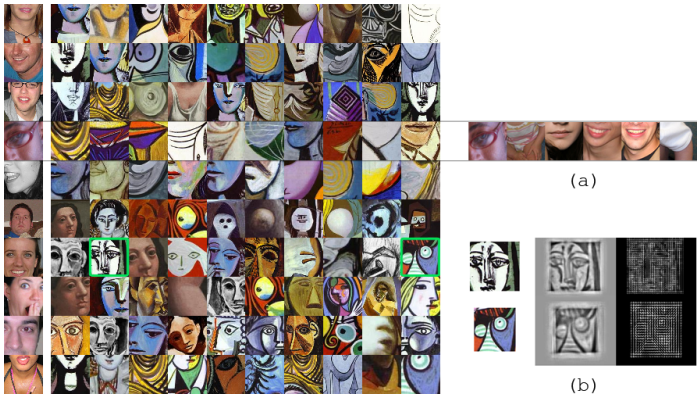


Fig. 2: Computer vision methods trained on natural images are able to detect discriminative parts of person figures in Cubist paintings. (Left) Each row displays the top ten discriminative patches activations[1]. The leftmost column shows the top activation on the training data. Most of the detectors find corresponding face parts in the natural and painting images, although many are false positives. The fourth patch-detector from the top detects patches with little visual consistency on the paintings as well as the training data (a). Some false positive activations (b)(Bottom) seem more similar to true positives (b)(Top) in Hoggles space [22] (b)(Middle) than in HOG (b)(Right) or the original RGB (Left, marked in green) spaces.

Poselets: Poselets is a similar part-based model that considers extra human supervision during training[13][14]. Here, parts are not discovered but learned from body-part annotations. Poselets do not necessarily correspond to anatomical body parts like limbs or torsos as these are often not the most salient features for visual recognition. In fact, a highly discriminative Poselet for person detection corresponds to "half of a frontal face and a left shoulder".

R-CNN: R-CNN replaces the earlier rigid HOG features with features learned by a deep convolutional neural network [19][23]. While R-CNN does not have an explicit representation of parts, it is trained under a detection objective to be invariant to deformations of objects by using a large amount of data. Deep methods outperform previous algorithms by a large margin on natural data. Here we compare this state-of-the-art method to other methods and human perception on abstract paintings.

5 Experimental Setup

We compare the above methods to human perception in two ways. First, we study the human and algorithm performance on person detection over a corpus of Cubist paintings. Second, we examine the degradation in performance of human perception and detectors as the distortion of person figures in the paintings increases. We conduct all comparisons using the PASCAL VOC evaluation

mechanism, in which true positives are selected based on a 50% overlap between detection and ground truth bounding box [21].

5.1 Picasso Dataset

In the experiments described below we used a set of 218 Picasso paintings, which have a title indicating that they depict people, as our test data. These ranged from naturalistic portraits to abstract depictions of people as a collection of distorted parts. The set of paintings we used is highly biased in comparison to PASCAL person class images [21]. Given the nature of the art form, Cubist paintings usually depict people in full frontal or portrait views. Since in a PASCAL-style evaluation true positives are selected based on a 50% overlap, this results in higher average precision scores. For example, portrait paintings devote most of the canvas area to the torso of a person. In such paintings, a random detection that contains over 50% of the image would count as a true positive. This issue exists in any PASCAL VOC evaluation, but it is especially pronounced in this case.

5.2 Human Perception Study Setup

We conducted two experiments as part of our perception study. First, we recorded human detections of person figures in Cubist paintings. Second, we asked participants to bucket the paintings by the degree to which the figures in them were deformed compared to photorealistic depictions of people. For each painting, raters were asked to pick a classification on a 5-point Likert scale, where 1 corresponded to "*The figures in this painting are very lifelike*" and 5 corresponded to "*The figures in this painting are not at all lifelike*".

Participants We recruited eighteen participants to partake in our perception study. Sixteen participants were undergraduate students at our institution, one was a graduate student and one a software engineering professional. Seventeen participants were male and one was female.

Mechanism Participants completed the study on their personal laptops using an online graphical annotation tool we wrote for the purpose. Each participant spent an hour on the study and received a compensation of $15. Each participant annotated 146 randomly chosen paintings out of the total 218, so that every painting was annotated by 14 - 15 unique participants.

5.3 Detector Study Setup

We compare the human recognition performance we measured during the perception study to four object detection methods. We train all methods using the PASCAL 2010 "person" class training and validation images [21]. We train the

methods using natural images so that they do not enjoy an advantage over humans by training on the paintings. However, some research suggests that human recognition in Cubist paintings does improve with repeated exposures [2]. We set all parameters in all four methods to the same settings used in the original papers, except for the Poselets detection score which we set to 0.2 based on cross validation on the training data.

5.4 Ground Truth

Because Picasso did not explicitly label the human figures in his paintings, there is no clear cut gold standard for human figure annotations in our image corpus. As a result, we rely on our human participants to form a ground truth annotation set. We do so by capturing the average rater annotation as follows. Since each painting might have more than one human figure, we use k-means clustering to group annotations by the human figure they correspond to. For each cluster, we obtain a ground truth bounding box by taking the median of each corner of the bounding boxes in that cluster along every dimension. This yields one ground truth bounding box per human figure per image, which we can now use to evaluate both human and detector annotations. There is one subtlety omitted in the above description: since it would be unfair to allow a human rater's annotations to influence ground truth when evaluating that rater, for each human rater we construct a modified leave-one-out ground truth from the annotations of all other raters, withholding the annotations of the rater we intend to evaluate. When evaluating detectors, however, we include annotations from all human raters in the ground truth.

It is worth noting that since humans themselves are error-prone (especially when recognizing objects in more abstract paintings), our ground truth cannot be a perfect oracle. Rather, all evaluation is comparing performance to the average, imperfect human. From this perspective, our evaluation of human raters can be seen as a measure of their agreement, and our evaluation of detectors can be seen as a measure of similarity to the average human.

5.5 Evaluating Humans and Detectors

Unlike detectors, humans provide only one annotation per figure without confidence scores, so we cannot compute average precision. Our primary metric for humans is the F-measure (F1 score), which is the harmonic mean of precision and recall. In order to combine the F-measures of all participants, we consider both (qualitatively) the shape of the distribution of the scores, and (quantitatively) the mean F-measure.

To compare detectors with humans, we pick the point on the methods' precision-recall curve that optimizes F-measure, and return the precision, recall, and F-measure computed at this point. This is generous to the detectors but captures their performance if they were well tuned for this task.

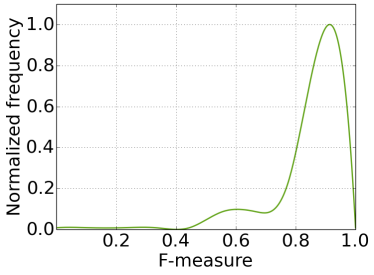


Fig. 3: Frequency distribution of human F-measure scores for recognizing person figures in all 218 Cubist paintings against the leave-one-out ground truth bounding boxes. Due to the small number of participants, the curve has been smoothed for clarity.

6 Detection Performance on the Picasso Dataset

In the first part of the comparison we evaluate the performance of the humans and the four methods at detecting human figures in Picasso paintings.

6.1 Human Performance on the Picasso Dataset

First, we evaluate our human participants against the leave-one-out ground truth to determine how effective people are at recognizing human figures in Cubist art. Figure 3 displays the distribution of human F-measures for this task. Qualitatively, we see that humans perform quite well, as the distribution has its peak around 0.9, and there is little variance among the scores. It is worth noting the bump in the distribution around 0.6—there were a few raters whose annotations were significantly different from the ground truth annotation. This may have been due to either failure to recognize the images, or a misunderstanding of the annotation interface and instructions. Quantitatively, the first row of the table in Figure 7(Right) shows the mean human precision, recall, and F-measure. These numbers confirm our impressions of the distribution–humans score high on average when detecting human figures in the paintings.

6.2 Detector Performance on the Picasso Dataset

Qualitative Results Part-based models trained purely on photographs performed surprisingly well when tested against the Picasso data. As can be seen in Figures 4 and Figure 5, Poselets and DPMs successfully produce bounding boxes for figures in the paintings, though they have their fair share of false positives and misses. We additionally note semantically significant Poselet activations as well as false positives in tricky cases where non-human image elements take on human shapes–an expected consequence of the complexity of the paintings (Figure 5). The non-part-based methods do not perform as well, as is evident in Figure 6 where each row displays the top ten detections of a method over the entire painting dataset, sorted by each method's confidence score.

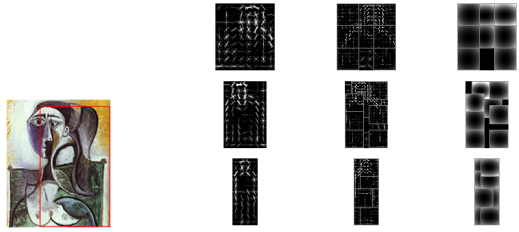


Fig. 4: (Left) DPM detects one of the two viewpoints of the split face, resulting in a shifted localization of the person as a whole. (Right) The DPM model trained on natural images learns sub-face parts in three different scales. This provides insight as to why DPM is able to detect a split-face patch resulting in the bounding box detection on the left.

Fig. 5: The Poselets method is able to find person-parts in Cubist paintings and use them to detect person figures as a whole. (Left) Poselet bounding box detections in Picasso's "Girl Before a Mirror 1932" with a false positive detection in the center. (Middle) A true positive Poselet activation on the painting together with the corresponding Poselet activation on the training data. In both image domains the Poselet detects a downward-angled face in profile. (Right) The false positive activation that results in the incorrect bounding box detection on the left. Here the Poselet falsely detects an away-facing back, shoulders, and head in the painting.

Quantitative Results Figure 7(Right) displays the precision-recall curves for each method when evaluated on all paintings. There is a clear ordering in accuracy: DPM performs the best by far, Poselets and RCNN come next, and Dalal and Triggs, though characterized by high recall, has extremely low precision, leading to terrible performance. In general, all of the methods achieve recall of up to 0.8, which implies that they are capable of recognizing most human figures in the image. However, DPM is the only method which can maintain a reasonable precision as recall increases, which explains its significantly greater AP. The table in Figure 7(Right) confirms these insights. DPM's performance on this task is quite encouraging, as its AP (0.38) is not too far off from its performance on undistorted PASCAL 2010 photographs (0.41 without context rescoring) although the actual number may be higher in practice as the PASCAL test dataset, unlike ours, includes many images that do not contain people [23].

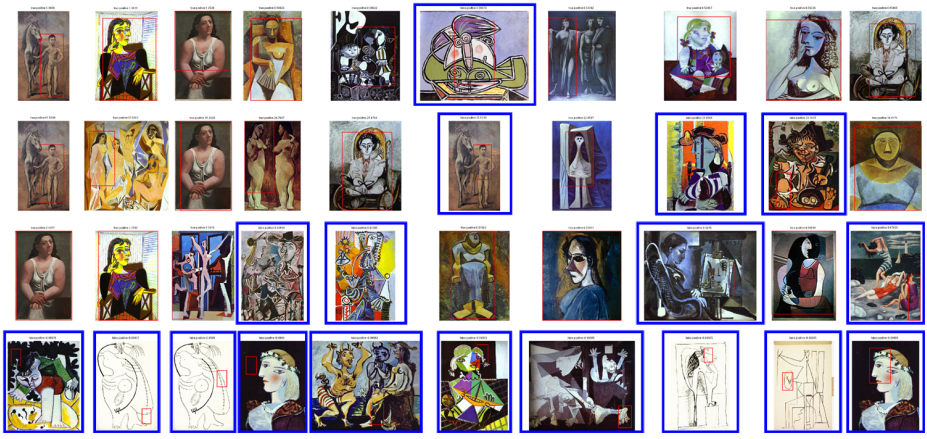


Fig. 6: Top ten detections for each method according to confidence from left to right. First row: DPM. Second row: Poselets. Third row: R-CNN. Fourth row: D&T. False positives are marked in blue. Qualitatively, it is evident that part-based models outperform the other approaches.

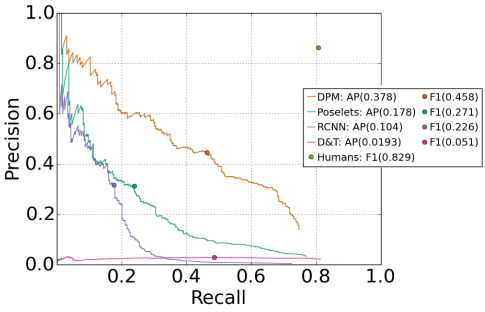
6.3 Comparing Human and Detector Performance

As is clear from the graphical and tabular data in Figure 7, humans are far better than detectors at recognizing person figures in Picasso paintings. The green dot in the upper right corner of the graph shows human precision and recall to be far higher than the orange dot of DPM, the highest-performing method.

6.4 Discussion of Performance on the Picasso Dataset

The comparison between the human participants and four methods on the task of detecting person figures in Picasso paintings demonstrates clearly that humans are highly skilled at this task, and that detectors are much less effective but can still achieve results within an order of magnitude of their detection performance on undistorted images. Among algorithms, part-based object detection methods perform better than object-level methods on distorted images. DPM and Poselets, our two part-based methods, demonstrate the best performance on the object detection task. This is likely due to the fact that part-based methods are able to recognize medium-level parts that remain intact even in Cubist paintings where standard human body parts are highly distorted. We emphasize the part-based approach here, as we have no reason to believe that the HOG features used by both DPM and Poselets carry any advantage over other image features organized in a part-based model.

Given its success in object detection on natural images, it is interesting that R-CNN does not perform well on this task. One reason for this could be that R-CNN is not a part-based model, however this is only partly true because the



Annotator	Precision	Recall	F-measure	AP
Human	0.804	0.860	0.829	N/A
DPM	0.444	0.464	0.458	0.378
Poselets	0.311	0.240	0.271	0.178
RCNN	0.315	0.177	0.226	0.104
D&T	0.027	0.486	0.051	0.019

Fig. 7: Performance comparison via precision-recall curves (Left) and tabular data (Right). In both the plot and the table: for detectors, precision, recall, and F-measure are the maxima over the entire precision-recall curve. For humans, these numbers are averages across the raters. While DPM outperforms other methods, none of the methods reach human performance in person detection in Cubist paintings.

convolutional filters can be thought of as parts, and the max pooling as performing deformations. A second factor might be the fact that R-CNN over fits to the natural visual world and fails at adapting to the domain of paintings. There has been little research into how CNN-based networks perform on deformed images, but an initial investigation suggests that tiny changes to an image may cause drastic changes to their output [24].

7 Performance Degradation with Increased Deformation

In the second part of the comparison we study the degradation of performance of humans and methods with increased painting abstractness.

7.1 Classifying Images by Degree of Deformation

In our user study (described in Section 5.2), each rater labeled images on a scale from 1 (*The figures in this painting are very lifelike*) to 5 (*The figures in this painting are not at all lifelike*). By taking the rounded average of user labels for each image, we divide the images into five 'deformation buckets' in order to evaluate object detection performance as image distortion increases. An example of a painting with an average rating of 1 is Picasso's "Seated Woman 1921"', and an example of a painting with an average rating of 5 is Picasso's "Nude and Still Life 1931" (copyrighted paintings not reproduced in this text). The histogram of the number of images in each bucket is shown in Figure 8(Top Left).

7.2 Human Performance Degradation

Figure 8(Top Right) demonstrates the impact of deformation on human object detection performance. As the images become more distorted, the distribution

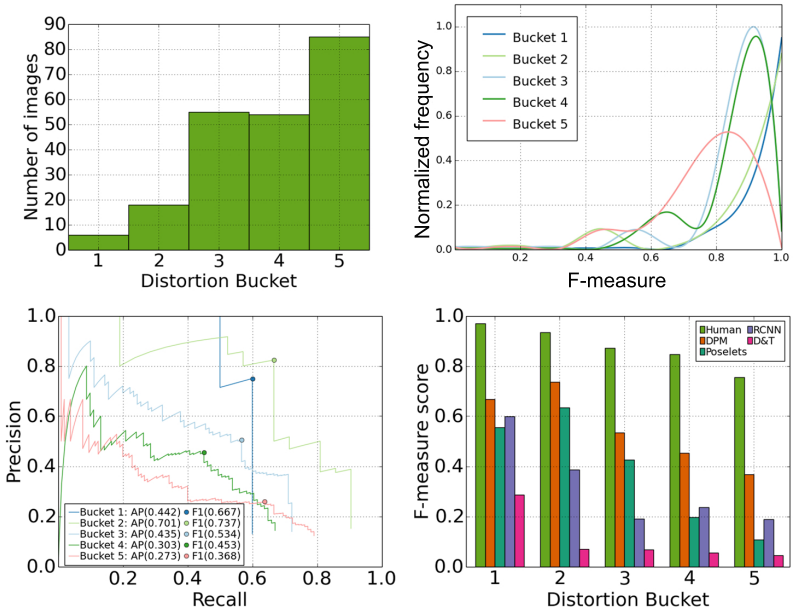


Fig. 8: Impact of increasing image distortion on object detection performance. Bucket 1 contains images that appear most natural, and bucket 5 contains images that appear most distorted. (Top Left) A histogram showing the number of images per bucket. (Top Right) Human F-measure distributions by distortion bucket. Because there were a small number of human raters, the curves have been smoothed for clarity. (Bottom Left) DPM precision-recall curves by distortion bucket. (Bottom Right) Comparing humans and methods by bucket. Here, part-based models show a similar degradation behavior to human performance as the distortion of the image increases.

of human F-measures shifts clearly to the left. This indicates that human performance worsens on more deformed human figures, which is consistent with previous results [8][5]. As noted in Section 6, there are a few annotators with low F-measures in each of the curves, which implies that these raters' errors are independent of image distortion. This supports the theory that they are due to a failure to follow the instructions rather than an inability to recognize objects.

7.3 Detector Performance Degradation

Qualitative Results In Figure 9 we compare the top detections per method on paintings from bucket 2 versus bucket 5. The top discriminative-patches activations on these two buckets (Right) help visualize the difficulty in detecting meaningful mid-level parts as the paintings become more abstract.

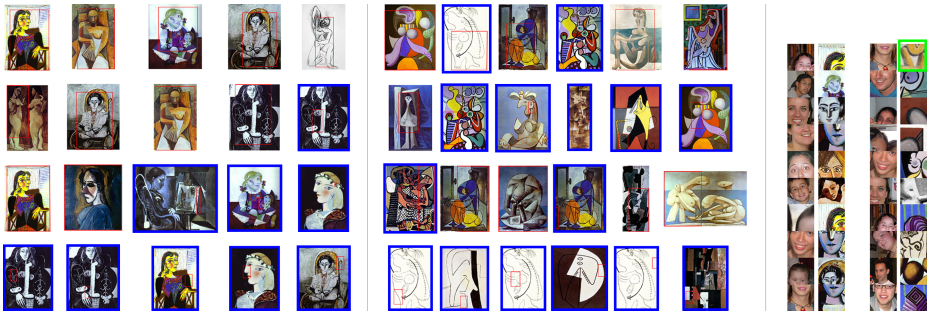


Fig. 9: The degradation in performance with image difficulty. Top five detections per method (rows correspond to: DPM, Poselets, R-CNN, D&T) on images from bucket 2 (Left) and bucket 5 (Middle). False positives are marked in blue. This comparison shows that all detection methods perform worse on more distorted images. (Right) The degradation in performance is also evident in the detection of the parts in isolation. Top 10 discriminative patches activations for images of bucket 2 and bucket 5, with corresponding activations on PASCAL images. All top activations on bucket 2 are true positives compared to only one from bucket 5 (in green).

Quantitative Results Figure 8(Bottom Left) shows the precision recall curves for the DPM method with varying distortion buckets. As with human performance, increasing the image distortion causes a pronounced decrease in method performance. The overlap between the curves for buckets 1 and 2 may be due to variance as a consequence of the low number of images in those two buckets (see Figure 8(Top Left)).

7.4 Comparing Human and Detector Degradation

Figure 8(Bottom Right) shows the F-measures for humans and each method with varying distortion buckets. This figure shows that detector performance degrades in a similar pattern to human performance as images become increasingly distorted. A closer look at the degradation patterns reveals that the part-based methods follow the human pattern most closely, which matches our intuition about the similarities between part-based object detection and the human visual system. In contrast, the template-based Dalal and Triggs method abruptly breaks down after bucket 1.

7.5 Discussion of Degradation with Increased Deformation

As we have demonstrated, part-based models for object detection show a smooth degradation in precision and recall as the distortion of the images increases. This is consistent with results from neuroscience, which indicate that humans are capable of detecting objects cut into parts, but that their ability degrades

significantly when the parts are scrambled. The correspondence between human and computational method performance on this task suggests that a part-based object representation might be a good approximation for the mechanisms of object detection in the human brain. The ability to model these mechanisms computationally further corroborates the neuroscience theory of part-based object detection strategies. This is encouraging, even though current methods cannot yet perform at the level of human vision.

We note that the correspondence between part-based models and human perception is a somewhat surprising one. At their core, the methods we used are based on HOG features that we expected to be highly dependent on image statistics. It was pleasantly surprising to observe the correlation between these methods, trained on natural images, and human perception on Cubist paintings with completely different statistics. We believe that better performance could be achieved using part-based models that rely on higher level features than HOG.

8 Conclusion

Since human perception acts as a natural benchmark for computer vision, we should strive to understand its performance at its limits. In this paper, we have argued that object detection under form deformation is an example of a challenging perception task that existing image benchmarks do not properly evaluate. We have proposed the use of Cubist paintings as a more appropriate benchmark, as they contain rearranged object parts that are nevertheless recognizable as whole objects. Using this benchmark, our evaluation comparing human performance to that of various object detection methods demonstrates that part-based models are a step in the right direction for modeling human robustness to object deformation, since their performance degrades comparably to humans as the abstractness in images increases. By showing that these models can be trained on photographic data yet still perform on abstract data we demonstrate that they are less over-fit to the natural world than template-based and deep models.

Part-reorganization in Cubism is one example that pushes the envelope of human perception, but there are other artistic movements with characteristic deformations, such as the use of blurring in Impressionism, that would provide rich grounds for study. Future work in the design of computer vision methods should be cognizant of the limitations of traditional camera snapshot datasets and look for complementary resources when evaluating computational methods' ability to mimic the wide range of human perception.

9 Acknowledgments

The authors would like to thank Mark Lescroart for his guidance and advice throughout this project and Bharath Hariharan, Carl Doersch and Katerina Fragkiadaki for their insightful comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1106400.

References

1. Laporte, P.M.: Cubism and science. The Journal of Aesthetics and Art Criticism **7**(3) (1949) pp. 243–256
2. Wiesmann, M., Ishai, A.: Training facilitates object recognition in cubist paintings. Frontiers in Human Neuroscience **4** (2010) 11
3. Ishai, A., Fairhall, S.L., Pepperell, R.: Perception, memory and aesthetics of indeterminate art. Brain Research Bulletin **73**(46) (2007) 319 – 324
4. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. Pattern Analysis and Machine Intelligence (PAMI) **32**(9) (2010)
5. Sinha, P., Torralba, A.: Detecting faces in impoverished images. Journal of Vision **2**(7) (November 2002)
6. Ullman, S., Vidal-Naquet, M., Sali, E.: Visual features of intermediate complexity and their use in classification. Nature Neuroscience **5**(7) (July 2002) 682–687
7. Lewis, M.B., Edmonds, A.J.: Face detection: Mapping human performance. Perception **32**(8) (2003) 903–920
8. Tsao, D.Y., Livingstone, M.S.: Mechanisms of face perception. Annual Review of Neuroscience **31** (2008) 411–437
9. Grill-Spector, K., Kushnir, T., Hendler, T.: A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Human Brain Mapping **6**(4) (1998) 316–328
10. Vogels, R.: Effect of image scrambling on inferior temporal cortical responses. Neuroreport **10**(9) (June 1999) 1811–1816
11. Freiwald, W.A., Tsao, D.Y., Livingstone, M.S.: A Face Feature Space in the Macaque Temporal Lobe. Nature Neuroscience **12**(9) (September 2009) 1187–1196
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2. (2005) 886–893
13. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2009) 1365–1372
14. Bourdev, L.D., Maji, S., Brox, T., Malik, J.: Detecting People Using Mutually Consistent Poselet Activations. In: Proceedings of the European Conference on Computer Vision (ECCV). (2010) 168–181
15. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Proceedings of the European Conference on Computer Vision (ECCV). (2012)
16. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? ACM Transactions on Graphics (SIGGRAPH) **31**(4) (2012)
17. Akselrod-Ballin, A., Ullman, S.: Distinctive and compact features. Image and Vision Computing **26**(9) (September 2008) 1269–1276
18. Nelson, R.C., Selinger, A.: A Cubist Approach to Object Recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (1998) 614–621
19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
20. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR 2014), CBLS (2014)

21. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
22. Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A.: HOGgles: Visualizing Object Detection Features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2013)
23. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/
24. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. CoRR **abs/1312.6199** (2013)