Generative regularization with latent topics for discriminative object recognition

Jose C. Rubio, Angela Eigenstetter, Björn Ommer

Heidelberg Collaboratory for Image Processing and Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany

ARTICLE INFO

Article history:
Received 3 January 2015
Received in revised form
1 June 2015
Accepted 23 June 2015

Keywords:
Visual object recognition
Part-based models
Non-Negative Matrix Factorization
Latent SVM
Mixture models

ABSTRACT

Popular part-based approaches to recognition are currently limited to few localized parts, which only poorly represent the fine-scale details and large variability of object categories. Extending to hundreds of specific part detectors helps to capture peculiar characteristics but due to their specificity, for each object instance different parts will be helpful and others will yield noisy responses that actually impair classification. While training the part-based model, we thus need to learn which parts are relevant for which training instances. To automatically discover these latent topics of parts and instances we employ generative non-negative matrix factorization and seek topics with low reconstruction error. To assure recognition performance this generative approach is embedded within a discriminative latent max-margin procedure that separates classes while optimizing the latent topics. Consequently, generative reconstruction is regularizing discriminative classification, while the latter ensures that topics actually help in recognition. Experiments on PASCAL VOC demonstrate the recognition performance of our model as well as the construction of meaningful topics.

1. Introduction

The complexity and large intra-class variability of present-day category-level object recognition datasets such as PASCAL VOC require representations that go beyond holistic models. Part-based representations [1–4] are the leading paradigm in the field because they effectively deal with articulation, occlusion, pose, and other variability. Currently very popular models are based on only a small number of informative parts (between 5 and 50) that describe the appearance of semi-local regions of an object category. To take into account category variability and the locality of parts, we instead sample a large number of parts (more than 1000 per category). Each part is specific to details of a local region in training instances, and trained with a single positive instance against negatives [5]. Consequently, different parts are helpful in different images. For instance, only some images of the *car* category feature a certain type of radiator grille, while others show radiator parts of vastly different shape and appearance, or no radiator parts at all.

The high *specificity* of parts helps to capture fine-scale details of the objects. However, for certain instances these parts may lead to noisy responses in the images where they are not meaningful. Therefore, when learning the part-based object model we need to learn which parts are relevant for which instances so that we can train with the right subset of parts on their corresponding samples. This goes beyond feature selection, where a fixed set of best parts for *all* instances is sought. In contrast, we aim to associate parts and instances to latent *topics*. These topics need to be learned automatically without relying on topic training annotations. The goal is then threefold: (i) discriminating positive instances of a category from negatives in order to detect objects amidst clutter, (ii) identifying which parts are meaningful for each instance and (iii) generating topics of instances and parts, which enable training models with the respective parts. These problems are directly inter-related and must be solved jointly. On one hand our overall setting is a discriminative classification problem (category-level recognition with labeled training bounding boxes). On the other hand, discovering topics of parts and instances so that the parts are meaningful for the respective instances is a generative scenario without labeled data for the topics.

We propose a latent max-margin approach that jointly solves the categorization problem and infers topics of parts and instances. The discriminative model assures competitive categorization performance, while the learning process is regularized with a generative model for the grouping, which seeks topics with optimal reconstruction performance. For regularization and topic formation we employ Non-Negative Matrix Factorization (NMF), which has been shown to be an effective unsupervised grouping procedure [6,7]. The NMF penalizes the reconstruction error of a low-rank decomposition of the training data. This decomposition naturally represents the strong coupling that arises when grouping instances and parts, and it is integrated in the latent

http://dx.doi.org/10.1016/j.patcog.2015.06.013
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

max-margin learning as latent variables that represent soft-assignments of parts and instances to topics. We demonstrate the performance of our model on the challenging PASCAL VOC dataset and show that jointly tackling categorization and the grouping of parts and images to topics is superior to commonly used part-based models that do not integrate grouping in the learning process.

2. Related work

A simple popular approach to grouping instances is clustering based on views or aspects. In [1] objects are clustered into three modes according to their aspect ratio. [8] exchanged the aspect ratio clustering by a k-means grouping with HOG features. Exemplar SVM (ESVM) was introduced in [5] where instead of grouping positive instances, each instance represents its own mode. While [5] use exemplar classifiers to describe whole objects, we use part classifiers trained on individual instances to describe specific local constituents of objects. In [2] and [9] exemplar part classifiers are employed in the context of recognition, where multiple parts are combined individually to form less-specific part classifiers. Instead, we retain a large number of specific parts and group them by learning the overall object model.

Other previous works aim to integrate the learning of aspects into a discriminative model [1,10], usually in a latent SVM framework. The work of [10] finds the modes of the positive data in two steps: grouping and then learning from the groups. In [11] the optimal number of mixture components is learnt using a group-sparsity inducing norm. An important difference of our approach compared to these works is that we model the grouping of instances as a continuous soft-assignment instead of a discrete labeling. Moreover, instead of implementing *a-posteriori* categorization as other works concerned with latent aspect modeling (i.e., max decision over all topics) we learn the weighted combination of topic classifiers.

In [6,7] Non-negative Matrix Factorization (NMF) has been shown to be an effective unsupervised grouping procedure and so we employ it for for regularization and topic formation. As shown in [12], NMF is equivalent to the probabilistic version of Latent Semantic Analysis (LSA) when the objective function is the KL-divergence. So in principle LSA, which much like NMF has been widely applied to grouping problems in vision, language processing, and beyond, is another grouping procedure that is also conceivable. However, the optimization of the KL-divergence is computationally burdensome and its probabilistic nature prevents us from relaxing the non-negativity constraints which, as we will see later, is a necessary step to improve the discriminativity of topics. Moreover, the LSA decomposition requires orthogonality on the discovered bases, while NMF does not impose this restriction. Our approach in contrast to LSA not only aims at creating meaningful topics, but it optimizes the recognition performance. In contrast to LSA we therefore follow a discriminative approach to separate different categories and regularize it using the generative model. Contrary to LSA we are thus explicitly optimizing the discriminativity of our representation.

NMF was originally proposed by [13], and Lee and Seung [6] highlighted its semantic decomposition properties. Since then, NMF has found wide applicability in various recognition tasks in computer vision, including face recognition [14], object recognition [15] and action recognition [16]. The benefits of combining generative and discriminative models into hybrid approaches have been pointed out in several works [17,18]. The integration of discriminative models with NMF has been investigated in [19,12], where the NMF objective is coupled with a SVM classifier. These approaches profit from the NMF decomposition to map features to a low-dimensional space that favors the separability of the data. However, we are not interested in performing feature learning but we aim to tighten the coupling between the NMF and a latent max-margin approach. Not only does NMF act as a regularizer for the classifier, but the elements of the decomposition themselves are modeled as latent variables that intervene in the learning process.

3. Method

In this section we present a latent max-margin model that jointly infers groupings of parts and instances and learns to discriminate between positive and negative examples. First we introduce the part-based representation on which the rest of the method is based. Then we review the NMF formulation and present the generative component. Finally we introduce the joint model and discuss the details of the training process.

3.1. Part-based object representation

We represent an image as a set of responses from part classifiers, each specifically trained on a local region of one training image. We start by randomly sampling M squared regions from positive training images of a category at different locations and sizes. Then we train M region classifiers φ_k in an Exemplar SVM (ESVM) fashion [5] using HOG features extracted from each of those local patches. As in [5] a single positive is trained against a large set of hard negatives mined from training images. For each training sample I_i, each φ_k is evaluated densely on the image I_i. Then, I_i is divided into a regular 4×4 grid and we retain only the maximum of φ_k in each of those 16 cells. Concatenating the $D = 16M$ part classifier responses yields the object representation $\boldsymbol{\phi}(I_i) \in \mathbb{R}^D$. We abbreviate this vector of all localized part responses by $\boldsymbol{\phi}_i$. The goal is then to find a classifier $f(\boldsymbol{\phi}_i) = \mathbf{w}^\top \boldsymbol{\phi}_i$ trained on those part-based representations that is able to distinguish between positive and negative objects.

3.2. Latent topics of parts and instances

For a training sample we seek its subset of meaningful parts and to train the category classifier on a subset of parts we need a set of training samples where those parts are meaningful. Thus, the problems of finding meaningful parts and grouping training samples into *aspects* or *topics* that share the same parts are directly coupled. Since no training annotations are provided for identifying relevant parts or for the groups of instances they occur in, we need to tackle both problems jointly. Subsequently, we present an approach that jointly assigns parts to topics, assigns training instances to the topics, and trains discriminative classifiers for the topics that resolve the original categorization problem.

Non-negative matrix factorization minimizes the reconstruction error of the original representations $\boldsymbol{\phi}_i$ by a decomposition into topics. Given a set T^+ of N positive training samples, let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N)^\top \in \mathbb{R}^{D \times N}$ denote the matrix of all part-based representations. The NMF aims to find a low-rank decomposition of the data matrix $\boldsymbol{\phi}$ into two matrices $\mathbf{G} \in \mathcal{R}^{D \times K}$, and $\mathbf{H} \in \mathcal{R}^{K \times N}$. The K columns of the matrix $\mathbf{G}$ are basis vectors that can be interpreted as the groups of relevant parts of each of the K topics. The dimensions of the basis vectors with high values indicate a strong contribution of localized parts, while zero elements indicate non-contributing parts. By taking all non-zero elements of the basis j we identify a group of parts that are relevant for the topic j. Analogously, the coefficients of $\mathbf{H}$ serve as indicators of the extent to which an image is represented by each of the K topics, and

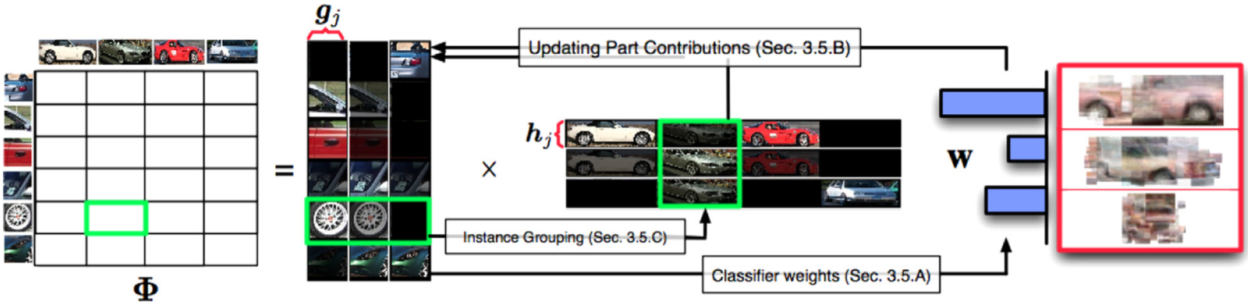
$\boldsymbol{g}_j$
Updating Part Contributions (Sec. 3.5.B)
=
×
$\boldsymbol{h}_j$
$\mathbf{w}$
Instance Grouping (Sec. 3.5.C)
Classifier weights (Sec. 3.5.A)
$\boldsymbol{\Phi}$

Fig. 1. General scheme of the approach. Data matrix $\boldsymbol{\phi}$ decomposes in the matrix of part groups **G** and the matrix **H** of image groups. Black or shaded areas correspond to matrix elements with low values. The arrows show the optimization steps, with references to the corresponding subsections of Section 3.5. The green boxes show the relationship between an element on the original matrix and its corresponding row and column in **G** and **H**. The histogram in the right side represents the topic classifier weights **w** for 3 topics, and the red box show visualizations of the part-based models of each topic. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

therefore provide a soft assignment of instances to topics. Formally,

$$\min_{G,H} \|\boldsymbol{\phi} - \mathbf{GH}\|_F^2, \quad \mathbf{G} \geq 0, \ \ \mathbf{H} \geq 0 \qquad (1)$$

where the Frobenius norm is used to measure the reconstruction error. We can interpret the columns of **G** as topic classifiers. By discovering those K topic classifiers $\boldsymbol{g}_j$ we are dividing our original classification problem into K simpler sub-problems $f_j(\boldsymbol{\phi}) = \boldsymbol{g}_j^\top \boldsymbol{\phi}$ that only focus on the parts relevant for the respective topic j.

Retrieval phase: Now, to classify a new test sample I we need to project its $\boldsymbol{\phi}(I)$ first into the space of the part-topics, and linearly combine the contributions of each topic classifier in a final decision $f(\boldsymbol{\phi}) = \mathbf{w}^\top \mathbf{G}^\top \boldsymbol{\phi}$. Note that compared to the D-dimensional original classifier weights, the weight vector **w** now has K dimensions, since samples are now categorized in the space of part-topic classification scores f_j instead of all parts $\boldsymbol{\phi}$.

3.3. *Combined discriminative and generative approach to recognition and reconstruction*

The topic classifiers $\boldsymbol{g}_j$ in the previous section have been obtained using a generative approach that aims at an optimal reconstruction of latent topics of positive training samples. However, for the final problem of classifying image regions and detecting objects we are not only interested in these generative abilities but also in $\boldsymbol{g}_j$ classifiers that discriminate objects from clutter. Therefore we now combine the generative model from (1) with a discriminative approach. However, this is not merely a discriminative matrix decomposition. The groupings of parts and instances inferred by the generative process should have an active role in the learning of topic classifiers and of the overall classifier f that combines all topics. We propose a latent max-margin approach that jointly learns the matrix decomposition, the topic classifiers and the final overall classifier. The generative component supports topic discovery. Moreover, by ensuring that topics can actually reconstruct the original representation, it also regularizes the discriminative classifier, which itself enables object recognition. Fig. 1 shows an overview of the approach.

Latent SVM: Let us first review the general approach of Latent Support Vector Machines (LSVM). Assume that a set T of training instances is provided, where each instance I_i is described by a feature vector $\boldsymbol{\phi}_i$, and has corresponding label $y_i = \{1, -1\}$ indicating if it belongs to a particular category. In the general setting, the scoring of the LSVM has the form $f(\boldsymbol{\phi}) = \max_{h \in \mathcal{H}} \mathbf{w}^\top \psi(\boldsymbol{\phi}, h)$, where the joint feature vector ψ depends on the data point $\boldsymbol{\phi}$ and the latent variables h. The learning problem is then formulated as

$$\operatorname*{argmin}_{\mathbf{w},\xi} \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + \sum_{i \in T} \xi_i \qquad (2)$$

$$\text{s.t} \quad y_i f(\boldsymbol{\phi}_i) \geq 1 - \xi_i, \quad \forall i \in T \qquad (3)$$

$$\xi \geq 0. \qquad (4)$$

From an optimization perspective, the usual procedure is to alternate between finding the best parameters **w** given an assignment h of latent variables, and inferring the best assignment of latent variables given the current state of **w**.

Joint model: Let us now incorporate the generative NMF into the latent max-margin approach for discriminative classifier training. In the latent formulation part classifier parameters **G**, assignments of instances to topics **H** and the overall classifier weights **w** are jointly optimized. The model is formulated as follows:

$$\arg\min_{\mathbf{G},\mathbf{H},\mathbf{w},\xi,\zeta} \frac{\lambda_1}{2} \|\mathbf{w}\|_2^2 + \sum_{i \in T} \xi_i + C \sum_{i \in T} \sum_{j=1}^{K} \hat{h}_{i,j} \zeta_{i,j} + \frac{\lambda_2}{2} \|\boldsymbol{\phi} - \mathbf{GH}\|_F^2 \qquad (5)$$

$$\text{s.t} \quad y_i (\mathbf{w}^\top \mathbf{G}^\top \boldsymbol{\phi}_i) \geq 1 - \xi_i, \quad \forall i \qquad (6)$$

$$y_i \boldsymbol{g}_j^\top \boldsymbol{\phi}_i \geq 1 - \zeta_{i,j}, \quad \forall i, j \qquad (7)$$

$$\mathbf{H} \geq 0, \zeta_{i,j} > 0, \xi_i > 0, \quad \forall i, j \qquad (8)$$

$$\mathbf{G} \geq 0 \qquad (9)$$

where $\hat{h}_{i,j} = h_{i,j} \quad \forall i \in T^+$, and $\hat{h}_{i,j} = 1 \quad \forall i \in T^-$. This distinction between positive samples in T^+ and negative samples in T^- is necessary because the entries $h_{i,j}$ of the matrix **H** are only defined for positive samples. Recall from Section 3.2 that negative samples are not related to any topic. The constraint (6) assures discriminative power in the original categorization problem, and the constraint (7) enforces individual topic classifiers to be discriminative. The latent assignments of instances to topics determine during training which samples are selected to train each of the part-topics $\boldsymbol{g}_j$. To achieve this, we introduce slack variables $\zeta_{i,j}$ which are weighted according the coefficients of **H**: if the positive instance i has a strong affinity with topic j, then the matrix entry $h_{i,j}$ has a high value, and the topic classifier $\boldsymbol{g}_j$ gets penalized when misclassifying the instance i (since $\sum_{i,j} \hat{h}_{i,j} \zeta_{i,j}$ is minimized and $y_i \boldsymbol{g}_j^\top \boldsymbol{\phi}_i \geq 1 - \zeta_{i,j}$ with positive $\zeta_{i,j}$). Contrarily, low values of $h_{i,j}$ allow the classifiers to ignore the positive training samples that do not belong to the topic. The negative samples always contribute with a cost when misclassified, given that their corresponding $\hat{h}_{i,j}$ is fixed to 1. The parameter λ_1 regularizes the weights of the overall classifier, while the parameter λ_2 regularizes the topic classifiers. The parameter C indicates the importance of fulfilling the topic-classifier loss against the overall classifier loss.

The number of topics should be sufficiently large to adequately represent the variability of category instances and their parts. In particular, we need a finer granularity than the few, generic views of DPM [1]. We set the initial number of topics to be $1/15$ of the number of positive training instances and let the overall classifier boost or suppress topics acoording to their importance. Note that topic classifiers $\mathbf{g}_j$ that fail at categorizing training samples correctly penalize their corresponding coefficients in $\mathbf{H}$ due to the margin violations expressed by the slacks ζ. Therefore, topic classifiers which repel instances lose most of their training data and eventually converge to very few outlier images. Consequently the overall classifier $\mathbf{w}$ will assign to those topics a near-zero weight, thus minimizing their influence on the decision function, and achieving a data-driven adaptation of topics utilized per category.

3.4. Constraints on the topic classifiers

So far we have followed the rationale of NMF and constrained $\mathbf{G}$ to be strictly positive. Thus, parts with high weight in $\mathbf{G}$ contribute to the presence of an object. Conversely, parts with low weight do not help in deciding about the presence/absence of an object. However, in the discriminative setting parts should also be able to vote for the *absence* of objects.

To let parts contribute negatively and thus gain discriminative power, we relax the NMF factorization by dropping the non-negativity constraint on $\mathbf{G}$ (Eq. (9)) retaining the non-negativity on $\mathbf{H}$ of assignments from instances to topics. This increases the discriminative power of the approach, but at first sight one might fear that this is at the cost of reducing the interpretability of the basis vectors $\mathbf{g}_j$. However, performing the matrix decomposition while allowing negative values in the matrix $\mathbf{G}$ is directly related to clustering, as shown in [20]. The columns of $\mathbf{G}$ denote cluster centroids and the rows of $\mathbf{H}$ are the assignments of data points to clusters. Thus the approach is grouping instances, relating them to parts, and discovering common topics, while trading classification performance against the reconstruction error.

Let us now consider other potential constraint relaxations. We started our implementation of the approach by directly using the non-negative bases as classifiers (non-negativity constraint on $\mathbf{G}$ and $\mathbf{H}$). The results were not competitive (around 20% drop in accuracy), since constraining the classifier weights means limiting the classifier model space (the hyperplane cannot be arbitrarily aligned) and thus restricting the ability to separate the data. Regarding dropping non-negativity of $\mathbf{H}$, since the matrix $\mathbf{H}$ expresses the degree to which training samples belong to classifiers, negative values in $\mathbf{H}$ would not make sense and need to be prevented.

3.5. Training

The optimization of the objective in (5) is a non-convex problem due to the coupling between the unknowns $\mathbf{w}$, $\mathbf{G}$ and $\mathbf{H}$. We approximate its solution by alternating three convex optimization problems.

(A) *Contributions of topics to the category* ($\mathbf{w}$): First we initialize the matrices $\mathbf{G}$ and $\mathbf{H}$. The strategy used for initialization is discussed in detail in Section 4.1. With both $\mathbf{G}$ and $\mathbf{H}$ fixed we solve for the parameters of the general classifier $\mathbf{w}$ using a support vector machine in the topic space by projecting features into topics: $G^\top \boldsymbol{\phi}$.

(B) *Updating Part Contributions* ($\mathbf{G}$): With $\mathbf{w}$ and $\mathbf{H}$ fixed, we can denote the objective over the elements of the matrix $\mathbf{G}$ (the topic classifiers) as follows:

$$\arg\min_{\mathbf{G}} \frac{\lambda_1}{2} \| \mathbf{w} \|^2 + \sum_{i \in T} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{G}^\top \boldsymbol{\phi}_i)) + C \sum_{i \in T} \sum_{j=1}^{K} \hat{h}_{i,j} \max(0, 1 - y_i \mathbf{g}_j^\top \boldsymbol{\phi}_i) \tag{10}$$

$$+ \frac{\lambda_2}{2} \| \boldsymbol{\phi} - \mathbf{GH} \|_F^2 \tag{11}$$

We optimize the previous objective using a stochastic sub-gradient descent (SGD) algorithm in the primal. As in standard SVM we use the hinge loss as a surrogate, which is convex and sub-differentiable. Following common practice, we compute the sub-gradient with respect to the model parameters,

$$\nabla_{\mathbf{G}} = -\sum_{i \in A} y_i \sum_{j}^{K} \mathbf{w}_j \boldsymbol{\phi}_i - C \sum_{\{i,j\} \in B} \hat{h}_{i,j} y_i \boldsymbol{\phi}_i - \lambda_2 (\boldsymbol{\phi} \mathbf{H}^\top - \mathbf{GH}^\top)$$
$$A = i \,|\, y_i(\mathbf{w}^\top \mathbf{G}^\top \boldsymbol{\phi}_i + b) < 1, \quad B = \{i,j\} \,|\, y_i \mathbf{g}_j^\top \boldsymbol{\phi}_i < 1,$$
$$i \in T, \ \ 1 \le j \le K \tag{12}$$

The sets A and B define the domain where the objective function is differentiable. In practice, during optimization the SGD assigns a zero gradient to any incoming sample that does not belong to sets A or B.

(C) *Updating instance groupings* ($\mathbf{H}$): This step consists of determining the optimal soft-assignment of positive instances to topics given the current state of the topic classifiers $\mathbf{G}$. If a positive image is inside the margin of a given topic-classifier j, that is $\mathbf{g}_j^\top \boldsymbol{\phi}_i < 1$, then the latent variable $h_{i,j}$ is penalized with a cost proportional to the degree of margin violation of the training sample. Samples correctly classified will induce zero cost on their corresponding latent variables. Eqs. (10) and (11) comprise a linear and a quadratic term respectively over the variables $h_{i,j}$. Given the matrix of topic classifiers $\mathbf{G}$ computed in Section 3.5. B, inferring the latent values of $\mathbf{H}$ amounts to solving a standard quadratic program. After some algebraic manipulation the objective can be compactly defined as

$$\arg\min_{\mathbf{h}} \frac{1}{2} \mathbf{h}^\top (\lambda_2 \boldsymbol{I} \otimes \mathbf{G}^\top \mathbf{G}) \mathbf{h} + vec(\boldsymbol{E} - \lambda_2 \mathbf{G}^\top \boldsymbol{\phi}) \mathbf{h}, \tag{13}$$

where $\mathbf{h} > 0$. The term $\boldsymbol{E}(i,j)$ denotes the matrix of $C \max (0, 1 - y_i \mathbf{g}_j^\top \boldsymbol{\phi}_i) \quad \forall i,j$, $\boldsymbol{I}$ is the identity matrix of size $|T^+|$, $\otimes$ is the Kronecker product, and $vec(\cdot)$ is an operator that stacks the columns of a matrix above another.

4. Experiments

We utilize the challenging PASCAL VOC 2007 [21] benchmark dataset to evaluate the proposed discriminative recognition approach with generative regularization based on latent topics. Images of VOC exhibit high intra-class variability in terms of visual appearance, object deformation and pose. The dataset contains more than 12,000 objects for training which are divided in 20 classes. We train our models on training and validation set and show Average Precision (AP) results on the test set. The AP is given by the area under the precision/recall curve. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class.

4.1. Experimental set-up

We follow the standard protocol and train our model on training and validation set and evaluate performance in the object detection challenge on novel images from the test set (measured with the usual 50% PASCAL overlap criterion for AP). Object models are trained on the positive object bounding boxes and using a set of candidate regions extracted with DPM [1] (using a conservative, low threshold of -1.1) from the background as hard negatives. We are only using these boxes, but no ranking information or DPM scores. In each candidate window we extract the part-based representation from Section 3.1 (1000 parts for a category) and compute $\boldsymbol{\phi}$. We follow the same setup during testing to compute $\boldsymbol{\phi}$ for all candidates. Parameters λ_1, λ_2 and C are obtained by cross-validation on the training data. As it is common practice, the step size of the stochastic gradient descent (SGD) algorithm is set to decrease each iteration as $\eta_t = 1/\lambda_2 t$. The number of *epochs* of the SGD is set to 5 and the number of iterations of the overall learning process is set to 40. To initialize **G** and **H** and start the learning from Section 3.5 we first assign random values and run 10 rounds of NMF optimization (Eq. (1)) by applying the multiplicative update rules of [7]. This initialization has shown better performance than initializing **H** with k-means clustering on $\boldsymbol{\phi}$ ($h_{ij} = 1$ if instance i belongs to cluster j) and training a topic model to initialize **G** with its weights. This latter strategy boosts the performance very early in the learning process but ends in local minima without reaching significant gain over the baseline.

4.2. *Topics of parts and instances*

Our model not only addresses the categorization problem but also infers latent topics of parts and instances. Fig. 2(A) shows 75 instances of the *bicycle* category picked at random from the most important 6 topics (with highest bfw). The columns of **H** serve as a low-dimensional representation of the images by capturing their degree of correspondence to each of the topics. We compute a 2-dimensional embedding from that k-dimensional representation using the Isomap algorithm to produce the scatter plot, and assign each instance to the topic with maximum value according to matrix **H**. Fig. 2(B) shows a single prototypical instance per topic that has smallest average distance over **H** to all other instances in the topic. Fig. 2(C) and (D) shows visual representations of instance topics and part topics respectively, computed by running the part filters of each topic over all images of that same topic and averaging all detections. The topics discovered by the algorithm reveal subtle conceptual information beyond a mere aspect ratio or view clustering. For instance, the red group represents bikes with no person riding them, while the green group contains bicycle riders. The yellow set groups bikes that are tilted. The blue set shows frontal racing bikes, while the pink set gathers frontal street or mountain-bikes. The cyan group covers rare instances like kid's bikes and cropped bicycle parts like handlebars. As expected, the visual representation of this outlier topic looks cluttered and structureless due to the heterogeneity of the instances belonging to the topic. Examples of topic visualizations of other categories are presented in Figs. 7 and 8.

The optimum number of topics varies per category. With categories like *potted plant* or *sheep* is hard to discover multiple fine-grained groups that improve overall categorization, and therefore such categories can be represented by a small number of topics. However, categories like *car* or *aeroplane* are favoured by a large number of topics. In order to avoid hand-picking the number of topics per category we follow the strategy of choosing a sufficiently large number of topics, proportional $(1/15)$ to the number of positive training samples available. During training, if a category suffers from over-clustering, the weaker topic classifiers will loose training samples until remaining with very few outliers. These classifiers will be assigned with a low weight by the general classifier and therefore adapt dynamically to the optimal number of groups. Fig. 3 analyzes the impact of the initial number of topics on performance. Accuracy saturates at about $1/15$ of the number of positive training samples with minor fluctuations afterwards. This suggests to simply select a sufficiently large number of initial topics (e.g., proportional to $1/15$ of the number of positive training samples) so that the proposed algorithm can afterwards automatically infer relevant topics.

4.3. *Convergence of the optimization*

In Fig. 4 (left) we empirically confirm the convergence of the optimization from Section 3.5. The red curve shows that overall cost is decreasing rapidly in the first iterations. The black curve gives the area under the precision/recall curve (AUC) after

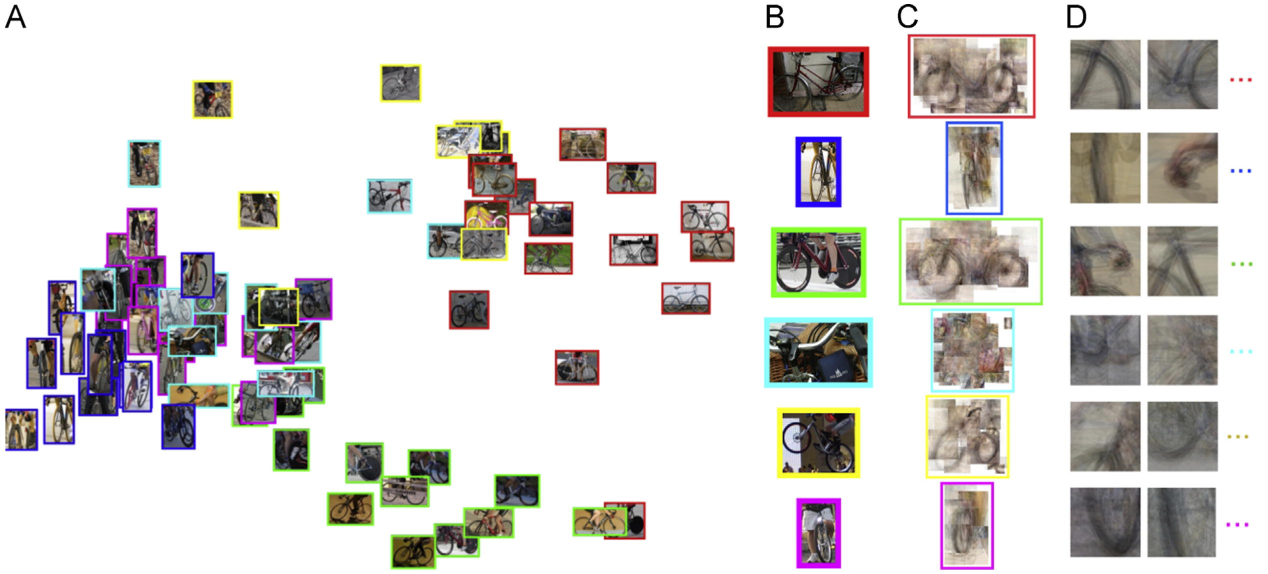

Fig. 2. (A) Scatter plot illustrating the instance grouping. Different colors denote different topics. (B) shows one prototypical representative for each topic to provide a more concise overview. (C) instance topics: images with high values in each row of the instance grouping matrix **H**. (D) Examples of the corresponding part topics, each of them defined by those parts with highest values in each column of part grouping matrix **G**. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

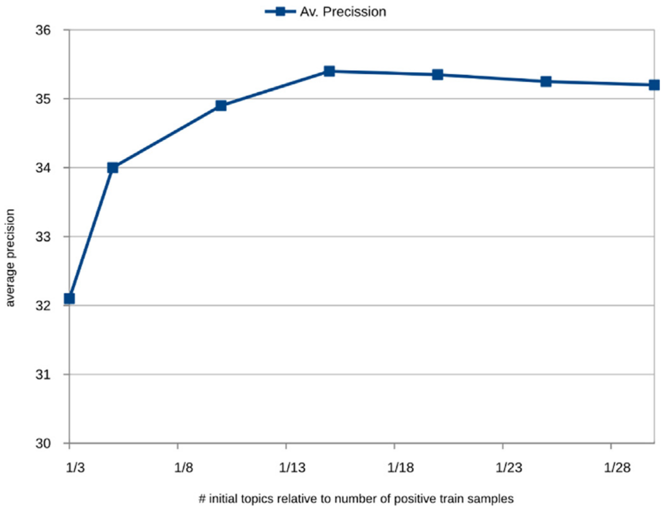

Fig. 3. Variation of the mean average precision in PASCAL VOC depending on the number of initial topics (relative to the number of positive training samples of the respective object category). When the number of initial topics is sufficiently large, e.g., around 1/15 of the number of positive instances, the precision saturates, as the proposed approach can automatically discover relevant topics and discard irrelevant ones.

applying the classifier on the test data. In order to evaluate the contribution of the NMF regularization we repeat the experiment by removing $\mathbf{G}$ from the NMF regularization in Eq. (5) and instead regularize $\mathbf{G}$ using L-2 Norm, $\sum_j^K \|\mathbf{g}_j\|_F^2$. When removing $\mathbf{G}$ from the NMF term, $\mathbf{H}$ might become unbounded. Thus we insert a surrogate $\mathbf{G}_2$ in $\|\Phi - \mathbf{G}_2\mathbf{H}\|_F^2$, which is obtained by solving Eq. (1) once at initialization, and keep $\mathbf{G}_2$ fixed afterwards throughout training. We denote this baseline as BS-nmf. The convergence results after removing the regularization are presented in Fig. 4 (right), showing a significant drop of performance (8%) as well as a slower convergence of the cost function. This suggests that the topic classifiers are more likely to end in bad local minima with L-2 instead of NMF regularization.

The convergence speed depends on the number of training samples N and the number of topics K. The time complexity is $O(NK)$, but since N is much larger than K, the number of training samples dominates over the number of topics.

4.4. *Results on PASCAL VOC*

Table 1 shows results on the PASCAL 2007 dataset. On the top part of the table we compare a baseline version of our approach

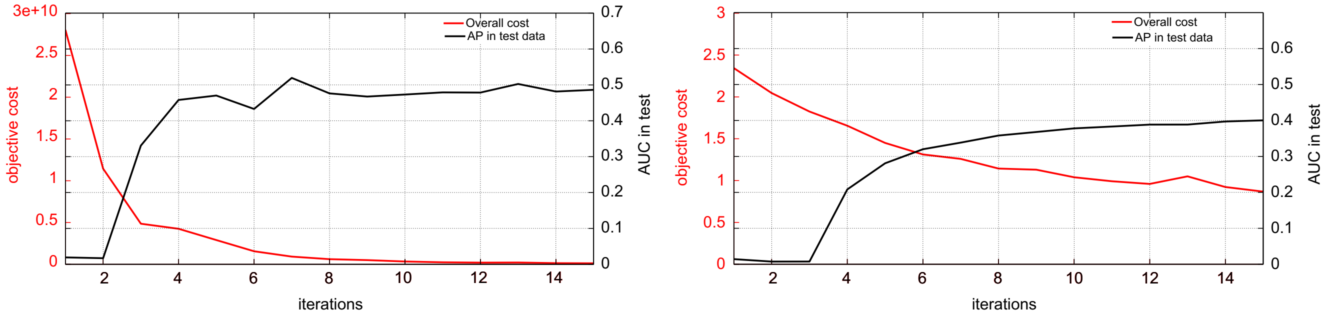

Fig. 4. Left: convergence of the cost function against area under precision/recall curve (AUC) for the class *person*. Right: convergence and AUC of BS-nmf. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
Test results on PASCAL VOC 2007. The BS-nmf column provides baseline results of running our part-based model without NMF regularization (see Section 4.3). In BS-km we pre-cluster instances using *k*-means and train classifiers per each group (using all parts). TCa shows the results of our approach without DPM filters. TCb includes DPM filters. KLSO refers to the work of [10] on mixture learning. GN is the Group Norm learning of [11], and AOT refers to the And-Or-Tree models of [4].

Category	BS-km	BS-nmf	TCa	TCb	DPM	KLSO	GN	AOT
Aeroplane	36.1	36.5	37.2	**37.8**	33.2	33.3	33.6	35.3
Bicycle	58.5	60.9	63.2	**63.2**	60.3	53.6	57.6	60.2
Bird	5.9	6.8	9.5	10.1	10.2	9.6	9.4	**11.0**
Boat	13.1	13.0	15.2	15.1	16.1	15.6	15.5	**16.6**
Bottle	21.2	22.7	22.0	22.3	27.3	22.9	28.9	**29.5**
Bus	52.1	53.2	59.1	**59.2**	54.3	48.8	51.7	53.0
Car	58.0	58.6	59.6	**59.7**	58.2	51.5	55.3	57.1
Cat	23.7	24.4	26.1	**26.4**	23.0	16.3	20.2	23.0
Chair	19.1	20.2	20.5	20.6	20.0	16.3	22.1	**22.9**
Cow	24.1	24.5	26.8	27.4	24.1	20.0	**30.4**	27.2
Table	28.7	29.0	29.8	**29.9**	26.7	23.8	28.9	28.6
Dog	17.5	17.2	19.3	**19.5**	12.7	11.0	11.5	13.1
Horse	52.4	53.2	54.7	55.1	58.1	55.3	58.1	**58.9**
Motorbike	47.2	48.8	52.5	**53.1**	48.2	43.8	46.4	49.9
Person	39.7	40.0	41.3	**47.7**	43.2	36.9	38.8	41.4
Plant	11.5	11.9	12.8	13.2	12.0	10.7	14.1	**16.0**
Sheep	20.3	20.7	20.8	21.5	21.1	22.7	16.2	**22.4**
Sofa	36.3	36.9	39.2	**39.8**	36.1	23.5	32.3	37.2
Train	45.0	45.7	47.8	47.4	46.0	38.6	45.6	**48.5**
Tvmoitor	42.1	43.5	45.1	**45.2**	43.5	41.0	43.8	42.4
Mean	32.5	33.4	35.1	**35.4**	33.7	29.8	33.0	34.7

without topics with our approach with Topic Classifiers (TC) in two different set-ups. In TCa we do not include the DPM parts (only the randomly sampled 16 K localized part classifiers), while in TCb we include the strong filters trained by the DPM model. Adding DPM filters does not alter significantly the results (+0.3% AP).

The bottom part of Table 1 compares with DPM and three recent works that also deal with instance grouping and part-based models. In some of the categories the topics do not necessarily play an important role (e.g bottle), cases in which the gain against the monolithic baseline is not compelling. Moreover, categories in which parts are not very meaningful, or which are very often occluded or cluttered (e.g potted plant) our structured topics of parts are outperformed. We clearly improve upon the results of [10], which also tackles the problem of latent learning of groups of positive instances. In comparison to [4] our model shows superior performance, improving on 11 categories.

We have focused our comparison on other part-based models that deal with the problem of grouping (either instances or parts). There are other methods that without dealing explicitly with the grouping problem also provide state-of-the-art results, as is the case of [22] which holds the best performance in PASCAL VOC (mAP 41.7). However, the results reported in [22] vary significantly depending on which types of features used. When using HoG as we and the majority of approaches on this challenge, they report a mAP of 35.1%, which is in line with our performance (35.4%). Some detection results can be seen in Fig. 5, together with the visual representation of the topic classifier that contributed the most to the detection score.

Regarding the computational effort, our most expensive step is to compute the responses for the 1000 part classifiers. However, due to the linear nature of the classifiers, evaluation for such a large number of parts is very efficient, taking around 13 s for a thousand parts. Actually, most of the computation time is spent on HOG feature computation and extracting sliding-windows, which is independent of the number of parts.

4.5. *Results on scene classification*

We report results on Scene Classification using the MIT indoor database introduced in [26]. We provide results in terms of average precision, and compare our performance to 7 other approaches (see Table 2). Our method in its standalone set-up (TC) shows an accuracy on line with the state-of-the-art. If we focus our comparison on approaches that rely solely on HOG features and part-based representations, we outperform mid-level-patches [25] as well as BoP [9] by 11% and 5.6% respectively. In this experiment, given that the amount of positive samples per category is significantly lower than in PASCAL VOC (around 80 training positive images per category), we fix the number of groups $K = 3$ in all categories. Having a larger number of groups

Table 2
Scene categorization results on MIT indoor dataset.

Method	Av. precision
Object Bank [23]	37.6
RBoW [18]	37.9
DPM+GIST-color+SP [24]	43.1
Patches+GIST+SP+DPM [24]	49.4
Mid-Level Patches [25]	38.1
BoP [9]	43.5
IFV+BoP [9]	63.1
Ours (baseline)	46.5
Ours (baseline – nmf)	44.2
Ours (TC)	49.1
Ours (TC+IFV)	61.3

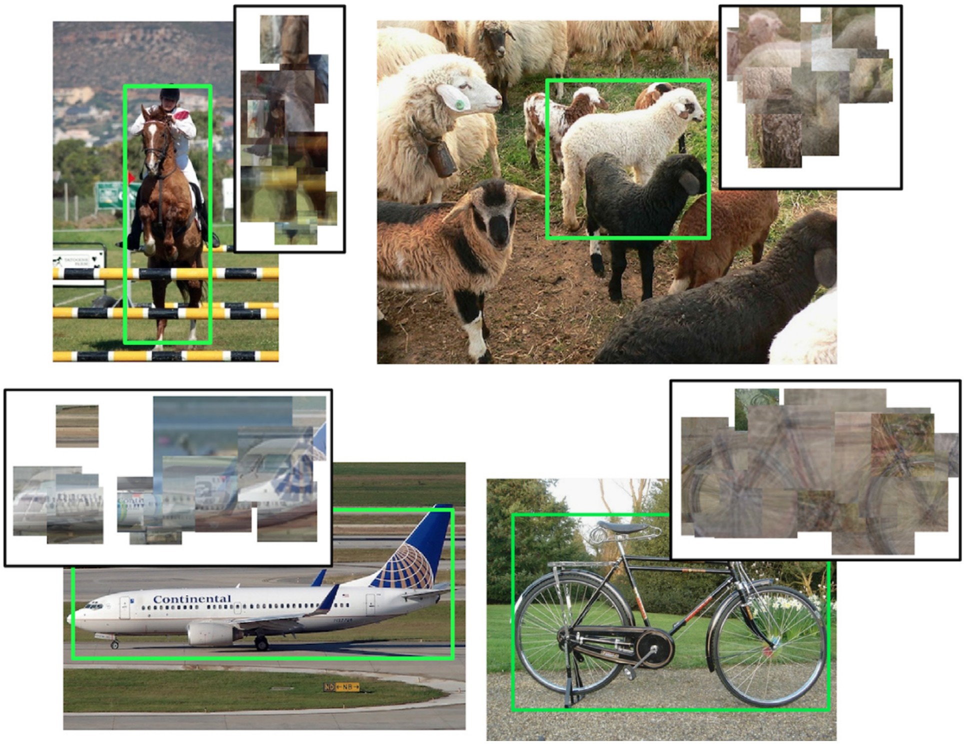

Fig. 5. Example detections (green boxes) of categories *horse, sheep, airplane and bicycle*. The black box at the upper corner shows a visual representation of the topic classifier that scored the highest for that object instance. The representations are constructed by averaging part detections across the whole topic of instances, and placed in the highest scoring locations of the test image. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

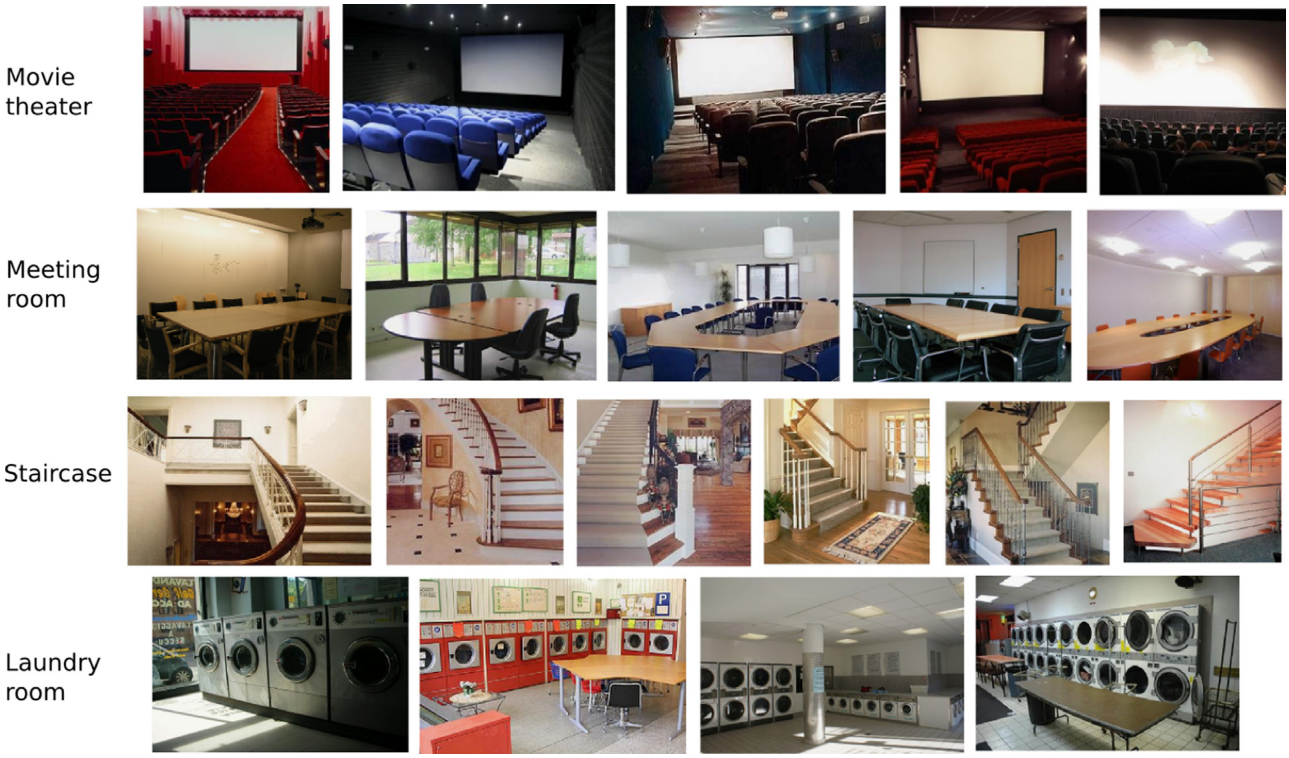

Fig. 6. Classification results on the indoor MIT dataset. For each category we show examples of the top ranked testing instances.

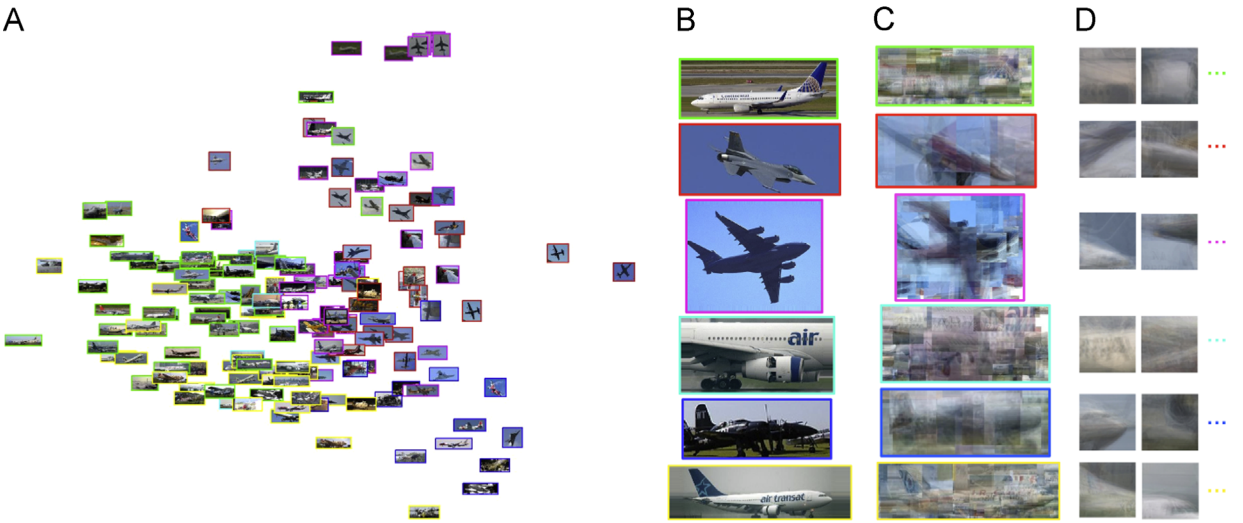

Fig. 7. Visualization of topics of the airplane category. The decomposition of instances distinguishes planes flying (red and pink groups) from planes landed in an airport (green and yellow groups). The cyan group gathers planes that are occluded. Other types of semantics can also be spotted: the red group tends to gather modern war planes with sharp edges, while the pink group covers planes with soft shapes such as seaplanes or old-school war planes. a) Scatter plot of grouped instances, b) Topics: prototypical representative, c) Instance topics and d) Part topics.(For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

hampers the quality of the topic classifiers due to an insufficient amount of positive training samples, ultimately resulting in a lower average precision of the joint model. The performance boost obtained by combining our model with Fisher Vectors is of 11%, significantly lower than the one reported by BoP (19%). Note that we are using our own implementation of Fisher Vectors and that in [9] there has been additional fine tuning of the parameters of this postprocessing stage, which is, however unrelated to our actual contribution. Fig. 6 shows examples of the top ranked testing images for different categories.

5. Conclusions

In this work we have shown that jointly grouping parts and instances is a beneficial avenue when dealing with high visual variability in the context of object recognition. Part-based models provide an excellent framework to tackle this problem, especially if the parts are very specific and a large number of them are available. The problems of grouping instances and parts into topics present two main challenges: first, the strong mutual inter-dependency between them. Second, the high specificity of the parts, that entails noisy

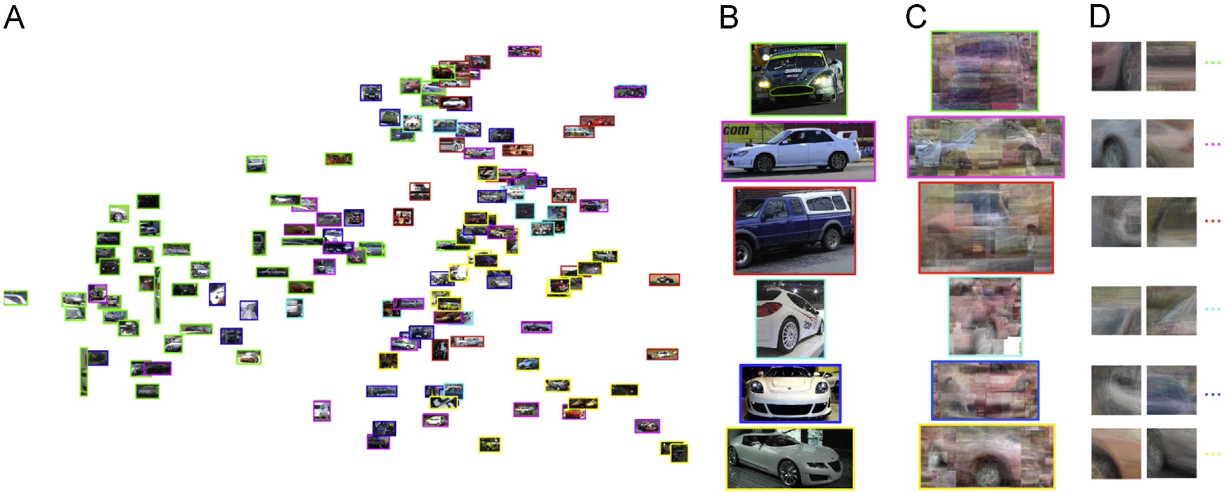

Fig. 8. Visualization of topics of the car category. Red and pink groups cover cars seen from the side, while green and blue gather cars seen from the front/back. The yellow topic gathers cars with tilted perspective. The cyan group in this case gets the outlier instances, that happen to be either occluded cars or weird examples like Formula1 cars or vintage old cars. a) Scatter plot of grouped instances, b) Topics: prototypical representative, c) Instance topics and d) Part topics.(For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

observations when inferring specialized topics of instances. We alleviate both problems using a latent max-margin classifier that jointly solves the categorization problem and the grouping problem. A generative regularizer (NMF) guides the grouping process, while the discriminative (Max-Margin classifier) contributes to categorization task. Our results in the PASCAL VOC 2007 dataset show that the grouping of parts and instances significantly improves the performance of a monolithic detector. We provide results comparable with the state-of-the-art, as well as showing that the generated grouping of parts and instances is semantically meaningful.

Conflict of interest

None declared.

References

[1] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, in: PAMI, 2010, pp. 1627–1645.
[2] I. Endres, K.J. Shih, J. Jiaa, D. Hoiem, Learning collections of part models for object recognition, in: CVPR, 2013, pp. 939–946.
[3] L.L. Zhu, et al., Latent hierarchical structural learning for object detection, in: CVPR, 2010, pp. 1062–1069.
[4] Y.J. Xi Song, Tianfu Wu, S.-C. Zhu, Discriminatively trained and-or tree models for object detection, in: CVPR, 2013, pp. 3278–3285.
[5] T. Malisiewicz, A. Gupty, A.A. Efros, Ensemble of exemplar-svms for object detection and beyond, in: ICCV, 2011, pp. 89–96.
[6] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization., in: Nature, 1999, pp. 788–791.
[7] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: NIPS, 2000, pp. 556–562.
[8] S.K. Divvala, A.A. Efrox, M. Hebert, How important are “deformable parts” in the deformable parts model ?, in: ECCV Parts and Attributes Workshop, 2012, pp. 31–40.
[9] M. Juneja, A. Vedaldi, C.V. Jawahar, A. Zisserman, Blocks that shout: distinctive parts for scene classification., in: CVPR, 2013, pp. 923–930.
[10] O. Aghaszadeh, H.A.J. Sullivan, S. Carlsson, Mixture component identification and learning for visual recognition, in: ECCV, 2012, pp. 115–128.
[11] D. Chen, D. Batra, W.T. Freeman, Group norm for learning structured svms with unstructured latent variables, in: ICCV, 2013, pp. 409–416.
[12] B.G.V. Kumar, I. Kotsia, I. Patras, Max-margin non-negative matrix factorization, in: Image and Vision Computing, 2012, pp. 279–291.
[13] P. Paatero, U. Tapper, Positive matrix factorization: a non-negative factor model optimal utilization of error estimates of data values, in: Environmetrics, 1994, pp. 111–126.
[14] S.Z. Li, X.W. Hou, H.J. Zhang, Q.S. Cheng, Learning spatially localized part-based representations, in: CVPR, vol. 1, 2001, pp. I-207–I-212.
[15] W. Liu, N. Zhen, Non-negative matrix factorization based methods for object recognition, in: Pattern Recognition Letters, 2004, pp. 893–897.
[16] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: CVPR, 2008, pp. 1–8.
[17] A.D. Holub, M. Welling, P. Perona, Combining generative models and fisher kernels for object class recognition, in: ICCV, vol. 1, 2005, pp. 136–143.
[18] S.N. Parizi, J.G. Oberlin, P.F. Felzenszwalb, Reconfigurable models for scene recognition, in: CVPR, 2012, pp. 2775–2782.
[19] M.D. Gupta, J. Xiao, Non-negative matrix factorization as a feature selection tool for maximum margin classifiers, in: CVPR, 2011, pp. 2841–2848.
[20] C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, in: PAMI, 2010, pp. 45–55.
[21] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, in: International Journal of Computer Vision, 2010, pp. 303–338.
[22] X. Wang, M. Yang, S. Zhu, Y. Lin, Regionlets for generic object detection, in: ICCV, 2013, pp. 17–24.
[23] L. Jia Li, H. Su, L. Fei-fei, E.P. Xing, Object bank: a high-level image representation for scene classification and semantic feature sparsification, in: NIPS, 2010, pp. 1378–1386.
[24] M. Pandey, S. Lazebnik, Scene recognition and weakly supervised object localization with deformable part-based models, in: ICCV, 2011, pp. 1307–1314.
[25] S. Singh, A. Gupta, A.A. Efros, Unsupervised discovery of mid-level discriminative patches, in: ECCV, 2012, pp. 1378–1386.
[26] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: CVPR, 2009, pp. 413–420.

Jose C. Rubio received his B.S. degree in computer science from the University de València in 2007, and his M.Sc. in computer vision and artificial intelligence from the University Autonoma de Barcelona (UAB) in 2009. He got his Ph.D. with the UAB as a member of the Advance Driver Assistance Systems research group. Later he joined the University of Heidelberg as a post-doc with professor Bjorn Ommer.

Angela Eigenstetter received her B.S. and M.S. degrees in the Technical University of Darmstadt. She is currently working towards her Ph.D. degree in the University of Heidelberg under the supervision of professor Björn Ommer. Her research interests include part-based models, visual recognition and machine learning.

Björn Ommer received a diploma in computer science from the University of Bonn, Germany and the Ph.D. degree in computer science from ETH Zurich, Switzerland in 2007. Thereafter, he held a postdoctoral position in the computer vision group at the University of California, Berkeley. In 2009, he joined the University of Heidelberg, Germany, where he is a full professor for scientific computing in the Department of Mathematics and Computer Science, and he is also on the faculty of the Department of Philosophy. He is heading the computer vision group that is affiliated with the Heidelberg Collaboratory for image processing. His research interests include computer vision, machine learning, and cognitive science. Particular research topics are object and action recognition in images and video, shape analysis, perceptual organization, compositionality, and visual analysis in the life sciences and of cultural heritage.