Regularizing Max-Margin Exemplars by Reconstruction and Generative Models

Jose C. Rubio and Björn Ommer
jose.rubio@iwr.uni-heidelberg.de, ommer@uni-heidelberg.de
Heidelberg Collaboratory for Image Processing
IWR, Heidelberg University, Germany

Abstract

Part-based models are one of the leading paradigms in visual recognition. In the absence of costly part annotations, associating and aligning different training instances of a part classifier and finding characteristic negatives is challenging and computationally demanding. To avoid this costly mining of training samples, we estimate separate generative models for negatives and positives and integrate them into a max-margin exemplar-based model. The generative model and a sparsity constraint on the correlation between spatially neighboring feature dimensions regularize the part filters during learning and improve their generalization to similar instances. To suppress inappropriate positive part samples, we project the classifier back into the image domain and penalize against deviations from the original exemplar image patch. The part filter is then optimized to i) *discriminate against clutter, to* ii) *generalize to similar instances of the part, and* iii) *to yield a good reconstruction of the original image patch. Moreover, we propose an approximation for estimating the geometric margin so that learning large numbers of parts becomes feasible. Experiments show improved part localization, object recognition, and part-based reconstruction performance compared to popular exemplar-based approaches on PASCAL VOC.*

1. Introduction

The large intra-class variability of real world object categories is one of the primary challenges of visual recognition. Part-based models are presently the most popular and powerful paradigm to learn representations that characterize all instances of a category despite their variation in appearance, articulation, occlusion, etc. In contrast to holistic, rigid templates [19, 4] and pooling strategies that provide some local shift invariance [17], explicitly modeling object parts [9, 20, 26] not only can explain away object deformation and missing parts. It also provides an explicit parsing of objects (e.g. [11, 23, 3, 24, 6]) and reasoning about their pose and articulation.

Learning part-based models benefits from intensive supervision [3, 11]. Manual labeling and aligning of all instances of a part is, however, very laborious and becomes prohibitive when dealing with large numbers of categories and training samples. Therefore, a common strategy is to automatically learn a small number of parts within object bounding boxes [9]. However, without part annotations, alignment of all training instances of a part is challenging, so that the variability of training samples for a part hampers learning the part representation. A recent solution to this problem are exemplar-based approaches [21] where a classifier is trained with only a single positive sample for a part and a large corpus of negatives. To achieve optimal performance, this exemplar-SVM approach requires several rounds of computationally demanding hard-negative mining, where a set of characteristic negatives are selected based on the single positive. [14] circumvent this mining by employing Linear Discriminant Analysis (LDA) as a very efficient training method for exemplar classifiers at the cost of significantly lower performance. A main limitation of LDA is the implicit assumption that the positive and negative classes share the same covariance. Typically the negative distribution is also utilized for the positive class (a very crude simplification).

Our goal is to avoid negative mining, while retaining the discriminative performance of the part classifier and improving its ability to generalize to other part instances, which are related to the original exemplar. Rather than selecting hard negatives, we represent the negative by a generative model and we learn another generative model for features from within object bounding boxes of a category. The distribution of positives couples spatially neighboring feature dimensions (e.g. HoG cells) and thus reduces its degrees of freedom. When adding further part instances from other training images to the original exemplar patch, we need to prevent the classifier from drifting away from the original exemplar part due to badly aligned samples and clutter. Otherwise we would end up with a representation corresponding to an uninformative, averaged patch. Therefore, we propose a perceptual regularization that projects

the part classifier back onto a wavelet representation of the image patch corresponding to the part. Then we penalize against distortions from the original exemplar patch in the (wavelet representation of the) image domain. We suggest a optimization strategy as a sequence of Second Order Cone problems. Additionally we propose a sub-optimal approximation that speeds-up the training process significantly (faster than popular ESVM [21] or DPM[9]) while retaining the object recognition performance. Figure 1 shows a scheme of the approach.

The result of our learning process is a set of exemplar part detectors that trade-off specificity against generalization and that directly correspond to object image regions. Their localization performance and recall is boosted significantly by the addition of the generative model as well as additional positive instances. Since our approach does not require positive or hard negative mining, training is fast (a few seconds per part). Thus part models with large numbers of parts become feasible. We evaluate the improvement of part localization as well as of object recognition performance on the challenging PASCAL VOC data-set. Moreover, our model provides part-based reconstructions of objects. Here qualitative and quantitative comparisons illustrate the robustness of our method against part misalignment.

2. Related Work

Middle-level representations based on classifiers trained on localized parts have recently experimented an increase of popularity due to their simplicity and competitive performance applied to a large variety of problems. In [1] exemplar parts are used to represent 3D objects and to align them with their 2D correspondences, while in [6] object detection is performed by pooling the filter responses of a large set of exemplar classifiers.

There has been several recent works [16, 5, 8] aiming for improving the generalization of those weak classifiers by means of gathering additional training parts and therefore, moving from exemplar towards multiple-positive part classifiers. In both [16] and [8] parts are refined incrementally starting from a single exemplar after filtering the additional candidates using validation data. The approach of [5] looks for discriminative *modes* on a density representation of the training data to then draw candidate parts from those modes.

In a slightly different avenue of research, there are methods that include generative models in their formulation in order to improve the classifiers regarding efficiency or accuracy. In [14], a generative representation of the negative data distribution is proposed to train exemplar classifiers very efficiently, by simply decorrelating HoG features. The follow-up work [13] uses this same technique to accelerate the training of DPM. Other approaches such as [10] propose instead to learn a generative model of the positive class to perform transfer learning and train classifiers from few examples. In [7] covariance features inspired in such generative models are used to determine *good* filters and improve the ranking of parts.

3. Method

Opposite to the approaches that focus on obtaining a consistent set of candidates to avoid hampering the classifier, we aim to directly improve the generalization performance of exemplar parts while retaining their specificity. We propose a discriminative learning framework that includes a generative model to represent the distribution of the positive and negative categories. Contrarily to LDA our model allows for distinct covariances for each positive and negative class. The negative distribution allows for fast training by avoiding mining hard-negatives. The positive distribution serves to reduce the classifier degrees of freedom, imposing a category-specific structured prior that assures the quality of the filters (Sect. 3.1). We also propose a novel *perceptual* regularizer that prevents the classifier from changing the concept represented by its original exemplar, when training with multiple positive instances. Additionally, this regularization term links our model to the image domain, thus enabling the generation of visual representations of our part-based models (Sect. 3.2). Finally, our optimization strategy trains models efficiently (on the order of LDA) while achieving performance superior to exemplar-SVM with hard negative mining (Sect. 3.3, and 3.4).

3.1. Generative Regularization of Max-Margin Exemplars

Even though the recent success of LDA exemplars is remarkable [14], their standalone performance is still inferior to SVM with hard mining. An important limitation of LDA is the assumption that the distributions of each category share the same covariance with their mean shifted. While Quadratic Discriminant Analysis overcomes this limitation, its decision boundary is defined by a quadratic function, which makes it computationally intractable for large-scale visual recognition problems.

The intuition behind the success of LDA for exemplar-based models is that a uni-modal covariance is able to capture generic spatial statistics of the background distribution, so that a simple decorrelation boosts the discriminative feature dimensions of the exemplar classifier decision surface. However, the model solely focuses on the distribution of the negatives. The distribution of the positives is simply assumed to be the same. This assumption is obviously incorrect and contradicts the analysis of recent works such as [7], that identify a good HoG filter to have strong correlations of neighboring cells.

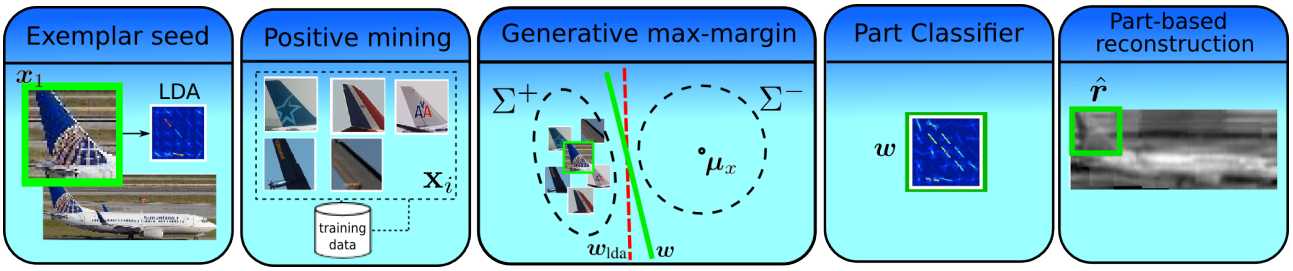

Figure 1. Scheme of the approach. First local patches are extracted from positive boxes. With each, an LDA classifier is trained. Then, positive mining is performed using that classifier on positive bounding boxes. Following, our generative max-margin classifier is trained using those positives as well as the distributions Σ^+ and Σ^-. Note that the red dashed line indicates the classifier obtained with regular LDA, while the green solid line represents our classifier, that *aligns* acoording to the shape of positive and negative covariances. Finally, the improved part filters are obtained, and part-based reconstruction of the original instance can be performed.

Our first goal is to determine a hyper-plane $\boldsymbol{w}$ with maximal Mahalanobis distance to the distribution of negatives centered at the mean $\boldsymbol{\mu}_x$ of the negatives and to the positive distribution centered at the single positive exemplar feature $\boldsymbol{x}_1$. This family of classifiers is closely related to Support Vector Machines [15, 18], and aims to establish a decision hyper-plane based on *local* information (exemplar features) as well as *global* evidence (probability densities). Let us denote the covariances of the positive and negative distributions as Σ^+ and Σ^- respectively. The positive covariance is specific for an object category and it is trained using all the HoG features present within the bounding boxes of the category objects. The negative covariance is common for all categories and it is trained using all available HoG windows in the data-set. In [14] is pointed out the difficulty of learning category-specific covariances due to the high dimensionality of the covariance matrix. However, we aim for decorrelating HoG features whose size is relatively low (5 by 5 cells) in comparison to whole object filters (12 by 12 cells). Additionally, we also regularize both covariances by adding a small value (.01) to its diagonal, which corresponds to adding an isotropic prior to Σ. We formulate the optimization problem as

$$\begin{aligned} \underset{\gamma,b,\boldsymbol{w}}{\arg\max} \quad & \gamma - \|\boldsymbol{w}\|_1 \\ \text{s.t} \quad & \boldsymbol{w}^\top \boldsymbol{x}_1 - b \geq \gamma\sqrt{\boldsymbol{w}\Sigma^+\boldsymbol{w}} \\ & -\boldsymbol{w}^\top \boldsymbol{\mu}_x + b \geq \gamma\sqrt{\boldsymbol{w}\Sigma^-\boldsymbol{w}} \\ & \gamma \geq 0 \end{aligned} \tag{1}$$

where we aim to explicitly maximize the functional margin γ. Note that we can not impose any scaling constraint on the margin and maximize the geometric margin instead, because given that $\Sigma^+ \neq \Sigma^-$ we cannot assume $\gamma\sqrt{\boldsymbol{w}\Sigma^+\boldsymbol{w}} = 1 = \gamma\sqrt{\boldsymbol{w}\Sigma^-\boldsymbol{w}}$.

To facilitate the overall optimization it is convenient to reduce the degrees of freedom of the classifiers to drive them towards good quality optima. We include a L1 normalization for this purpose, which encourages structured activations within HoG cells. It is also tempting to further exploit the positive distribution to induce a structured prior over the positive class to encourage strong correlations of the filter weights $\boldsymbol{w}_{\text{lda}}$ in nearby locations. Those two characteristics, cell covariance and cell cross-covariance, are indicators of a *good* filter, as pointed out in [10, 7]. We impose a structured prior in the positive covariance as

$$\Sigma^+ = \Sigma^{\boldsymbol{x}^+} \circ \mathbf{K} \tag{2}$$

where $\mathbf{K}$ is a sparse block matrix having ones only on the elements corresponding to neighboring HoG cells in 4 directions: horizontal, vertical, diagonal-left, diagonal-right, to encourage high cross-covariance on neighboring cells. The matrix $\Sigma^{\boldsymbol{x}^+}$ is a covariance matrix computed on features extracted exclusively from objects of the positive class (as previously explained), and $\circ$ denotes the Hadamard product.

3.2. Regularization by Reconstruction

Let us now further improve the generalization performance (their recall) of our filters while preserving their discriminative power. Since the filters are trained with only a single positive we need to add additional positive samples to improve their recall and have them generalize better to other similar samples. Since part-based annotations re not available for this large number of parts, automatically gathering additional positive samples inevitably implies outliers and spatially badly aligned samples, which reduce classifier performance as can be seen in Figure 5. The effect in the filters themselves can be seen in Figure 2, where naively gathering multiple positives and training an LDA classifier (MLDA) averages out miss-aligned features, with the loss of the original exemplar specificity.

To make the models robust to badly chosen samples, we need a regularization term that prevents a classifier from drifting away from the original object region that the exemplar represented in the first place. Thus, we need to estimate the image region that the filter $\boldsymbol{w}$ originally corre-

Figure 2. Visual comparison of part classifiers. Each row shows a part classifier. The first column shows the original patch, the central column the 5 nearest neighbors (out of 10 used for training), and the third column shows a visual representation of the positive weights of the resulting classifier for ELDA (1 positive), MLDA (10 positives) and ours (10 positives). In the first row, the miss-alignment of plane-rudders produces a wider edge in the MLDA while ours is compact. In the second row, the MLDA filter *hides* the details of the bent wing-tips present in the original exemplar, while our filters keep those details. The third row shows an example of a part with very low repetitivity and plenty of false positives. The response of the MLDA filter is averaged out, while ours preserves most of the signal of the original ELDA.

sponds to, and penalize deviations therefrom. [25] have utilized a mapping from HoG space to the image domain using ridge-regression to primarily visualize and study the filters. We now invert the HoG filter and map it back into the same domain to regularize the learning of the part classifier. Since several images map to the same HoG feature, the inversion is not bijective and the inverted representation resembles an averaging of several patches obtained by shifting few pixels in every direction. This amplifies details that are robust to misalignment and weakens those that are likely to get lost when assembling a large set of related patches.

Let $\boldsymbol{r}$ denote the wavelet transformation of an image region and $\boldsymbol{x}$ is the corresponding HoG feature. The wavelet representation consists of the approximation coefficients of the *daubechies* wavelet decomposition. This transformation removes high-frequency data not interesting for the sake of the reconstruction and facilitates the covariance estimation by reducing to half the dimensionality of the image domain. We stack $(\boldsymbol{x}, \boldsymbol{r})^\top$ and compute the covariance matrix $\Sigma = \left(\begin{smallmatrix} \Sigma^{xx} & \Sigma^{xr} \\ \Sigma^{rx} & \Sigma^{rr} \end{smallmatrix}\right)$, and $\boldsymbol{\mu} = [\boldsymbol{\mu}_x \boldsymbol{\mu}_r]$ right over all features sampled from all training images. We denote with r the image domain and with x the HoG domain. The upper-left part of Σ is the covariance of all negative features $\Sigma^{xx} = \Sigma^-$ from Eq.(1). Now we can map $\boldsymbol{w}$ back into the image domain by projecting it onto

$$\hat{\boldsymbol{r}} = \Sigma^{rx}\boldsymbol{w} + \boldsymbol{\mu}_r \tag{3}$$

and then apply the inverse wavelet transformation.

Let us denote the set of related positive samples (including the original exemplar $\boldsymbol{x}_1$) as $\{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_N\}$, and let $\boldsymbol{r}_1$ be the wavelet representation of the image patch corresponding to the original exemplar. The set of candidate positive features is obtained by running the classifiers within a validation set of images containing objects of the category and keeping the N nearest neighbor detections that lie within the boundaries of the ground-truth bounding boxes (Figure 2 shows examples of candidate sets). Since the inversion is in closed form we can easily include it in the optimization of Eq.(1) as a regularization term,

$$\begin{aligned} \underset{\gamma,b,\boldsymbol{w},\xi}{\arg\max} \quad & \gamma - \left[\|\boldsymbol{w}\|_1 + C\sum \xi_i + \beta\|\hat{\boldsymbol{r}} - \boldsymbol{r}_1\|_2^2\right] \\ \text{s.t} \quad & \boldsymbol{w}^\top \boldsymbol{x}_i - b \geq \gamma\sqrt{\boldsymbol{w}\Sigma^+\boldsymbol{w}} - \xi_i \quad \forall i \\ & -\boldsymbol{w}^\top \boldsymbol{\mu}_x + b \geq \gamma\sqrt{\boldsymbol{w}\Sigma^-\boldsymbol{w}} \\ & \hat{\boldsymbol{r}} = \Sigma^{rx}\boldsymbol{w} + \boldsymbol{\mu}_r \\ & \xi \geq 0,\ \gamma \geq 0 \end{aligned} \tag{4}$$

Since we are considering multiple positive features and some of them could be outliers, we need to take the separability into account by introducing a slack variable ξ_i for each positive feature. This is not necessary in the negative side, since the mean of the negative distribution will never violate the margin. The regularization term encourages the inverted classifier $\hat{\boldsymbol{r}}$ to stay close to the wavelet transform of the original exemplar patch $\boldsymbol{r}_1$.

3.3. Optimization

The coupling of γ and $\boldsymbol{w}$ makes the problem non-convex. We approach it in an iterative fashion, similarly to [15], as

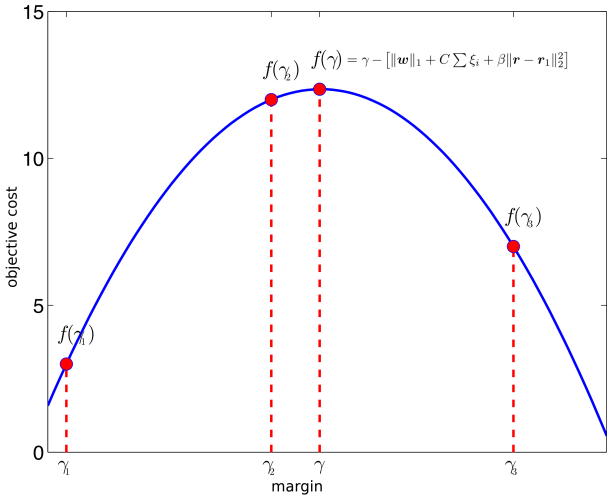

Figure 3. Three point pattern to perform a Quadratic Line Search to estimate the maximum margin, with a fixed hyperplane $\boldsymbol{w}$. The blue continuous line denotes the unimodal distribution of the objective as a function of the margin γ.

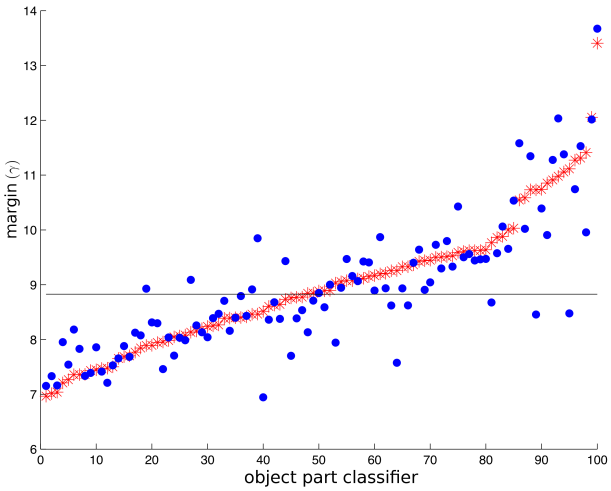

Figure 4. Optimal margin of 100 bicycle parts against the predicted margin. The red crosses indicate the optimal margin trained using the sequential SOCP optimization (Sect. 3.3). The blue circles indicate the margin predictions using linear regression on filter characteristics Ω (Sect. 3.4). The black line indicates the training average as a naive margin predictor.

a sequence of Second Order Cone Programming (SOCP) problems. When the margin γ is fixed, the optimization is equivalent to minimizing $\|\boldsymbol{w}\|_1 + C\sum\xi_i + \beta\|\boldsymbol{r} - \boldsymbol{r}_1\|_2^2$ under the same constraints, which is feasible in polynomial time using interior-point methods. Then parameters $\boldsymbol{w}, \xi, \boldsymbol{r}$ are optimized, before searching for the next optimal margin γ using a line search strategy over the original objective function $f(\gamma) = \gamma - \left[\|\boldsymbol{w}\|_1 + C\sum\xi_i + \beta\|\boldsymbol{r} - \boldsymbol{r}_1\|_2^2\right]$. We approach the line search by fitting a quadratic function to a three point pattern spanned by three values $\gamma_1, \gamma_2, \gamma_3$, and their corresponding objective values $f(\gamma_1), f(\gamma_2), f(\gamma_3)$ (See Fig. 3). To generate the next γ we search for the one that maximizes the interpolated estimate and establish a new three point pattern using the new γ as the central point together with the two nearest previous values.

Since we need to establish a range for the line search a priori, we use a trust-region approach. We assume $f(\gamma)$ to be unimodal in a interval $\gamma \in (0, \infty)$ of the margin. We set the three initial values of γ to be uniformly distributed within a small *trust region* of size $2s$ ($s = 15$ in our implementation), as $\gamma_1 = \gamma_2 - s, \gamma_2 = \frac{\boldsymbol{w}^\top \boldsymbol{\mu}_x}{\|\boldsymbol{w}\|}, \gamma_3 = \gamma_2 + s$. At each step, if the maxima of the quadratic fit is found to be outside the trust region, we progressively expand it in the direction that the maxima is expected. The iterative process runs until reaching a maximum number of iterations or the stopping condition $tol > |\gamma_{prev} - \gamma_{new}|$ is satisfied.

3.4. Speeding up the training of part filters

A great advantage of representing the negative training set with a generative model (the parametric model of Σ^-) is to avoid performing several rounds of hard negative mining to train each of the part classifiers, which is the most expensive process of standard exemplar-SVM training. Although the proposed training procedure is faster than Exemplar-SVM, the iterative optimization presented in Sect. 3.3 is still computationally more costly than LDA.

Let us know propose an alternative to speed-up the training of the classifiers by directly estimating an approximation of the functional margin. We first sample 1000 regions from object bounding boxes (details of the sampling can be found in the Experiments section). For each of those regions we train a part classifier i using the method described above, inferring the classifier parameters $\boldsymbol{w}_i, b_i$ as well as the optimal functional margin γ_i. Intuitively, the value of the optimal margin will depend on the shape of the positive and negative data distribution, as well as on the set of positive instances and the classifier hyperplane. Thus, we aim for learning a margin predictor g, using a set of characteristics $\boldsymbol{\Omega}$ extracted from the filters themselves,

$$\boldsymbol{\Omega} = \left[d(\mathbf{X}^+), \boldsymbol{w}^\top \mathbf{X}^+, \|\boldsymbol{w}\|_2, \boldsymbol{w}\Sigma^+\boldsymbol{w}, \boldsymbol{w}\Sigma^-\boldsymbol{w}\right] \quad (5)$$

where $\mathbf{X}^+$ is the matrix of all positive instances $\boldsymbol{x}_i$ and $d(\mathbf{X}^+)$ is a vector of Mahalanobis distances of those instances to the negative distribution $< \boldsymbol{\mu}_x, \Sigma^- >$.

We perform a linear regression to map the feature characteristics of a part to its functional margin $\hat{\gamma}_i = g(\boldsymbol{\Omega}_i) + \epsilon$. In such way, we can perform the optimization in one sweep, which takes around 10 seconds per part, at the expense of a slight drop in performance (1%). Figure 4 shows the optimal margin obtained with the alternating optimization of Sect. 3.3 and that of linear regression estimate presented in this section on 100 parts of the bicycle category.

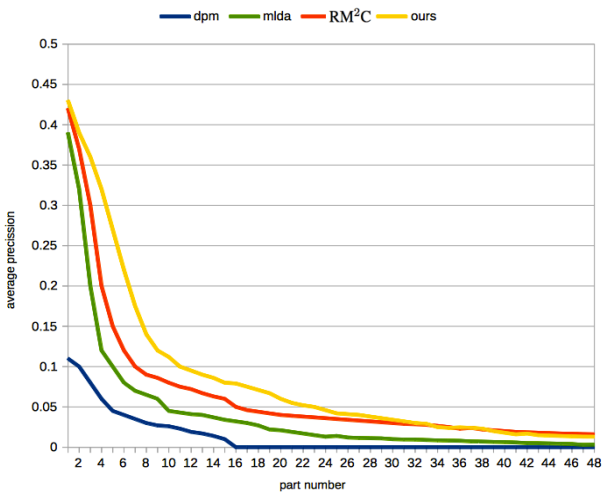

Figure 5. Part localization. Detection accuracy of *i)* the 48 DPM parts [9] against a set of 48 parts trained with *ii)* MLDA (10 positive instances), *iii)* ESVM on object patches [21] or, equivalently, the part filters of [6] (1 positive instance), and *iv)* our final approach (sequential SOCP). The average gain in AP between our iv) and iii) is 4%, although our training is roughly $4\times$ faster.

Category	ELDA [25]	MLDA-5	MLDA-10	Ours
airplane	.58	.51	.47	.60
bicycle	.47	.46	.42	.51
bird	.50	.47	.46	.54
boat	.61	.60	.55	.61
bottle	.82	.81	.81	.84
bus	.65	.56	.48	.66
car	.78	.75	.72	.82
cat	.62	.58	.57	.60
chair	.79	.70	.68	.81
cow	.67	.60	.56	.64
table	.74	.65	.56	.79
dog	.56	.50	.48	.58
horse	.62	.61	.58	.61
motorbike	.70	.67	.60	.73
person	.58	.55	.53	.61
plant	.76	.75	.72	.76
sheep	.56	.56	.54	.57
sofa	.81	.75	.73	.80
train	.79	.74	.69	.78
tvmonitor	.83	.82	.79	.85
mean	.67	.63	.60	.69

Table 1. Average reconstruction quality per category in PASCAL VOC 07 measured by maximum cross-correlation between reconstruction and original

3.5. Part-based Object Recognition

Given the set of trained part-classifiers of a category we need to combine them to represent the object instances of that same category. For each image, parts are evaluated by densely convolving the part-filters along the image region of an object box, at all locations and scales. Then we aggregate the contribution of all parts by max-pooling the filter responses in a regular grid within the object bounding box. In our implementation, we employ a three-level grid with 4×4, 2×2 and 1 grid cells in each level. Therefore, after pooling and concatenating we obtain a high-dimensional representation of an object instance of $N \times (16+4+1)$ dimensions, where N is the number of part filters. To provide a final ranking of objects we train a linear classifier (SVM) on top of those part-based representations.

4. Experiments

Subsequently we compare our exemplar-based approach against several popular exemplar and part-based models, which also need no part annotations. The experiments on the standard benchmark data-set of PASCAL VOC 2007 evaluate part localization performance, present a quantitative and qualitative comparison in context of object reconstruction, and investigate recognition performance.

Regarding the experimental set-up, in all our experiments we use HoG filters with a size of 5 by 5 cells. The C and β parameters of the classifiers are estimated by grid-search on a validation set and are set to 0.5 and 0.1 respectively. We run the sequential SOCP optimization for a maximum of 100 iterations, with a tolerance of 0.01. To estimate the wavelet representation and the ridge-regression model, we scale the image patches to fit 4x4 pixels per HoG cell. That is, patches of 20x20 wavelet approximation coefficients.

4.1. Part Localization Performance

A benefit of part-based models is that they not only detect objects but also accurately localize object parts, which is crucial for various applications ranging from pose estimation and gait analysis to quality control in industrial inspection. To evaluate the accuracy of part detection in novel images we employ the key-point annotations of Pascal VOC 2010 that were introduced in [3]. Note that contrary to [3] neither our model nor the part-based models that we compare to are trained with part annotations. We only use the key-point annotations in this section to evaluate the localization of individual part classifiers, but nowhere else.

We sample exemplar patches centered at the annotated key-points, with a random size smaller than half the object bounding box. For each of those parts, we train an Exemplar SVM with hard mining [21], use the LDA approach of [14] with one positive, an LDA classifier with 10 positives (abbreviated as MLDA) and our final discriminative approach with generative regularization and reconstruction with 10 positives, Sect. 3.2. Positive instances for our approach and MLDA are obtained by simply running exemplar LDA over the category bounding boxes of the validation set and keeping the 10 nearest neighbors for an exemplar. Evaluation follows the standard VOC protocol, considering detections correct if part overlap is greater .5.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	mean	time;parts
ESVM [21]	37.2	58.1	8.6	12.4	23.1	54.7	55.4	24.2	20.2	26.4	28.2	18.1	50.5	49.6	36.8	12.1	19.3	34.6	46.9	43.2	33.0	40h;1K
ELDA [14]	37.1	53.2	7.5	11.9	22.3	51.1	54.2	21.7	19.6	24.3	26.9	18.2	47.8	48.6	34.5	12.1	19.2	3.4	45.3	42.1	31.6	1sec;1K
MLDA	35.4	54.5	7.2	10.6	21.7	46.8	52.9	20.0	18.2	23.3	24.5	16.6	45.7	47.8	33.9	11.2	18.1	31.4	43.1	40.6	30.1	1sec;1K
w/o Σ^+	37.3	57.2	9.6	12.7	24.2	55.1	56.1	23.5	21.1	25.5	28.7	18.9	49.8	50.0	36.3	12.8	20.1	34.6	46.3	42.2	33.1	14h;1K
Final	39.9	57.8	9.2	14.1	27.7	57.8	56.4	25.2	22.5	26.2	29.8	21.2	49.7	50.5	37.7	13.1	21.9	35.5	46.9	43.8	34.3	14h;1K
Final/Fast	39.0	56.3	8.9	13.0	26.8	57.4	55.1	24.5	22.2	24.9	29.5	20.5	48.1	50.2	37.3	12.8	21.2	35.1	46.1	43.0	33.6	14min;1K
DPM [9]	33.2	60.3	9.8	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	19.5	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7	5h;48
LLDA [13]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.4	7min;48
RM^{2}C [6]	37.0	58.3	12.0	14.7	22.9	51.3	51.7	23.7	21.7	25.0	29.0	20.6	51.4	46.1	36.3	12.7	22.3	35.1	43.9	41.8	32.9	40h;1K

Table 2. Detection performance on PASCAL VOC 2007 (AP %). ESVM: exemplar-SVM with hard-negative mining. ELDA: LDA with one positive sample. MLDA: multiple (10) positive samples with no reconstruction regularization nor positive generative model. In w/o Σ^+: our model with regularization by reconstruction but without positive generative model. In Final: the approach of Sect. 3.2. In Final/Fast: the approach with accelaration by margin regression of Sec. 3.4. Related approaches: DPM [9], LLDA [13] and RM^{2}C [6]. The last row shows the average training time of each approach, which depends on the number of parts that need to be trained. Whereas DPM needs 5 hours to train 48 parts our fast method trains 1000 parts in 14 minutes.

Figure 5 evaluates the detection accuracy of individual parts. Parts are ranked according to their AP over all categories. Comparison against MLDA, the DPM parts of [9], and to the part classifiers of [6] shows an average gain of 4% over the previously best results by [6], which is an exemplar-SVM. This gain is mainly due to a considerable increase in recall of our parts.

4.2. Computational Performance

Training a single ESVM classifier [21] (or a part of [6]) on an Intel i7 CPU takes around 30 minutes per part due to hard negative mining on the training set. Our part detectors optimized with sequential SOCP take approximately 10 minutes per part. Speeding up our model as detailed in Sect. 3.4 reduces training time of a part to the order of 10 seconds, i.e., a speed-up of roughly $180\times$ over the popular ESVM. Also note that part training can easily be parallelized, so learning a category model with 1000 parts takes only a few minutes. This is a significant gain over [6] with the same number of parts (more than 24 hours) or DPM (around 5 hours on the same hardware, i.e., around $20\times$ slower). Regarding testing time, filter responses are computed with a single matrix multiplication (all at the same time). Evaluating $1K$ parts in a whole image takes around 13 seconds.

4.3. Visual Parsing

Our part detectors are optimized not only to discover object parts [6] and discriminate against clutter [22]. They are also trained to provide a good reconstruction of original object regions. Thus they can be backprojected into the image domain to parse and explain objects and provide more information than only a bounding box with class label. To generate an object visualization, in a query image we gather all parts that cover the object box and place the reconstruction of each part at the corresponding location, averaging over overlapping regions (see Fig. 6).

Related approaches to reconstruction such as hoggles [25] are typically holistic (reconstruct the full bounding box rather than learned object parts) and based only on a single positive exemplar sample for a part. Using multiple positive samples for training (MLDA) reduces details due to bad sample alignment and clutter as can be seen in Fig. 6. In contrast our regularization yields additional details as can for instance be seen at the tail of the cat, the cockpit of the airplane, the tail of the car, or the wheel of the bike. Moreover, the average reconstruction quality per category (measured by maximum cross-correlation between reconstruction and original object bounding box) is improved when comparing against standard ridge regression with 1,5, and 10 positive instances, in the PASCAL VOC 2007 dataset (see Table 1).

4.4. Object Recognition

Let us now evaluate our parts in the context of object detection. We sample initial exemplar patches from all object bounding boxes of all classes. Patches have a random size not larger than half the object bounding box. We first discard patches containing little detail based on their average image gradient to remove those that contain only noise or compression artifacts. Similarly to [2], each remaining HoG descriptor $\boldsymbol{x}$ gets a score

$$q(\boldsymbol{x}) = \frac{1}{(\boldsymbol{x}-\boldsymbol{\mu}_x)\Sigma^{xx\,-1}(\boldsymbol{x}-\boldsymbol{\mu}_x)} \tag{6}$$

thus ranking first those part filters that are further from the background distribution. We retain the 1000 highest scoring parts per class and with those we compute object representations as explained in Sect. 3.5. We apply the same process over test data by extracting object hypothesis following the same strategy as [6].

The state-of-the-art of object detection in PASCAL VOC is currently achieved with Deep Learning approaches [12] with an AP of 53.7%. However, such approaches require pre-training with extensive datasets (ILSVRC 1.2 million images). We focus our comparison with works based on HoG that follow the VOC *comp3* (only PASCAL training data). Table 2 evaluates the object detection performance obtained with different algorithms for unsupervised part

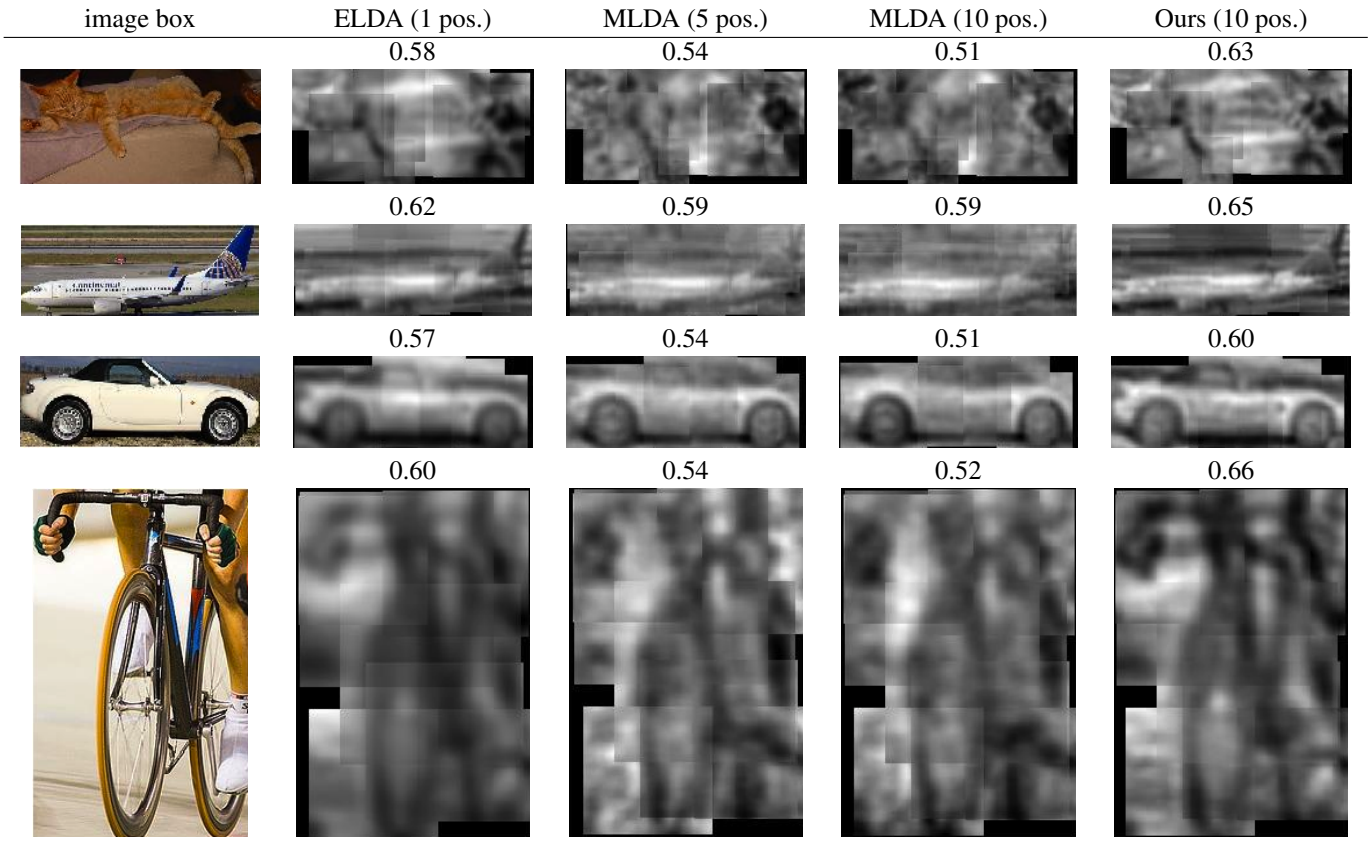

Figure 6. Part-based reconstruction of query images with LDA (ridge-regression) [25] with one positive (second column), with 5 instances per part (third column), 10 instances per part (fourth column), and with our final model (fifth column). On top of each reconstruction its quality is given by its maximum cross-correlation against the original image.

training on PASCAL VOC 2007. Although we have significantly improved the speed of part training, part localization performance, and visual reconstruction, the detection performance is not impaired. We emphasize the training time of individual part filters to show that our models with margin prediction achieve competitive performance with extremely fast training times. In approaches like DPM and its fast version Latent-LDA [13] the parts are tightly linked to the root so that all of them have to be trained jointly. Our model treats parts independently so that the training can be paralelized very efficiently and by adding additional parts a posteriori the object model does not have to be retrained from scratch. For the computational performance presented in Tab. 2 we used a parallel pool of a maximum of 12 threads. Having additional threads would further increase the time differences between our approach and DPM.

Consistently with the experiment in Sect. 4.1 the performance of models with multiple positives and lacking regularization (MLDA) suffers significantly. Only by including the reconstruction regularization (Tab. 2 fourth row) provides a performance boost of 3%. Including the structured prior of Eq.(2) in the positive class provides an additional accuracy increase of 1.2%. Moreover, we see that pooling a 1000 parts and combining them in a joint model at least partially compensates for weak part classifiers (2.9% performance loss between ESVM and MLDA) as opposed to a drop of 5% for individual parts in Sect. 4.1. However, this is computationally daunting without our speeding up of training.

5. Conclusions

We have addressed a key problem of part-based models, the efficient training of large numbers of part classifiers with multiple samples and without requiring part annotations or costly hard negative mining. This is an orthogonal direction to previous work [16, 5, 8] that aims at improving the positive mining process. We integrate a generative regularization into the discriminative part training and enforcing good visual reconstruction. The proposed part training algorithm is significantly faster than popular ESVM [21] or DPM [9], while retaining object recognition performance and improving part localization and the ability to reconstruct and explain images of objects.

References

[1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 2

[2] M. Aubry, B. C. Russell, and J. Sivic. Painting-to-3d model alignment via discriminative visual elements. *ACM Trans. Graph.*, 2014. 7

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009. 1, 6

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *ICCV*, 2005. 1

[5] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 494–502, 2013. 2, 8

[6] A. Eigenstetter, M. Takami, and B. Ommer. Randomized max-margin compositions for visual recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 6, 7

[7] G. S. Ejaz Ahmed and S. Maji. Knowing a good hog filter when you see it: Efficient selection of filters for detection. In *European Conference on Computer Vision (ECCV)*, 2014. 2, 3

[8] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. In *CVPR*, 2013. 2, 8

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1, 2, 6, 7, 8

[10] T. Gao, M. Stark, and D. Koller. What makes a good detector? – structured priors for learning from few examples. In *European Conference on Computer Vision (ECCV)*, October 2012. 2, 3

[11] G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes. Parsing occluded people. In *CCVPR*, 2014. 1

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[13] R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 2, 7, 8

[14] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 2, 3, 6, 7

[15] K. Huang, H. Yang, I. King, and M. R. Lyu. Maxi-min margin machine: Learning large margin classifiers locally and globally. *IEEE Transactions on Neural Networks*, 2008. 3, 4

[16] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2, 8

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012. 1

[18] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 2003. 3

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*. 1

[20] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision (ECCV)*, May 2004. 1

[21] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 2, 6, 7, 8

[22] A. Monroy and B. Ommer. Beyond bounding-boxes: Learning object shape by model-driven grouping. In *ECCV*, 2012. 7

[23] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*. 2007. 1

[24] G. Sharma, F. Jurie, and C. Schmid. Expanded Parts Model for Human Attribute and Action Recognition in Still Images. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[25] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. *ICCV*, 2013. 4, 6, 7, 8

[26] P. Yarlagadda and B. Ommer. From meaningful contours to discriminative object shape. In *European Conference on Computer Vision (ECCV)*, 2012. 1