Density-Aware Part-Based Object Detection with Positive Examples

Ekaterina Riabchenko
Machine Vision and Pattern Recognition Laboratory
Lappeenranta University of Technology

Joni-Kristian Kämäräinen and Ke Chen
Deparment of Signal Processing
Tampere University of Technology

***Abstract*—Part-based models have become the mainstream approach for visual object classification and detection. The key tools adopted by the most methods are interest point detectors and descriptors, shared codes for object parts (visual codebook) and discriminative learning using positive and negative class examples. Distinction of our method from the existing part-based methods for object detection is the use of sparse class-specific landmarks with semantic meaning. The landmarks are the additional distinguished information of object location in the proposed framework. Additionally, localising semantic and discriminative landmarks (object parts) is significant in other related applications of computer vision, such as facial expression recognition and pose/orientation estimation of objects. Therefore, we propose a model which deviates from the mainstream by the fact that the object parts' appearance and spatial variation, constellation, are explicitly modelled in a generative probabilistic manner. With using only positive examples our method can achieve object detection accuracy comparable to state-of-the-art discriminative method.**

I. Introduction

Discriminative learning, utilising intensive parameter and model optimisation, has turned out to be the approach for visual object class detection and classification with popular benchmarks (Caltech-101 [1], Caltech-256 [2], Pascal VOC [3]) in terms of precision-recall. Many of the state-of-the-art methods ([4], [5], [6], [7]) require only training images with manually annotated bounding boxes. As a downside, the search space of parameters and model variables can be large: selection of the best class parts, selection and parametrisation of visual features, estimation of locations of the parts in training images, and finally, optimisation of the discriminant function parameters. Assumption of existing discriminative part-based models, that sufficient amount of training samples is available to learn the boundary of the function, is invalid in practice. In other words, data sparsity problem [8], [9] and scalability to a large-scale number of object categories [10] will significantly increase the difficulties to learn a discriminative model.

To overcome the limitation of the existing discriminative methods (e.g. extensive optimisation), generative learning is adopted in this paper. In this work, we propose a simple training pipeline from the low level features, biologically inspired complex-valued Gabor filter responses, to full part-based object detection. We form part appearance models and probabilistic spatial constellation models in a fully probabilistic manner. Our method has only one learning phase, estimation of probability density functions (pdfs) by unsupervised Gaussian mixture model. As a result, training time is linearly dependent on the number of object classes and

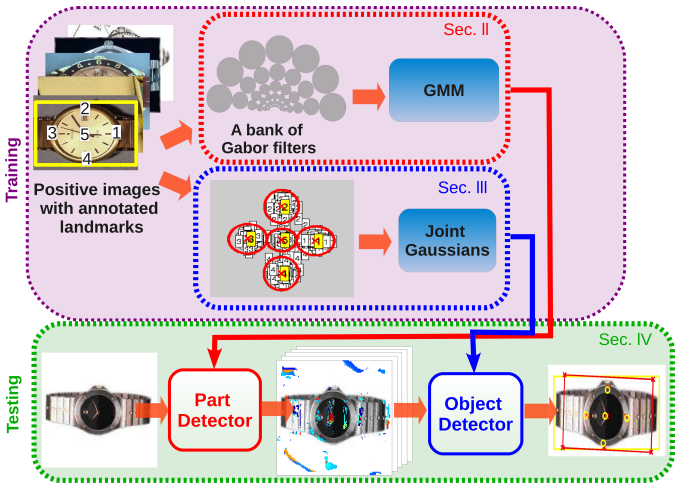

Figure 1. The workflow of our method.

examples. Figure 1 shows a flow chart of our method. In experiments, our method, learning from positive examples only, achieves performance comparable to the-state-of-art discriminative method [4], proving its better efficiency in training. Our main contributions are:

- A probabilistic part-based object model consisting of 1) probability density driven part model and detector, and 2) probability density based constellation model in "aligned object space".
- Explicit definition of the probability terms in the model and their unsupervised estimation from training data.
- A supervised generative learning approach which only needs positive examples and performs well with just tens of training examples of complex visual classes.

A. Related Work

Privileged learning – Aside from the mainstream, our approach does not utilise bounding boxes in training but class specific landmarks, object parts (Figure 2). Apparently, those manually or automatically selected and annotated landmarks are the distinguishable parts of the objects with semantic meaning. For example, the facial landmarks in Figure 2 can be further extended to facial expression recognition [11]. Given these strong label information, i.e. *learning using privileged information* [12], our framework can get benefits from additional source of useful information related to object detection task and significantly boost the training step. Given example images with landmarks, our method is practically parameter-free and

learns probabilistic models for the both class landmarks and their spatial variation, constellation of the class parts.

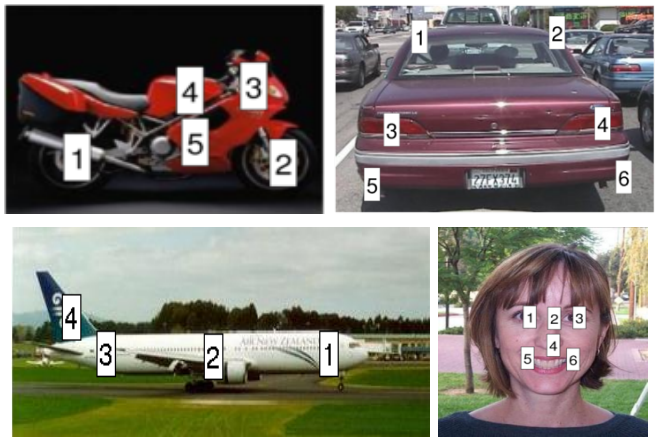

Figure 2. Examples of annotated object parts (landmarks).

Supervised part-based models – Our approach is aside from the mainstream where the methods are typically interest point driven. For example, the Bag-of-Features (BoF) approach with no/ad-hoc spatial model (e.g., [13], [14], [15]), methods using a spatial model over shared "parts" (BoF codebook) (e.g., [16], [17], [18]), and finally methods which learn explicit models for the local parts, but utilise interest point detectors in the part selection [5], [19], [20]. Supervised methods which do not utilise interest points or the BoF structure but learn explicit models for the parts and their spatial variation are only a few (Table I). Our choice of the local part model, multi-resolution Gabor bank and Gaussian mixtures, is similar to Felzenszwalb and Huttenlockher [21] who used steerable filters and a single diagonal Gaussian. Our spatial model is similar to Burl et al. [22] who represented variation with Gaussians in a normalised object space; we also use Gaussians, but do not use fixed landmarks to transform points to the object space.

II. A Probabilistic Model of Local Parts

A probabilistic local part model requires two processing stages: 1) extraction of local image features $\boldsymbol{f}(x, y)$ and 2) their representation in a probabilistic form $p(\boldsymbol{f}, P_j)$ (P_j denotes a specific part). In the bag-of-features approach the first stage is typically sub-divided to the consequent stages of detection and description (e.g., [13], [14]), but that introduces the problems of finding a good object part detector and selecting a good detector-descriptor combination [31]. In many recent works, the interest points have been replaced with dense sampling [32], and accordingly we want to fetch the most likely local part candidates over the whole image. Our method of choice is the supervised method from [33] (Figure 3), which uses a multi-resolution Gabor filter bank to extract features and converts features to probability density values using an unsupervised Gaussian mixture model pdf estimation methods by Figuiredo and Jain [34].

Our detector learns features in aligned object spaces (see Section III) using special procedure to handle limited data (see Section II-A) which help part representation and detector to be more discriminative and invariant to rotation and scaling. The method is very effective and for an input image provide a full

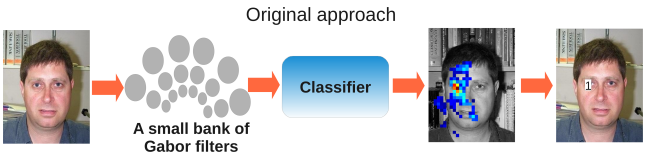

Figure 3. Workflow of the probabilistic part detector used in our work [33]

map of probability likelihoods which can be used to pick N best part candidates (Figure 4). This is the reason why we call our system "density-aware".

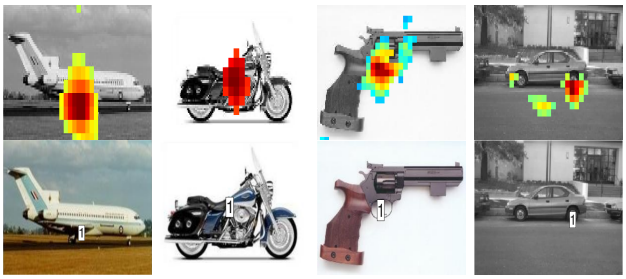

Figure 4. Examples of the part likelihoods (only the 10% best fractile shown for clarity) and the best candidate selected by the largest likelihood value.

A. Learning Parts from a Few Examples

Difficulties occur with the unsupervised estimation of GMM parameters $\boldsymbol{\theta}$ with the Figuiredo-Jain method [34] if there are too few training examples. In our case, with 4 frequencies and 6 orientations in the Gabor space, the estimation typically failed with less than 200 examples. However, we would like to keep the size of training set as small as possible to avoid laborious annotation and therefore need to circumvent the limitation. There are two standard solutions in the case of the lack of training data: *feature randomisation* [35] or *data randomisation (bootstrapping)* [36].

By data randomisation, also referred as *enrichment*, we understand replication of the data by adding Gaussian noise to the landmark locations. Thus we repeat each training image with slightly shifted landmark position, to make the EM part in [34] to avoid singularities. However, this approach alone cannot avoid overfitting.

The second approach, feature randomisation [33], can be implemented during training. Instead of a single GMM-based pdf which deals with all features ($|\boldsymbol{g}| = 4 \times 6 = 24$) we can estimate, for example, 5 GMMs with 9 randomly selected features. Each $j = 1 \ldots 5$ of the pdfs define own probability density function: $p_j(\boldsymbol{g}|\boldsymbol{\theta})$. By the assumption of independence, the final pdf is formed as the product: $p(\boldsymbol{g}|\boldsymbol{\theta}) = \prod_j p_j(\boldsymbol{g}|\boldsymbol{\theta})$. It is noteworthy that this novel approach is similar to the idea of forming random forests for classification [37].

III. A Probabilistic Model of Part Constellation

To remove the effects of translation, scaling and rotation, the local parts are mapped to a normalised space referred to as "object space". In the object space, spatial variation of the parts is characteristic for a class and can be statistically captured by density estimation. Burl et al. [22] fixed two landmarks

Table I. "STRONGLY" SUPERVISED PART-BASED OBJECT CLASS MODELS (D: DISCRIMINATIVE, G: GENERATIVE).

Ref-Year	Feature	Const. model	Type	Test data
[23]-1995	Gaussian derivative	Spatial voting	G	A few simple objects
[22]-1998	Sliding window	P parts in canonical space	G	Own face images
[24]-2005	Edge features	K-fan representation	G	Caltech-4
[21]-2005	Steerable filters	Pair-wise energy model	G	Yale face + own torso images
[25]-2009	General + part detector	Prob. model on a frame	G	Torso ("Buffy"), VOC2008
[26]-2009	Boosted boundary feats.	Boundary model	G	Few "googled"
[27]-2009	Color and HOG	Tree structure	D	Buffy & sign lang.
[28]-2009	HOG	No model (vote sum)	D	Human detection
[29]-2010	Sliding window	Graphical model	D+G	Caltech-4 face + torso
[30]-2010	Edges (Gabor)	Edge map	G	Few own classes

to transform objects into the object space (isometry). In that approach, detection of the fixed landmarks becomes a bottleneck and their spatial variance is undesirably re-distributed to others. In this work, we adopt a variant of the mean shape algorithm by Cootes et al. [38]. We iteratively transform part constellations to a mean shape space which is simultaneously updated. A pseudo code is given in Algorithm 1 and examples in Figure 5.

Algorithm 1 Construct an aligned object space.

1: Select a random image and use its part locations as the initial object space.
2: **for all** images **do**
3: Estimate the point correspondence homography $\boldsymbol{H}$ to the object space using [39] (using 2 points for similarity or isometry, 3 points for affinity homography transformation). *// Store also to the set $\boldsymbol{H}_{prior}$ used in Algorithm 2*
4: Transform parts to the object space.
5: Refine the space by computing the average part locations.
6: **end for**
7: Return the final object space part locations.

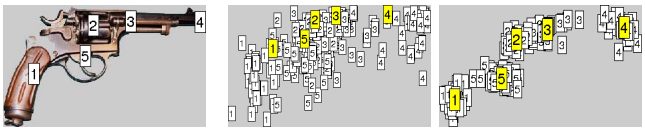

Figure 5. Left-to-right: annotated object parts; all training examples plotted in the original space; plotted in the object space (see Algorithm 1). Yellow labels denote the mean locations.

For rigid objects, a single 2D Gaussian at each part is sufficient to represent their spatial variation (see Figure 5). The joint probability density is $p(\boldsymbol{x}'_1, \boldsymbol{x}'_2 \dots \boldsymbol{x}'_N|\boldsymbol{\theta}_c)$, where $\boldsymbol{x}'_i$ are part locations in the aligned space and $\boldsymbol{\theta}_c$ is the set of parameters (mean and variance) of N 2D Gaussians of the class c. Note that the number of parts N is class specific. Assuming independence between the parts' locations, the constellation pdf gets the form (the object space notation $'$ and indexing with c are omitted for clarity):

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_N|\theta_c) = p(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \Sigma_1) \dots p(\boldsymbol{x}_N|\boldsymbol{\mu}_N, \Sigma_N) \quad (1)$$

In our generative model, the constellation probability in (1) forms the model prior. For efficient search, we also estimate a set of allowable transformations $\{\boldsymbol{H}_{\text{prior}}\}$ which can be constructed from the set of transformations $\boldsymbol{H}$ in Algorithm 1.

IV. PROBABILISTIC OBJECT DETECTION

Our local parts are learned from only positive examples and described by probability densities (Section II, Figure 4) - these form the observation probability. The spatial constellation of the parts is described by joint coordinate probability density learned as well from positive examples (Section III) - this forms prior probability. Our desired output is the set of most likely hypotheses of part locations $\{hyp_{BEST}\}$ - the nominator (likelihood) in the Bayes' theorem. We may compute likelihoods over the whole parameter space (coordinates of each landmark), but that would be inefficient. Therefore we propose an efficient and effective random sampling based algorithm (Algorithm 2). The inputs for Algorithm 2 are the N_{bestLM}

Algorithm 2 Detect object

1: Initialise the set of best hypotheses $\{hyp_{BEST}\}$ to null and set score values to zero
2: **for** all minimum combinations of the detected parts (2 for isometry/similarity) **do**
3: Estimate the homography $\boldsymbol{H}$ from the image space to the object space
4: **if** $\boldsymbol{H} \notin \{\boldsymbol{H}_{\text{prior}}\}$ **then**
5: Skip this hypothesis.
6: **end if**
7: Transform all detected parts to the object space using $\boldsymbol{H}$
8: For each transformed part compute the spatial likelihood (the single terms in (1)).
9: Select the parts with the highest likelihoods (omit if below $P_{landmark}$).
10: *Compute the detection probability* p (see Section IV-A)
11: **if** p is better than for any in $\{H_{best}\}$ **then**
12: Using all selected parts, estimate $\boldsymbol{H}^{-1}$ from the object space to the image space
13: Transform the selected parts to the image space (replace omitted with the mean shape).
14: Add hypothesis (parts' coordinates) to $\{hyp_{BEST}\}$ (remove the worst if the max number exceeded).
15: **end if**
16: **end for**
17: Return the best hypotheses in $\{hyp_{BEST}\}$

extracted local parts for each class (see Figure 4), *object space* (Algorithm 1) and $\boldsymbol{H}_{\text{prior}}$, estimated from training examples. With a small number of the best detected parts, for example only the five best, the algorithm requires about a hundred iterations to provide very accurate hypotheses. Moreover, for the most of iterations only the operations up-to the line 3 will be executed due to transformation prior omitting too awkward transformations. The only remaining question is the form of the detection probability computed in the line 10 in Algorithm 2. Note that for speedup we do not include the term $\boldsymbol{H}_{\text{prior}}$ into the detection probability but fix the extreme rotation and scaling values to $\pm 20\%$. All hypotheses outside the limits will be omitted.

A. Detection Probability p in Algorithm 2

The detection probability consists of the two elements: probability of spatial locations of the parts in the object space and likelihoods of the parts' appearance. At first glance, we are attempted to try the Bayesian rule of posteriors similar to [22], but that has two problems. First of all, we would need a good estimate of the *X_not_at_that_location* density, i.e. piles of negative examples. Secondly, object detection is not a classification problem, but should be based on likelihood values (nominator in the Bayes' theorem). Detection precedes the classification stage, which is a Bayesian task. Detection must be based on the highest likelihoods found for each class.

The full likelihood consists of the two elements mentioned above and, assuming independence of the appearance and spatial location, can be written as (C denotes class number and i its part number)

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{g}|\theta_{c,i}, \theta_c) &= p(\boldsymbol{x}|\theta_c) \times p(\boldsymbol{g}, \theta_{c,i}) = \\ &= \underbrace{p(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \Sigma_1) \cdot \ldots \cdot p(\boldsymbol{x}_N|\boldsymbol{\mu}_N, \Sigma_N)}_{\text{constellation}} \underbrace{p(\boldsymbol{g}|\theta_{c,i})}_{\text{appearance}}. \end{aligned} \quad (2)$$

For appearance, we may assume conditional independence between the different parts, but that does not help the problems due, for example, occlusion or simply detection failure. Therefore, we must consider all possible combinations of 1, 2, ..., N-1, N parts visible out of N possible. That leads to 1st, 2nd, ..., Nth order probability terms:

$$\begin{cases} p(\boldsymbol{g}_1|\theta_{c,1}) + p(\boldsymbol{g}_2|\theta_{c,2}) + \ldots + p(\boldsymbol{g}_N|\theta_{c,N}) + \ldots \\ p(\boldsymbol{g}_1|\theta_{c,1})p(\boldsymbol{g}_2|\theta_{c,2}) + p(\boldsymbol{g}_1|\theta_{c,1})p(\boldsymbol{g}_3|\theta_{c,3}) + \\ p(\boldsymbol{g}_2|\theta_{c,2})p(\boldsymbol{g}_3|\theta_{c,3}) + \ldots \\ p(\boldsymbol{g}_1|\theta_{c,1})p(\boldsymbol{g}_2|\theta_{c,2})p(\boldsymbol{g}_3|\theta_{c,3}) + \ldots \\ \ldots \end{cases} \quad (3)$$

which enforces us to perform summation of $\sum_{k=1\ldots N} \binom{N}{k}$ product terms. Good iterative implementations perform well upto 10-15 parts after which approximations (e.g., Stirling's approximation formula based) must be utilised. In our experiments the number of parts was sufficiently small for exact computation.

In Equation 2, a strong product constraint is used for spatial locations of the parts, but a rather weak sum constraint for their appearance. That is understandable as the part detection is always unreliable due to detection failures (background clutter) and occlusion. The appearance part in Equation 2 consists of low, mid and high order probability terms exemplified in Equation 3, but cannot be reduced since the high order terms dominate if almost all local parts are detected correctly and the low order terms if only a small number of the parts are detected. It should be noted, that Algorithm 2 recovers from occlusion (the lines 11 and 15) since the parts with too low constellation likelihoods, which are likely to be false alarms, are replaced with the mean shape parts and mapped back to the query image.

V. Experiments and Results

We report classification and detection results for the two popular datasets, Caltech-4 and Caltech-101. The obtained results are reported using confusion matrices and precision-recall-curves. Since the training data is limited with Caltech-101, we present our results using feature randomisation methods with GMM-based pdf estimation (Section II-A). For the both datasets the provided images were randomly divided into equal size training and test sets.

A. Performance Measures

Precision-recall curves and confusion matrices, in our work, were used in order to compare our results with other works. We also utilise average precision, which is defined as mean value of precision at 11 levels of recall (from 0 to 1) and describes the shape of a curve.

$$AP = \frac{1}{11} \cdot \sum(p(r)).$$

A successful detection is defined by the overlap ratio A [3]

$$A = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p},$$

where B_{gt} is the area of a ground truth bounding box and B_p of a predicted bounding box. Detection is accepted if the overlap ratio is greater than 0.5. The detection rate corresponds to the proportion of correctly located objects (with $A \geq 0.5$) in the images.

B. Caltech-4 Object Classification

Our main interest is object detection, but since no public implementations of other generative part-based methods are available and they have reported classification results only for outdated or ad hoc datasets (Table I) we selected the Caltech-4 which seems to be the only common benchmark. Caltech-4 contains the categories: $faces$, $airplanes$, $motorbikes$ and $cars\ rear$ with 435, 800, 798 and 1155 images, respectively. We cannot compare to the most recent method by Bergtholdt et al. [29] since they report results only for the faces and their own background images. Also the other similar work by Crandall et al. [24] omits the most difficult class, cars rear, which was replaced by a separate background class. We, however, report their results in Table II where the diagonal elements denote the proportions of correctly classified images.

Table II. Confusion matrices for Caltech-4.

	Our method				Crandall et al. [24] method			
	faces	airplanes	mbikes	cars rear	faces	airplanes	mbikes	bg
faces	91,23	0,00	3,51	5,26	94,90	1,40	3,70	0,00
airplanes	0,00	99,75	0,00	0,25	0,00	90,50	1,25	8,25
motorbikes	0,00	0,00	98,57	1,43	0,00	1,00	96,00	3,00
cars rear	0,00	0,00	0,00	100,00	-	-	-	-

With Caltech-4 our method, utilising only positive examples, performs favourably as compared to other similar methods. Moreover, we simply used the detection scores which are not optimal for the classification task. Table II also reveals the importance of landmark choice. It can be noticed, that most of the misclassified images are confused with cars rear. The errors happen because of too general nature of cars' rear landmarks (simple corners), which can be found everywhere with high probability, showing the necessity of proper landmark choice.

C. Caltech-101 Object Class Detection

In this experiment we omit the other generative part-based methods since no implementations are available and compare our method to the state-of-the-art discriminative method by Felzenszwalb et al. [4]. We report precision-recall-curves for several Caltech-101 categories which show the performance highs and lows of the both methods.

Felzenzwalb et al. method was executed in two modes: 1) with 400 images from the background class as negative examples and 2) with only a single randomly selected background image - the latter intended to approximate our setting of positive examples only.

The result graphs in Figure 6 demonstrate that our method performs comparably to the state-of-the-art method without any optimisation beyond the unsupervised EM of Gaussian mixture model pdf estimation and by using only positive examples. It can be noticed that the categories with a sufficient number of training images (no data or feature randomisation needed), Motorbikes, Faces, stop sign and car side, the both methods achieve almost perfect results. The airplanes category also contains sufficient number of examples, but the results for the both methods are worse, Felzenszwalb completely failing without negative examples. For categories with a small number of examples (only tens of images) our method (blue curves) performs as good or even better as Felzenszwalb et al. with 1 negative example (red curves) in terms of precision.

D. Results for Caltech-101 with Dense Grid Landmarks

To show universality of our method and importance of the landmark choice, we suggest two more experiments with Caltech-101. Both of them are based on the dense grid generated landmarks. The first set of landmarks represents object parts, so we generate a grid of 3×3 points inside the bounding box (Figure 7 top left). The second set of points corresponds to the object contours, thus we generate evenly spaced points along bounding box edges plus one point in the centre (Figure 7 top right). Our experiments show that results with points inside bounding box outperform those on its edges, though are still worse than manually annotated semantically meaningful object parts (Figure 7 bottom). From the Table III we can see that motorbikes can be described with their contours as well as with manually selected parts. Yin yang and dollar bill categories have better detection results with the dense grid generated inside the bounding box. In general, representation with manually annotated parts outperforms both types of the dense grid leaving the optimal automatic landmark selection as an open problem for future.

Table III. DETECTION AVERAGE PRECISION FOR CALTECH-101.

	car side	dollar bill	stop sign	revolver	dragonfly	grand piano
manual ann	95,21	86,74	94,12	89,03	82,32	91,89
grid inside BB	55,97	91,67	81,69	80,26	83,17	82,94
grid on BB border	50,13	70,40	47,03	73,20	65,46	67,81
	menorah	yin yang	faces easy	watch	airplanes	motorbikes
manual ann	81,59	92,87	99,56	93,89	88,58	95,16
grid inside BB	78,40	100,00	98,42	82,16	67,69	87,44
grid on BB border	51,74	96,33	97,76	54,68	63,81	95,02

VI. CONCLUSION

We proposed a probabilistic part-based object class model combining pdfs of part appearances and spatial constellation (Equation 2). Object parts were encoded by Gabor bank features and transformed to probabilities by a training set esti-

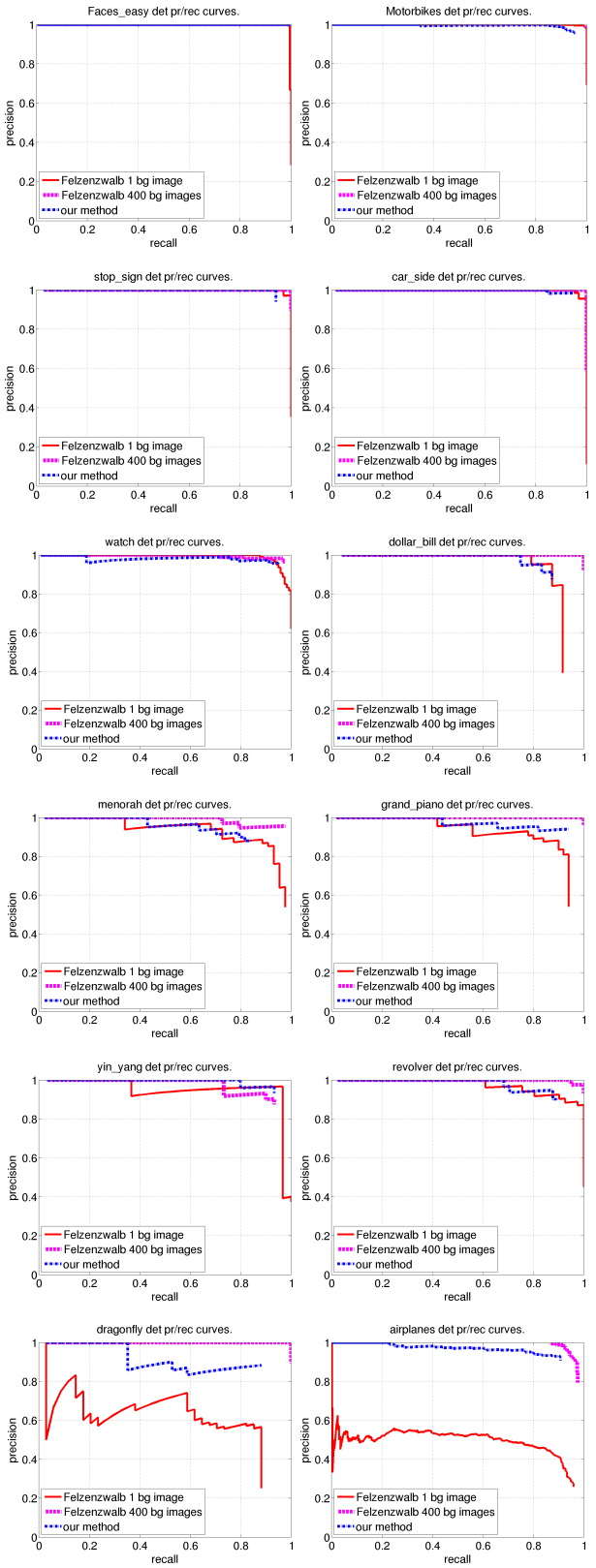

Figure 6. Comparison of our generative positive examples only method and the state-of-the-art discriminative method (Felzenszwalb et al. [4]): faces, motorbikes, revolver, car side, watch, dollar bill, stop sign, airplanes (from left-to-right and top-down).

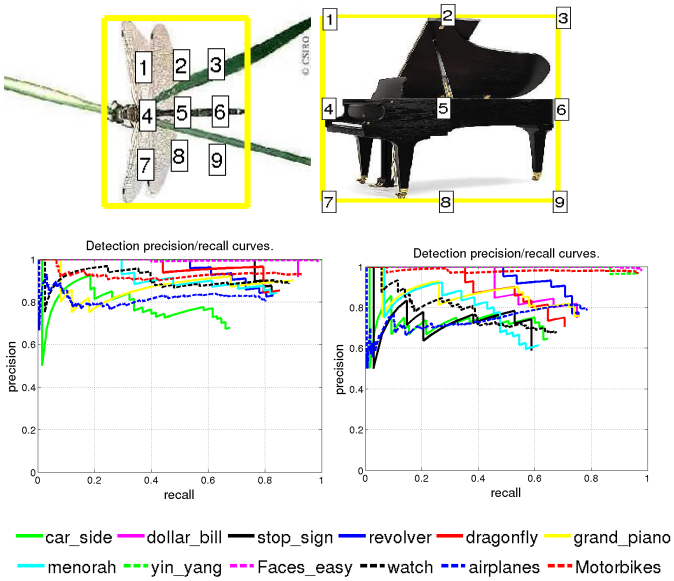

Figure 7. Top: Automatic landmarks inside the bounding box (left) and on the box contour plus centre (right). Bottom: results, correspondingly.

mated Gaussian mixture model (Figure 4). Part constellations (priors) were encoded by Gaussians in the normalised object space (Figure 5 and Algorithm 1). For learning, our method needs only positive examples and we proposed a fast sampling based detection method (Algorithm 2). In the experimental part, our generative and fully probabilistic model achieved object detection accuracy comparable to the state-of-the-art discriminative method using privileged information (annotated parts) and generative procedure (positive only). In our future work, we will address the important problem of automatic part selection. Moreover, we will apply our method to big data problems, such as the ImageNet dataset.

REFERENCES

[1] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Transactions on PAMI 28 (4) (2006) 594–611.

[2] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Tech. rep., California Institute of Technology (2007).
URL http://authors.library.caltech.edu/7694

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, Int J Comput Vis 88 (2010) 303–338.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on PAMI 32 (9) (2010) 1627–1645.

[5] Y. Chen, L. Zhu, A. Yuille, H. Zhang, Unsupervised learning of probabilistic object models (POMs) for object classification, segmentation, and recognition using knowledge propagation, IEEE Transactions on PAMI 31 (10) (2009) 1747–1761.

[6] A. Bar-Hillel, D. Weinshall, Efficient learning of relational object class models, Int J Comput Vis 77 (2008) 175–198.

[7] A. Holub, M. Welling, P. Perona, Hybrid generative-discriminative visual categorization, Int J Comput Vis 77 (2008) 239–258.

[8] D. Parikh, K. Grauman, Relative attributes., in: ICCV, 2011.

[9] O. Russakovsky, F.-F. Li, Attribute learning in large-scale datasets., in: ECCV Workshops (1), 2010.

[10] J. Deng, A. Berg, K. Li, L. Fei-Fei, What does classifying more than 10,000 image categories tell us?, in: ECCV, 2010.

[11] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Transactions on PAMI 29 (6) (2007) 915–928.

[12] V. Vapnik, A. Vashist, A new learning paradigm: Learning using privileged information, Neural Networks 22 (5-6) (2009) 544–557.

[13] G. Csurka, C. Dance, J. Willamowski, L. Fan, C. Bray, Visual categorization with bags of keypoints, in: ECCV Workshop on Statistical Learning in Computer Vision, 2004.

[14] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006.

[15] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, Spatial bag-of-features, in: CVPR, 2010.

[16] B. Leibe, A. Ettlin, B. Schiele, Learning semantic object parts for object categorization, Image and Vision Computing 26 (1) (2008) 15–26.

[17] B. Ommer, J. Buhmann, Learning the compositional nature of visual object categories for recognition, IEEE Transactions on PAMI 32 (3) (2010) 501–516.

[18] P. Carbonetto, G. Dorko, C. Schmid, H. Kuck, N. de Freitas, Learning to recognize objects with little supervision, Int J Comput Vis 77 (2008) 219–237.

[19] S. Todorovic, N. Ahuja, Unsupervised category modeling, recognition, and segmentation in images, IEEE Transactions on PAMI 30 (12) (2008) 2158–2174.

[20] D. Crandall, D. Huttenlocher, Composite models of objects and scenes for category recognition, in: CVPR, 2007.

[21] P. Felzenszwalb, D. Huttenlockher, Pictorial structures for object recognition, Int J Comput Vis 61 (1) (2005) 55–79.

[22] M.C.Burl, M.Weber, P.Perona, A probabilistic approach to object recognition using local photometry and global geometry, in: ECCV, 1998.

[23] R. Rao, D. Ballard, An active vision architecture based on iconic representations, Artificial Intelligence Journal 78 (1995) 461–505.

[24] D. Crandall, P. Felzenszwalb, D. Huttenlocher, Spatial priors for part-based recognition using statistical models, in: CVPR, 2005.

[25] M. Eichner, V. Ferrari, Better appearance models for pictorial structures, in: BMVC, 2009.

[26] G. Heitz, G. Elidan, B. Packer, D. Koller, Shape-based object localization for descriptive classification, Int J Comput Vis 84 (2009) 40–62.

[27] M. Kumar, A. Zisserman, P. Torr, Efficient discriminative learning of parts-based models, in: ICCV, 2009.

[28] Z. Lin, G. Hua, L. Davis, Multiple instance feature for robust part-based object detection, in: CVPR, 2009.

[29] M. Bergtholdt, J. Kappes, S. Schmidt, C. Schnör, A study of parts-based object class detection using complete graphs, Int J Comput Vis 87 (2010) 93–117.

[30] Y. Wu, Z. Si, H. Gong, S.-C. Zhu, Learning active basis model for object detection and recognition, Int J Comput Vis 90 (2010) 198–235.

[31] K. Mikolajczyk, B. Leibe, B. Schiele, Local features for object class recognition, in: CVPR, 2005.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results, http://www.pascal-network.org/challenges/VOC/voc2010/workshop/.

[33] E. Riabchenko, J.-K. Kamarainen, K. Chen, Learning generative models of object parts from a few positive examples, in: ICPR, 2014.

[34] M. Figueiredo, A. Jain, Unsupervised learning of finite mixture models, IEEE Transactions on PAMI 24 (3) (2002) 381–396.

[35] B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fine-grained image categorization, in: CVPR, 2011.

[36] M. Villamizar, J. Andrade-Cetto, A. Sanfeliu, F. Moreno-Noguer, Bootstrapping boosted random ferns for discriminative and efficient object classification., Pattern Recognition 45 (9) (2012) 3141–3153.

[37] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[38] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models – their training and application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.

[39] S. Umeyama, Least-squares estimation of transformation parameters between two point patterns, IEEE Transactions on PAMI 13 (4) (1991) 376–380.