To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

Finding the Weakest Link in Person Detectors

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

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

Abstract

Detecting people remains a popular and challenging problem in computer vision. In this paper, we analyze parts-based models for person detection to determine which components of their pipeline could benefit the most if improved. We accomplish this task by studying numerous detectors formed from combinations of components performed by human subjects and machines. The parts-based model we study can be roughly broken into four components: feature detection, part detection, spatial part scoring and contextual reasoning including non-maximal suppression. Our experiments conclude that part detection is the weakest link for challenging person detection datasets. Non-maximal suppression and context can also significantly boost performance. However, the use of human or machine spatial models does not significantly or consistently affect detection accuracy.

1. Introduction

Object detection remains an open and challenging problem in computer vision. Historically, the subclass of detecting people has attracted increased attention given its importance to many real world applications, and its challenging level of difficulty. The wide variety of poses and shapes people exhibit, along with variations in clothing, creates a very challenging task for modeling and learning algorithms.

Recently, person detectors have made significant progress using part-based models. The appearance of each part such as a person's head, foot, or torso are represented by Histograms of Gradients (HoG) [5, 13], color [7] or Harr wavelets [8]. The spatial relationships of object parts can be represented using trees [13], k-fans [4] or constellation models [15]. Each of these approaches propose a complex set of interdependent components to provide final detection results. While the additional complexity of the approaches have led to increased performance, understanding the role of each component in the final detection accuracy is difficult.

In this paper, we propose a thorough analysis of parts-based models to gain insight into which components of the pipeline could benefit the most if improved. We accomplish this task by using human subjects to perform the individual components previously performed by the machine algorithm. For instance, instead of using a machine classifier such as a latent SVM trained on HoG descriptors [13] to detect object parts, we use human subjects to label whether a small image patch contains a human's head, foot, torso etc.

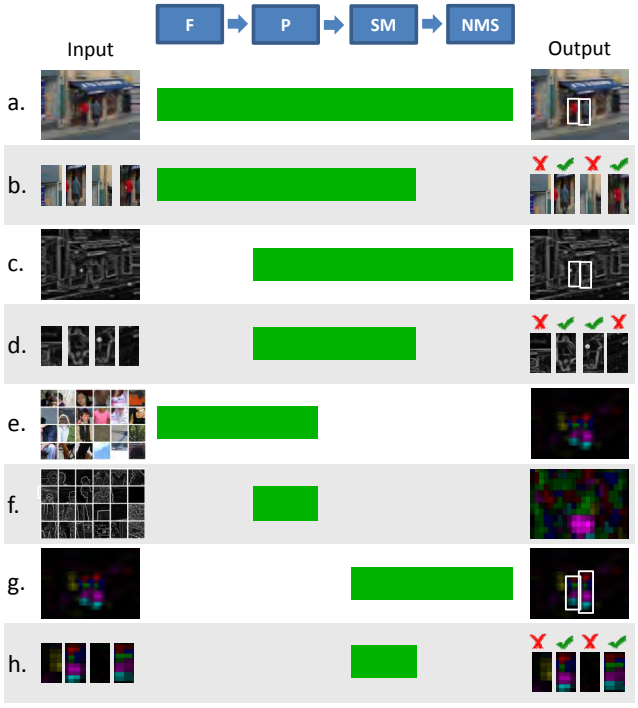


Figure 1: In order to gain insight into which components of a parts-based person detector could benefit the most if improved, we replace each component, *i.e.*, part detection (P), feature extract (F), spatial modeling (SM) and non-maxima-suppresion (NMS) in the pipeline with human subjects (green bars). Here we illustrate the various tasks performed by human subjects via example input/output pairs.

A parts-based detector can be roughly broken into four components: feature detection, part detection, spatial part scoring and contextual reasoning including non-maximal suppression. We combine numerous human and machine performed components to form complete person detectors and recognizers. The results indicate which components

lead to the greatest increase in accuracy over the standard machine approach. The experiments include the use of various feature types such as color, edges, and intensities for both detecting people and parts, the use of human detected parts with a machine spatial model, and machine detected parts with using a human's spatial model. Other experiments analyzing non-maximal suppression techniques and contextual information are also performed.

2. Related Work

We now discuss some existing techniques for person detection, as well works that conduct human studies to gain insights in computer vision. We comprehensively discuss works on parts, spatial models and contextual models in Section 3.

Person/Pedestrian detection: Given the importance of detecting people in images, numerous detectors have been proposed. A comparison of several approaches for pedestrian detection can be found in Dollar *et al*. [9]. Wojek *et al*. [44] analyzes several features and classifier types. Dalal and Triggs [5] first proposed the locally normalized histogram of gradients detector, which was improved upon by Felzenszwalb *et al*. [13] using deformable parts models. Increased performance was found using numerous feature types and boosting by Dollar *et al*. [7], and using multi-level features and intersection kernel SVMs by Maji *et al*. [23].

Human studies: An early example of designing computational models with similar behavior to humans is shown in David Marr's book [25]. Liu et al. [21] conducted human studies demonstrating that the high human performance in 3D object discrimination can only be explained if humans are using 3D information. Tarr *et al*. [38] and Hinton *et al*. [19] studied whether humans use mental rotation for recognition and determining if shapes have the same handiness. A comparison of human and machine algorithms for selecting regions-of-interest in images was conducted by Privitera *et al*. [32]. Fei-Fei *et al*. [11] demonstrated that human subjects can provide a large amount of detailed information about a scene even after viewing it for a very brief period of time. Bachmann *et al*. [2] show that humans can reliably recognize faces in images as small as 16×16 pixels, and Oliva *et al*. [27] present similar results for scene recognition. Torralba *et al*. [40] and Parikh *et al*. [30] show that humans can detect objects in 32×32 images with significantly higher performance than state-of-the-art machine algorithms using high resolution images. The work of Parikh *et al*. [29] uses human studies to determine if features, classification algorithms or the amount of training data is most likely to account for the superiority of humans over machines in recognizing objects and scenes.

3. Part-based detector

In this section, we describe machine models for various components in a part-based detector including feature extraction, parts modeling, spatial models, non-maximal suppression and contextual reasoning, before we describe the corresponding set-up for our human studies. For each stage, we follow the approach of Felzenszwalb *et al*. [13] that has shown recent state-of-the-art performance and briefly outline other approaches. Our studies are performed on subsets of the commonly used INRIA [5] dataset, and the more challenging PASCAL [10] dataset.

3.1. Feature extraction and modeling parts

Numerous low-level features and representations have been proposed for modeling objects and their parts. Representations have progressed from modeling textures [24] to histograms of gradients with global normalization [22] and local normalization [5]. The work of Felzenszwalb *et al*. [13] improved upon [5] to reduce its dimensionality and increase accuracy. Methods using color [7] and gradients without histograms [34] have also been proposed. Wavelet approaches [28, 42] have shown benefits in computational efficiency. Several methods combine various features using decision trees [41] or boosting techniques [7]. Representations may also be learned using random decision forests [37], feature mining [8], deep belief nets [18], mixture models [3] or biologically inspired models [26].

In this paper, we use the part detectors of Felzenszwalb *et al*. [13] trained via a latent SVM on histogram of oriented gradient features [5]. The models were pre-trained and supplied by Felzenzswalb *et al*. [13]. Each component of the model contains a root filter and six part filters. While [13] provides two component models, we only used one component, since slightly better results were achieved using a single component model on the datasets used in this paper. The part detections were obtained by independently applying the part filters.

3.2. Spatial model

The spatial relationship of parts can be modeled using several previously proposed techniques. Constellation models [15, 43] use Gaussian distributions to represent the relative positions of parts. More restrictive, but computationally efficient methods have been proposed using tree [14, 16], and k-fan models [4]. Tree-based deformable models called pictorial structures [14, 17] provide both efficient detection and learning. The 3D appearance of objects may also be represented using multiple templates [36], aspect graphs [31] or by linking parts from different viewpoints [35].

We use a star-graph spatial model similar to Felzenszwalb *et al*. [13]. The model assumes that the location of the parts are independent given the location of the person.

The locations of the parts relative to the person are modeled via a Gaussian distribution, with mean and co-variance parameters (μ_i, Σ_i) for the i^{th} part. All co-ordinates are normalized with respect to the hypothesized size of the person. Each person candidate window is scored as:

$$s = \sum_{i=1}^{K} \max_{x,y} s_i(x, y) \quad (1)$$

where K is the number of parts in the model, and s_i is the score associated with part i at location (x, y):

$$s_i(x, y) = \hat{s}_i(x, y) - (a_i \Delta x + b_i \Delta y + c_i (\Delta x)^2 + d_i (\Delta y)^2) \quad (2)$$

where $\hat{s}_i(x, y)$ is the score of the i^{th}part detector at location (x, y), and Δx and Δy are the positional offsets from the part's mean position. The coefficients $\{a_i, b_i, c_i, d_i\}$, which model the covariance, are learnt discriminatively via a linear-SVM to distinguish positive windows across the training dataset (with $> 50\%$ overlap with a ground-truth person bounding-box) from the negative windows (with $<$ 50% overlap with a ground-truth person bounding-box). The mean parameters of the star-graph μ_i are learnt through maximum likelihood estimation over the positive training windows using part detections that maximize $s_i(x, y)$. With the newly estimated mean parameters, a new set of covariance coefficients $\{a_i, b_i, c_i, d_i\}$ are learnt, resulting in an iterative learning procedure.

We initialize μ_i as the weighted mean of the part detections within the ground-truth person bounding-boxes in the training images. The weights correspond to the part detection scores. The coefficients $\{a_i, b_i, c_i, d_i\}$ are initialized to $\{0, 0, 0.3, 0.3\}$.

3.3. Context and non-maximal suppression

Recently, the use of context has received significant attention for object recognition and detection. Context provides a useful aid for determining likely positions of objects using scene information [39] or the location of other objects [20]. Pairwise interactions of objects can be modeled using CRFs [30, 33] or as a max-margin learning problem [6, 13].

The related problem of Non-Maximal Suppression (NMS) attempts to remove redundant detections of the same object. This can be viewed as contextual information shared between objects of the same class, i.e. two of the same object cannot typically occupy overlapping areas of the image. In fact, some approaches inherently solve NMS in their multi-object contextual models [6].

In this paper, we only use NMS and not more complex contextual models, as the performance gains provided by the complex models were minimal [6, 13] on the PASCAL dataset. We performed NMS by removing windows that overlapped with a higher scoring window. Overlap is computed as the ratio of the intersection and union of the two windows. We used an overlap threshold of 0.3.

4. Experimental setups for human studies

Our experiments involve replacing various components of the person-detector pipeline with human subjects. In this section, we describe the techniques we employ. The green bars in Figure 1 illustrate the various human studies we performed. For human testing we broke the pipeline into four stages; feature extraction, part detection, spatial modeling and NMS/context. There are 10 possible combinations of contiguous stages that the human could perform, of which we test 8. We do not perform the feature extraction stage alone, since we cannot get direct access to the features extracted by the human brain for further processing with a machine. In addition, we do not perform the NMS/context stage alone. All our human studies were performed on Amazon Mechanical Turk.

Our experiments were conducted on 50 INRIA and 100 PASCAL 2007 images containing 132 and 139 people respectively. We hand-labeled all the faces in the images, and re-scaled the images so that the faces were a canonical size. Fixing the scale reduces our search space making our human studies feasible. The machine implementation [12] with fixed scale gave an Average Precision (AP) of 0.7146 for our 50 INRIA images and 0.4625 for our 100 PASCAL images.

4.1. Feature extraction (F)

Given a natural image, human subjects may extract any low-level features necessary for recognition. However, if we pre-process images to retain only some of the information, we can constrain the low-level features accessible to the human subjects. In our experiments, we show subjects grey-scale images, normalized gradient images and colored images at both high and low-resolutions. A normalized gradient $\hat{g}(x, y)$ at pixel (x, y) with gradient $g(x, y)$ is computed as follows:

$$\hat{g}(x, y) = \frac{g(x, y)}{\bar{g}(x, y) + \epsilon} \quad (3)$$

where $\bar{g}(x, y)$ is a Gaussian weighted average with a standard deviation of 5, and $\epsilon = 4$ is used to ensure $\bar{g}(x, y)$ is above the level of noise. For visibility, the maximum normalized gradient within a patch is scaled to 255, see Figures 3 and 2 for examples. Figure 1 illustrates the settings where human subjects use their internal feature-extractor (a,b,e) or are constrained by machine extracted features (c,d,f).

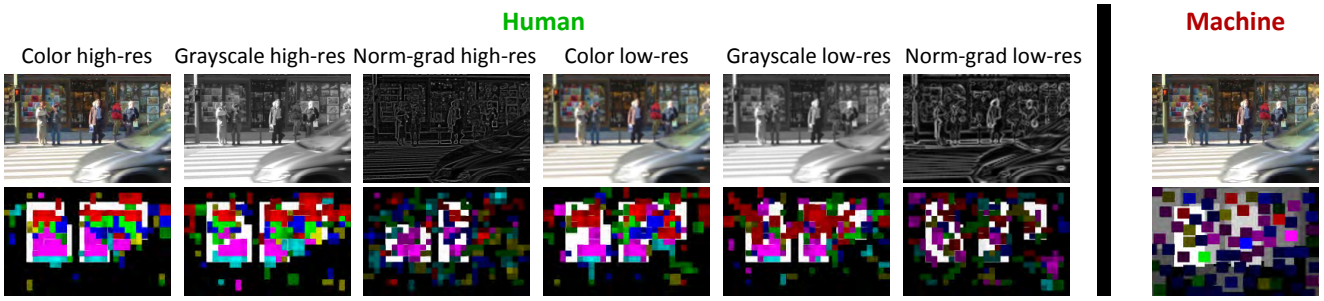
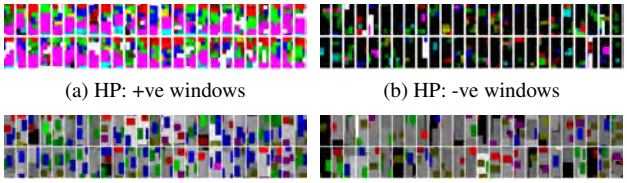


Figure 2: Part-detection visualizations created for human and machine detected parts.

Figure 3: Example patches classified by humans as (left to right) head, torso, legs and background in (top to bottom) regular, grey-scale low resolution and normalized gradient images.



(a) HP: +ve windows (b) HP: -ve windows

(c) MP: +ve windows (d) MP: -ve windows

Figure 4: Training data with positive (+ve) windows containing a person (left) and negative (-ve) windows not containing a person (right) shown to human subjects for learning a spatial model. Windows are cropped from part visualizations for (top) human detected parts (HP) and (bottom) machine detected parts (MP) in gray-scale PASCAL images.

4.2. Part detector (P)

Similar to the machine part detector, our human studies use a sliding window approach. Overlapping small patches are extracted from the images (Figure 1 (e, f)). Human subjects were randomly shown these patches across all images, so no contextual information was available. Subjects were asked to classify each patch as containing a head, torso, arm, hand, leg, foot, any other part of a person, or not a person at all. Each patch was classified by 10 subjects. Example patches shown to humans using color, grey-scale and normalized gradient images are shown in Figure 3. Visualizations of the detected parts aggregated across subjects are shown in Figures 1 and 2. The different colors correspond to different parts (red:head, blue:torso, green:arm, yellow:hand, magenta:leg and cyan:feet). The intensity of the color corresponds to the number of subjects that classified the local patch as the corresponding part. Analogous to the detector of Felzenszwalb *et al*. [13], we also detect "roots" in a similar sliding window fashion, which are low-resolution templates of a person (shown in white in Figure 2)[1].

We used root and part sizes similar to the machine implementation. Specifically, for INRIA the part (patch) sizes extracted from images were 38×38 pixels and the root (window) sizes were 50×150. For PASCAL, the part sizes were 80×80 and the root sizes were 114×314. In Felzenszwalb *et al*. [13] the spatial resolution of the root is lower than that of the other parts. Similarly, we downsample the root windows, leading to an effective resolution of 25×75 for INRIA and 29×80 for PASCAL. For low-resolution part detection in both INRIA and PASCAL, the resolution of the parts was reduced to 20×20, and the roots were scaled to 24 pixels in the largest dimension. For easy viewing, the parts and roots were displayed to subjects with the largest dimension scaled to 80 pixels. The part patches were sampled with 50% overlap between consecutive patches, and the root windows were sampled at 75% overlap.

4.3. Spatial model (SM)

To study the ability of the human subjects to reason about spatial relationships, we train the subjects using the colored part-visualizations, as shown in Figure 4. Subjects are then asked to classify windows using the same part-visualizations as containing a person or not (see Figure 1(h)). The set of windows are overlapping and randomly sampled. 10 subjects classified each window. A confidence score was computed as the average number of subjects classifying a window as containing a person. Standard non-maximal suppression can be performed by a machine using the confidence scores on all windows in an image. Finally, a precision-recall curve is computed to quantify the human subjects' performance. We note that similar part-detection visualizations can be created for both human and machine detected parts as shown in Figure 2[2]. This allows

[1]The root detections are not shown in Figure 1 for simplicity.

[2]To visualize the part detections of Felzenszwalb *et al*. [13] which contain highly overlapping detections, we perform non-maximal suppression among the parts. Each part is mapped to a color with intensity corresponding to the estimated likelihood of a person given the part score. Further evaluation indicates that our NMS processing of the parts actually increases the machine AP for INRIA by 0.018 and for PASCAL by 0.045.

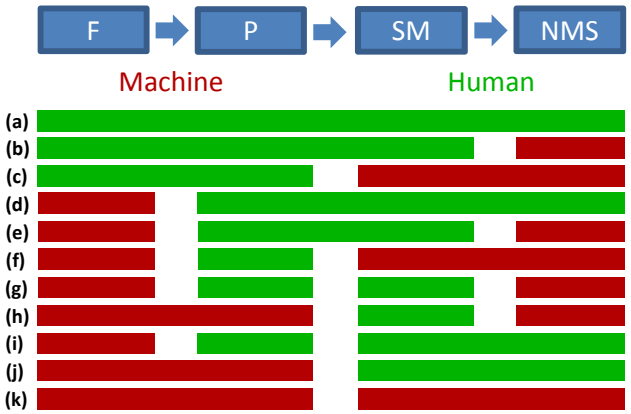


Figure 5: Summary of all experiments performed to test the use of humans and machines to perform various combinations of the components in a parts-based person detector.

us to evaluate the human spatial-model on machine or human detected parts. Inversely, the human detected parts can be fed into a machine spatial model.

One variation of the above experiments used for further analysis is to show subjects windows extracted from natural images instead of part visualizations (Figure 1 (b)). This is equivalent to using human extracted-features, parts as well as the human's spatial-model. We may also restrict the features available to the human subjects by pre-processing the images (Figure 1(d)), while still using human detected parts and spatial models.

4.4. Context and non-maximal suppression (NMS)

In the human studies on spatial modeling, we asked the subjects to classify cropped windows extracted from the images without context. We can study NMS and contextual reasoning by showing the subjects information over the entire image, and asking them to draw bounding boxes around detected persons (Figure 1(a,c,g)). By performing this task, subjects are implicitly performing NMS, and can use contextual information if provided. As shown in Figure 1, the information shown to the subjects is of three types; (a) original color images, (c) images after feature extraction, and (g) part-detections.

5. Results

In this section, we provide the results of numerous machine and human studies. We analyze the results with respect to the four detector components; feature extraction, part detection, spatial models and context/NMS. We also attempt to quantitatively compare the relative performance gains that may be achieved by improving each component of the detector. An illustration of the various combinations of human and machine experiments is shown in Figure 5.

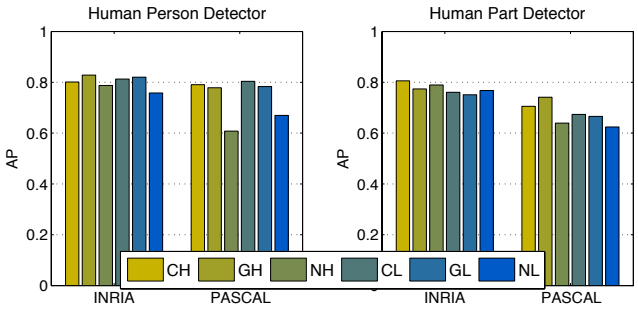


Figure 6: Effect of features: Accuracy of human person detector (left), and human part detector with machine spatial model and NMS (right) in high (H) and low (L) resolution color (C), grey-scale (G) and normalized-gradient (N) images.

5.1. Effect of features

We can analyze the effect of feature types using humans as person and part sliding window classifiers. We compare results using the original image (Figure 5 (b, c)) with results using different feature types (Figure 5 (e, f)). The results are summarized in Figure 6. We see that the loss of color or resolution does not significantly affect detection accuracies. Using normalized gradients alone did degrade human performance significantly, especially on the PASCAL dataset.

5.2. Effect of parts

In order to quantify the effect of better part detectors, we compare the performance of sliding window detectors on parts detected by humans and machines. We report results from showing grey scale image patches to the subjects, since the machine models do not use color. Similar results were found with other features types. There are several pairs of results we may consider as shown in Figure 5. These include using the machine spatial model with NMS (f, k), the human spatial model and machine NMS (g, h), and the human spatial model with human NMS (i, j). The results are shown in Figure 7. The use of human part detections significantly improves the performance of the detectors in most cases. For the challenging PASCAL dataset the improvement is as high as 0.24 for human detected parts over machine detected parts.

5.3. Effect of spatial models

We evaluate the effect of spatial models using a series of human studies on both machine and human detected parts. As shown in Figure 5, we compare machine to human spatial models using machine detected parts (h, k) and human detected parts (b, c) and (f, g). The results are shown in Figure 8. The results indicate that human spatial models do not significantly or consistently affect the AP scores across the various scenarios.

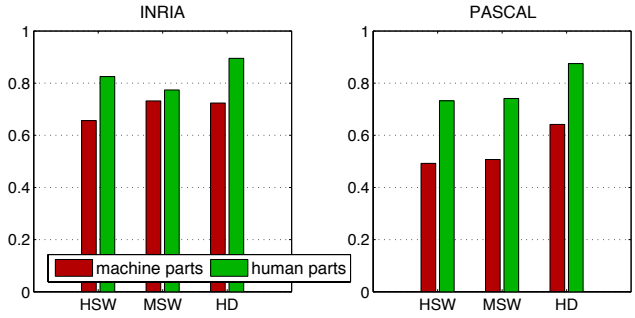


Figure 7: Effect of different part detectors: Accuracy of humans (HSW) and machines (MSW) spatial models followed by machine NMS, as well as using humans (HD) for both spatial models and NMS on part-visualizations. The human-parts are parts detected by subjects on grey-scale images.

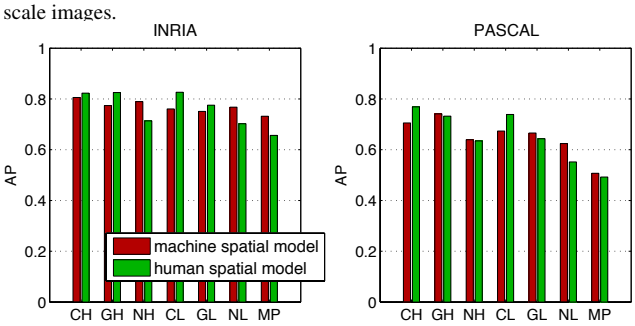


Figure 8: Effect of spatial models. Accuracy of human and machine spatial models on human part visualizations (CH, GH, NH, CL, GL, NL) and machine part visualizations (MP).

5.4. Effect of non-maximal suppression and context

Finally, we study the influence of context and non-maximal suppression (NMS) on detection performance. Two sets of experiments from Figure 5 can be compared. First, we can compare the human subjects as sliding window classifiers followed by machine NMS (e), to the human subjects as detectors using the entire image (d). The results are shown in Figure 9 (left) for grey-scale images. The use of the subjects' contextual models and NMS significantly increases performance by up to 0.15 on the PASCAL dataset over using the machine NMS.

Second, we can compare results using the part visualizations with human spatial models and machine NMS (g, h) to human spatial models and human NMS (i, j). The results are shown in Figure 9 (right). Since the subjects only have access to part detections, they are limited to performing intra-category contextual reasoning or NMS. As demonstrated in Parikh *et al*. [30], the use of inter-category contextual information may not be necessary given high resolution appearance information. A similar finding is found by comparing the experiments in Figure 9 that all use high resolution information. The amount of improvement in the left plot using all contextual information is similar to the right two plots that only use intra-category contextual information.

Figure 10 compares the accuracy of human subjects as person detectors (a, d) to humans as sliding window classifiers followed by machine NMS (b, e) across feature types. The accuracy of human subjects at classifying image windows in isolation decreases significantly as the appearance information becomes weak (*e.g*. normalized gradient low resolution images). However, their performance at detecting people in entire images remains quite robust. As previously shown in [30], this signifies the importance of contextual information under impoverished scenarios.

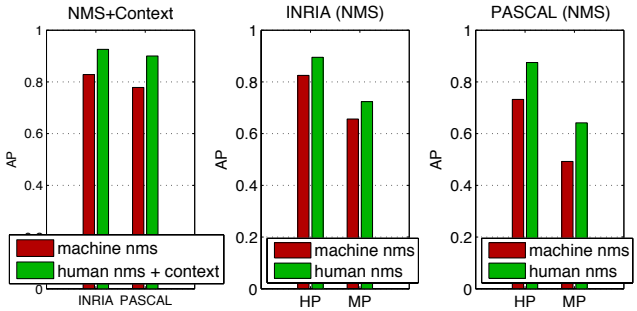


Figure 9: Effect of NMS + Context. Accuracy of human detectors using (left) high resolution grey-scale images and using (two right plots) human-part (HP) and machine-part (MP) visualizations.

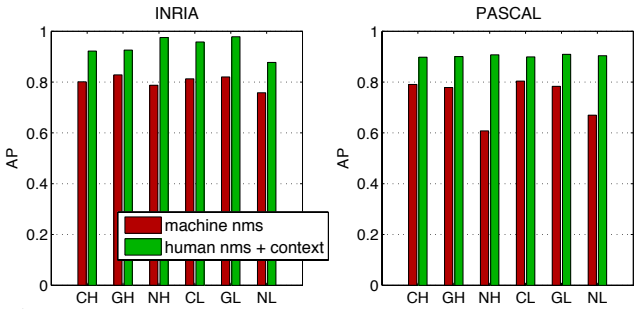
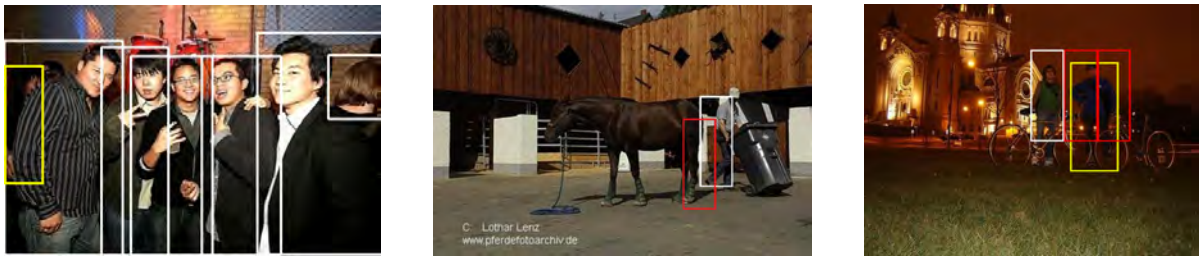


Figure 10: Effect of Context. Humans can reliably leverage contextual information and maintain robust detection accuracy even with impoverished appearance information.

5.5. Summary

We summarize the potential improvements in object detection accuracies based on our human and machine studies. In Figure 12, we show the average improvement in accuracies for part detection, spatial modeling and context/NMS and their variances. In reference to Figure 5, results are computed from (f, k), (g, h) and (i, j) for part detections; (b, c), (f, g) and (h, k) for spatial models; and (g, i) and (h, j) for context/NMS. We find parts to have the biggest impact, followed by non-maximal suppression. Spatial models do not have significant impact. The people in the PASCAL images as compared to those in the INRIA dataset demonstrate a wider variation in poses, and often exhibit challenging scenarios such as occlusion and truncation. Thus, we observe a higher potential for improvement on the PASCAL dataset, as seen in Figure 12. We note that the estimates of potential improvements can be viewed as lower-bounds, since our human studies were performed on Amazon Mechanical Turk where subjects may be distracted and often provide noisy



(a) Human detector (b) Human sliding-window (c) Machine spatial model

Figure 11: Example failure cases for scenarios with different amounts of human involvement. Correct detections are shown in white, false positives in red and false negatives in yellow. (a) Even when subjects are shown the entire image, highly occluded people in bad lighting are missed. (b) When subjects classify windows in isolation from the rest of the image as containing a person or not, lack of context leads to false positives when the windows locally appear to have parts of a person. (c) A machine spatial model applied to near-perfect human part-detections fails because of symmetric part detections. Subjects were asked to classify patches as containing arms, legs, etc. and were not asked to distinguish between left/right arms, legs, etc.

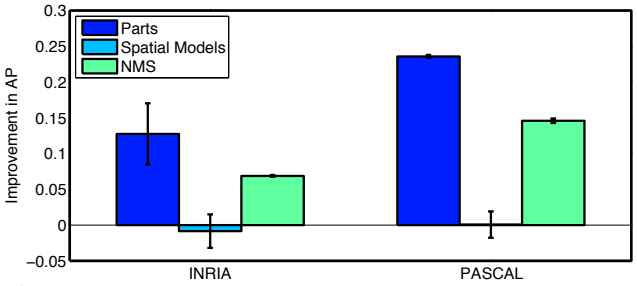
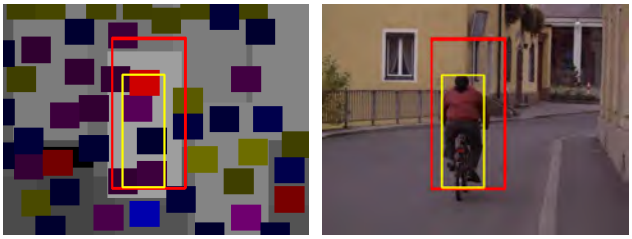


Figure 12: Summary of our results: The component of a parts-based person detector that can improve detection performance the most is the part-detection, followed by the NMS component. Spatial models do not affect the resultant performance significantly.



(a) Machine part detections (b) Resultant detections

(c) Human part detections (d) Resultant detections

Figure 13: Example detections where human detected parts allow for successful detection (white), while machine detected parts lead to false positive (red) and false negative (yellow) detections.

responses. Example failure cases for both human subjects and machines are shown in Figures 11 and 13.

6. Discussion

Our analysis is restricted to sliding-window parts-based models. It assumes a pipeline where the parts and spatial models are considered to be independent. Part models could be learnt jointly with spatial models as in [1, 13]. Moreover, the weaker the part models, the bigger role spatial models could play in the final detection performance. Our analysis does not account for such dependencies among various components in the pipeline.

In our human studies, subjects were instructed to find parts with semantic meaning (heads, torso, etc.). Since the patches were presented in isolation, we do not expect this semantic knowledge to provide contextual information to subjects. However, machine object detectors have the freedom to model parts without semantic meanings. This flexibility may allow for the use of better parts, but could also make the underlying learning problem intractable.

The accuracies of human subjects as person detectors on color and grey-scale images is higher than any experiment using a combination of machine and human components. This implies that the pipeline proposed for the machine detector may not be the same as the human subjects'.

In conclusion, we presented numerous studies combining both machine and human components for detecting people. By analyzing their relative performance we can determine which components could offer the greatest boost in overall performance if improved. Our results show that part detection is the weakest link on challenging datasets such as PASCAL, followed by non-maximal suppression and context. Human spatial models appear to offer negligible performance increase over machine spatial models. Grey-scale information provided the same level of accuracy as color. However, accuracies suffered when using only normalized gradients. Future work involves similar analysis for detecting generic object categories and other object detection models.

7. Acknowledgments

We thank Congcong Li for providing the machine part detections, and Ross Girshick for useful discussions on the machine spatial model.

References

[1] Y. Amit and A. Trouv. Pop: Patchwork of parts models for object recognition. *Int'l J. of Computer Vision*, 75:267–282, 2007.

[2] 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.

[3] E. J. Bernstein and Y. Amit. Part-based statistical models for object classification and detection. *IEEE Proc. of CVPR*, 2005.

[4] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. *IEEE Proc. of CVPR*, 2005.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Proc. of CVPR*, 2005.

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *IEEE Proc. of ICCV*, 2009.

[7] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009.

[8] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *IEEE Proc. of CVPR*, 2007.

[9] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *IEEE Proc. of CVPR*, 2009.

[10] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) http://www.pascalnetwork.org/challenges/voc/voc2007/.

[11] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), 2007.

[12] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 3. http://people.cs.uchicago.edu/ pff/latent-release3/.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on PAMI*, 32:1627–1645, 2010.

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int'l J. of Computer Vision*, 61:55–79, 2005.

[15] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *IEEE Proc. of CVPR*, 2, 2003.

[16] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *IEEE Proc. of CVPR*, 2005.

[17] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, C-22(1):67 – 92, 1973.

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 2006.

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

[20] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *Int'l J. of Computer Vision*, 80, 2008.

[21] Z. Liu and D. Kersten. 2d observers for human 3d object recognition? *Vision Research*, 38(15-16), 1998.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Computer Vision*, 60, 2004.

[23] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Proc. of CVPR*, 2008.

[24] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A*, 7(5), 1990.

[25] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., 1982.

[26] J. Mutch and D. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *Int'l J. of Computer Vision*, 80, 2008.

[27] A. Oliva and P. G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41(2), 2000.

[28] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. *IEEE Proc. of ICCV*, 1998.

[29] D. Parikh and C. Zitnick. The role of features, algorithms and data in visual recognition. In *IEEE Proc. of CVPR*, 2010.

[30] D. Parikh, C. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *IEEE Proc. of CVPR*, 2008.

[31] W. H. Plantinga and C. R. Dyer. An algorithm for constructing the aspect graph. In *Foundations of Computer Science, 1986., 27th Annual Symposium on*, 1986.

[32] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Trans. on PAMI*, 22(9), 2000.

[33] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE Proc. of ICCV*, 2007.

[34] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *IEEE Proc. of CVPR*, 2007.

[35] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *IEEE International Conference on Computer Vision.*, 2007.

[36] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. *IEEE Proc. of CVPR*, 2000.

[37] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int'l J. of Computer Vision*, 81, 2009.

[38] M. J. Tarr and S. Pinker. When Does Human Object Recognition Use a Viewer-Centered Reference Frame? *Psychological Science*, 1(4), 1990.

[39] A. Torralba. Contextual priming for object detection. *Int'l J. of Computer Vision*, 53, 2003.

[40] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. on PAMI*, 30, 2008.

[41] Z. Tu. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In *IEEE Proc. of ICCV*, 2005.

[42] P. Viola and M. J. Jones. Robust real-time face detection. *Int'l J. of Computer Vision*, 57, 2004.

[43] M. Weber, M. Welling, and P. Perona. In *IEEE Proc. of CVPR*, volume 2, 2000.

[44] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. In *Pattern Recognition*, volume 5096. 2008.