This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

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

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 approx imation 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 generaliza tion and that directly correspond to object image regions. Their localization performance and recall is boosted signif icantly 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 num bers of parts become feasible. We evaluate the improvement of part localization as well as of object recognition perfor mance on the challenging PASCAL VOC data set. More over, our model provides part based reconstructions of ob jects. Here qualitative and quantitative comparisons illus trate the robustness of our method against part misalign ment.

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 perfor mance 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 there fore, 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 addi tional candidates using validation data. The approach of [5] looks for discriminative *modes* on a density representation of t e training ata to then draw candidate parts from those modes.

In a slightly di ferent avenue of research, there are meth ods that include generative models in their for ulation in order to improve the classifiers regarding efficiency or ac curacy. In [14], a generative representation of the negative data distribution is proposed to train exemplar classifi rs 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 con sistent set of candidates to avoid hampering the classifier, we aim to directly improve the generalization performance of exemplar parts while retaining their specificity. We pro pose a discriminative learning framework that includes a generative model to represent the distribution of the posi tive and negative categories. Contrarily to LDA our model allows for distinct covariances for each positive and nega tive class. The negative distribution allows for fast train ing by avoiding mining hard negatives. The positive dis tribution 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 repr sented by it original xemplar, when training with multip e positive instances. Addition ally, this regularization term links our model to the image domain, thus enabling the generation of visual representa tions 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 exe pl r 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 cap ture generic spatial statistics of the background distribution, so that a simple decorrelation boosts the discriminative fea ture 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 as sumed to be the same. This assumption is obviously incor rect and contradicts the analysis of recent works such as [7], that identify a good HoG filter to have strong correlations of neighboring cells.

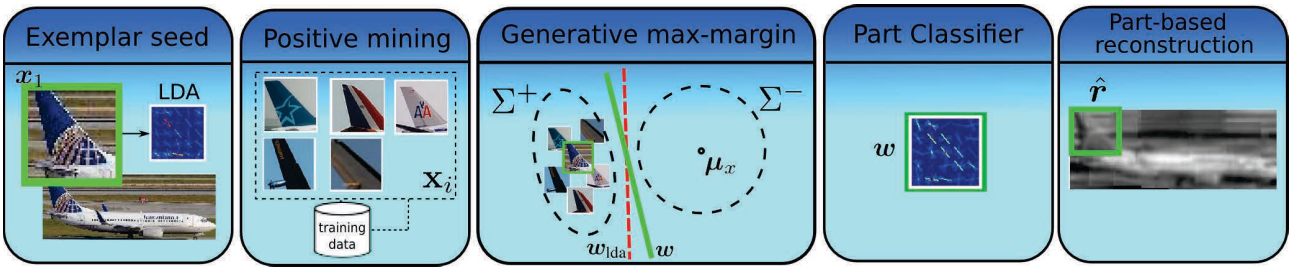


Figure 1. Scheme of the approach. First local patches are extracted from positive boxes. With each, an LDA classifier is trained. Then, positive mining is performed using that classifier on positive bounding boxes. Following, our generative max-margin classifier is trained using those positives as well 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 po itive and negat ve distribu tions as Σ^+ and Σ^- respectively. The positive covariance is specific for an object category and it is train d 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 win dows in the data-set. In [14] is pointed out the difficulty of learning category specific covariances due to the high di mensionality of the covariance matrix. However, we aim for decorrelating HoG features whose siz is relatively low (5 by 5 cells) in comparison to whole object filters (1 by 12 cells). Additionally, we also regularize both covariances by adding a small value (.01) to its diagonal, which corre sponds to adding an isotropic prior to Σ. We formulate the optimization problem as

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

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

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 nor malization 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 $\quad_{\text{lda}}$ 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 neighbori g HoG ells in 4 directio s: 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 perfor mance (their recall) of our filters while preserving their dis criminative 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 gath ering 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 ex emplar represented in the first place. Thus, we need to es timate 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 freq ency data not interesting for the sake of the reconstruction and facilitates the covariance estimation by reducing to hal the dimensionality of the image do main. 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 (includ ing 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 f candidate pos itive features is obtained by running the classifiers within a validation set of images containing objects of the cate gory 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 op timization 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 sepa rability into account by introducing a slack variable ξ_i for each positive feature. T is is not necessary in the negative side, since the mean of the neg tive 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

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/945232121243011304