Pascal VOC 2008 Challenge

Derek Hoiem
University of Illinois Urbana-Champaign.
dhoiem@cs.uiuc.edu

Santosh K. Divvala, James H. Hays
Carnegie Mellon University.
{santosh, jhhays}@cs.cmu.edu

1. Approach Overview

To tackle the challenging dataset presented in this challenge, we use the highly successful appearance-based detector of Felzenszwalb *et al.* [1] and augment it with rich contextual cues extracted from the image to further improve its performance. Specifically, we train detectors to obtain the confidence that a window contains an object based solely on global scene statistics [2, 3], nearby regions, the object position and size, geographic context [4] and boundaries [5, 6]. Our interest is to study how much each of these contextual cues can add to the performance of the local appearance based detector.

This report provides specific details of each of the individual cues used to tackle the classification, detection and segmentation competitions (more or less in a similar manner).

1.1. Local Appearance

To detect and localize the presence of objects of a generic category based on local appearance-based cues, we employ the method proposed by Felzenszwalb *et al.* in [1]. This detector has been very successful and had achieved top performance in most categories in the PASCAL VOC 2007 challenge. Qualitatively, we have observed that the results achieved by the detector are quite a bit better than could be interpreted from the reported numbers. This is because although, the detector does a good job in detecting the presence of an object correctly, it makes some mistakes in localizing it, due to the fixed aspect ratio of the bounding box and multiple firings on the same object. Thus, some false positives are due to mistakes in the appearance model (e.g., mistaking a lamppost for a person) but others are due to poor localization. We attempt to overcome these problems by augmenting the detector with global contextual information and improving localization using segmentation.

1.2. Global Context

The presence of an object at a particular location is believed to be influenced by its surroundings. We explore this hypothesis by developing detectors that predict whether an image contains the object and its likely location and size purely based on global contextual cues.

Object Presence To predict the likelihood of observing an object given the scene context, we train classifiers using contextual cues such as whole image gist [2], geometric context confidence maps (12×12 re-sized maps) [3] and "geographic context" (derived from the work of im2gps [4]). The former two cues have been shown in the literature to be good sources of contextual information.

The use of *geographic context* for object detection is a novel contribution in this work. The intuition is to provide geographic information to an object detector for each scene which will enhance or suppress object detections according to the co-occurrence of geographic properties and objects (e.g. 'boat' is frequently found in water, 'pedestrian' is more likely in high population density). Geographic properties such as land cover probabilities (e.g. 'forest', 'water', 'barren', or 'savanna'), population density estimates, light pollution estimates, and elevation gradient magnitude estimates are used. All the geographic properties are estimated as described in [4]. For each query image, any exact-duplicate Flickr images as well as any images from the same photographer are removed from consideration. The geographic properties are used to compute the likelihood that a scene contains an object of a certain class given the value of its geographic properties for each object class independently using logistic regression.

We also use the keywords associated with each image in the im2gps [4] dataset of Flickr images to predict object occurrence. The 500 most popular words appearing in Flickr tags and titles were manually divided into categories corresponding to all 20 VOC classes and 30 additional semantic categories. For instance, 'bottle', 'beer', and 'wine' all fall into one category, while 'church', 'cathedral', and 'temple' fall into another category. We use logistic regression to predict object class based on a count of the number of keywords falling into each of these categories in 80 nearest neighbor scenes.

Object Location The goal here is to predict **where** the

object(s) are likely to appear in an image (given that there is an indication of at least one object occurring in the image by the previous classifier). To train this location predictor, we divide the image into 5×5 grids and then train separate classifier for each grid using the whole image gist and geometric context cues. A grid is labeled positive if the bottom mid-point $(\frac{x_{left}+x_{right}}{2}, y_{bottom})$ of a bounding box falls within it.

Object Size The idea here is to predict the size (log pixel height) of an object, given its location in the image. This is learnt again using contextual cues based on depth from occlusion [6] (i.e., value at the bottom mid-point of an object bounding box), viewpoint estimates (relative y-value), whole image gist and geometric context. The true sizes are calculated using the ground-truth annotations provided for the objects in the training data. This regression task is reformulated as a series of classification tasks, where we first cluster object sizes into five clusters s_1, s_2, s_3, s_4, s_5 and then train a separate classifier for each size (i.e., size $< s_2$, size $< s_3$, size $< s_4$, size $< s_5$). At testing, we calculate $P(size = k)$ as $P(size < k+1) * (1 - P(size < k))$, with $\sum_k P(size = k) = 1$ and then compute the expected size as $\sum_k P(size = k) * center(k)$.

1.3. Object Segmentation

Localization error can cause multiple overlapping detections on a single object, or can cause an object to be missed entirely (in computing quantitative results) because the detector bounding boxes do not overlap sufficiently with the ground truth bounding box (due to aspect ratio differences). To remedy this, we apply graph cuts [7] segmentation to each bounding box above a threshold after performing non-maximum suppression. The segmentation can also be used to improve the appearance model with region-based features.

The unary potentials are based on class models of color, textons [8], geometric context [9], and a probability of background region detector trained on LabelMe. The unary potentials are learned by taking the log likelihood ratios of histograms on the training ground truth segmentations and learning a weighting of them using both the training and validation segmentations (only VOC2008 images were used). A shape prior was also learned over the training set using all candidate detections with at least 50% overlap. The pairwise potentials are based on probability of boundary [5] and probability of occlusion boundary [6] soft confidence maps. The pairwise parameters were set manually to be the same for each class (potential of -log(P(boundary))), except that occlusion boundaries were not used for chairs and bicycles. Given a bounding box, the image is resized so that the object length is 100 pixels, and graph cuts inference is performed to get the object mask.

For each object, we also train an appearance model based on histograms (normalized counts and entropies) of color, texture, discretized HOG features, and the segmentation quality. Given an object mask and its energy, we quantify the segmentation quality as the difference in energy from a purely background solution normalized by the number of object pixels.

After segmentation, the object bounding box is adjusted to the bounding box of the object mask, non-maximum suppression is performed based on region overlap ($> 50\%$ intersection over union), and the object score is updated as a weighted combination of its windowed detection score (including contextual information) and the segmentation-based score, with the weights learned on the validation set. Since the segmentation consistently undersegments or oversegments some objects (e.g., missing the legs of a chair), the bounding box is adjusted along each coordinate by the mean difference (with respect to object width or height), according to correct detections in the validation set.

2. Competitions

The task of recognizing objects in realistic scenes essentially requires the coordination of all of the above individual cues. In this submission, we have used a unified framework to integrate information obtained from each cue into the other.

Training and Datasets For extracting the geometric context and occlusion boundary information, we used the code and classifiers that are publicly available online (http://www.cs.uiuc.edu/homes/dhoiem/projects/software.html) as is. The geographic context, trained on the PASCAL VOC 2008 training set, uses the scene matches from Flickr but removing images that overlap with the VOC 2008 testset. The appearance-based detector provided by the authors [1] was trained on the PASCAL VOC 2007 trainval set.

2.1. Detection Competition

For detection, we combine the predictions from the *object presence*, *location*, *size*, *local detector* and *segmentation* classifiers. The location classifier was trained using VOC 2008 train-val and VOC 2007 test sets. The rest of the classifiers were trained using only the VOC 2008 train-val set. Logistic regression was used for training all of the context classifiers and feature weighting. A linear SVM classifier was used for training the segmentation-based appearance models. Table. 2.1 displays the detection results obtained on the validation set with and without using context information and after performing the segmentation. The results may be biased, since we used the validation set to tune

some parameters and feature weightings.

2.2. Classification Competition

For this competition, we combined the predictions from the *object presence* classifier and the above detector to predict the presence/absence of an object in the image. We also trained another classifier based on HOG [10] and SIFT [11] features in a typical Bag-of-Features paradigm to augment the above two scores. The final classification scores were obtained by linearly combining the individual classifier scores. For all the classifiers, logistic regression with L1-regularization [12] was used for training.

2.3. Segmentation Competition

We segment the objects as described in Section 1.3, with the difference that alpha expansion is used to make the objects compete for pixels.

Acknowledgments We thank Pedro Felzenszwalb and Deva Ramanan for kindly allowing us to use their detector.

References

[1] Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. Computer Vision and Pattern Recognition (CVPR) (2008) 1, 2

[2] Torralba, A., Oliva, A.: Statistics of natural image categories. Network: computation in neural systems **14** (2003) 1

[3] Hoiem, D., Efros, A., Hebert., M.: Recovering surface layout from an image. International Journal of Computer Vision **75** (2007) 1

[4] Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. Computer Vision and Pattern Recognition (CVPR) (2008) 1

[5] Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: Proc. CVPR. (2008) 1, 2

[6] Hoiem, D., Efros, A., Hebert., M.: Recovering occlusion boundaries from a single image. International Conference on Computer Vision (2007) 1, 2

[7] Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001) 1222–1239 2

[8] Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62** (2005) 61–81 2

[9] Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: Proc. ICCV. (2005) 2

[10] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR. (2005) 3

[11] Lowe, D.: Object recognition from local scale-invariant features. (1999) 1150–1157 3

[12] Koh, K., Kim, S.J., Boyd, S.: An interior-point method for large-scale l1-regularized logistic regression. In: Journal of Machine Learning Research. (2007) 1519–1555 3

Table 1. Detection Accuracies: From left to right: pedro/deva baseline, +context, +segmentation, +bboxadjustment, +segmentation-based appearance

	pd	pd-combined	segloc	seglocbbfit	comp4
Aeroplane	0.184	0.219	0.328	0.336	0.361
Bicycle	0.322	0.321	0.338	0.332	0.326
Bird	0.093	0.1	0.104	0.105	0.123
Boat	0.093	0.093	0.078	0.079	0.084
Bottle	0.239	0.252	0.254	0.253	0.247
Bus	0.206	0.203	0.253	0.255	0.262
Car	0.252	0.247	0.265	0.267	0.271
Cat	0.05	0.183	0.189	0.194	0.201
Chair	0.132	0.141	0.106	0.102	0.121
Cow	0.144	0.166	0.165	0.173	0.182
Dining-table	0.062	0.124	0.13	0.13	0.135
Dog	0.034	0.087	0.108	0.127	0.157
Horse	0.29	0.298	0.279	0.286	0.293
Motorbike	0.276	0.314	0.288	0.29	0.309
Person	0.301	0.351	0.36	0.372	0.384
Potted-plant	0.156	0.148	0.147	0.15	0.149
Sheep	0.11	0.118	0.105	0.112	0.061
Sofa	0.156	0.176	0.174	0.174	0.184
Train	0.182	0.192	0.219	0.219	0.262
Tvmonitor	0.329	0.368	0.38	0.379	0.415
Average	0.181	0.205	0.213	0.217	0.226