Computer Vision and Image Understanding 138 (2015) 114–123

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Part level transfer regularization for enhancing exemplar SVMs

Yusuf Aytar*, Andrew Zisserman

Visual Geometry Group, Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, 01865 273000, United Kingdom

ARTICLE INFO

Article history:
Received 2 July 2014
Accepted 14 April 2015

Keywords:
Exemplar SVMs
Transfer learning
Object detection
Image retrieval
Feature mapping

ABSTRACT

Exemplar SVMs (E-SVMs, Malisiewicz et al., ICCV 2011), where an SVM is trained with only a single positive sample, have found applications in the areas of object detection and content-based image retrieval (CBIR), amongst others.

In this paper we introduce a method of part based transfer regularization that boosts the performance of E-SVMs, with a negligible additional cost. This enhanced E-SVM (EE-SVM) improves the generalization ability of E-SVMs by softly forcing it to be constructed from existing classifier parts cropped from previously learned classifiers. In CBIR applications, where the aim is to retrieve instances of the same object class in a similar pose, the EE-SVM is able to tolerate increased levels of intra-class variation, including occlusions and truncations, over E-SVM, and thereby increases precision and recall.

In addition to transferring parts, we introduce a method for transferring the statistics between the parts and also show that there is an equivalence between transfer regularization and feature augmentation for this problem and others, with the consequence that the new objective function can be optimized using standard libraries.

EE-SVM is evaluated both quantitatively and qualitatively on the PASCAL VOC 2007 and ImageNet datasets for pose specific object retrieval. It achieves a significant performance improvement over E-SVMs, with greater suppression of negative detections and increased recall, whilst maintaining the same ease of training and testing.

Content based image retrieval (CBIR), the problem of searching digital images in large databases according to their visual content, is a well established research area in computer vision. In this work we are particularly interested in retrieving subwindows of images which are similar to the given query image, i.e. the goal is detection rather than image level classification. The notion of *similarity* is defined as being the same object class but also having similar viewpoint (e.g. frontal, left-facing, rear etc.). A query image can be a part of an object (e.g. head of a side facing horse), a complete object (e.g. frontal car image), or a composition of objects (e.g. person riding a horse). For instance, given a query of a horse facing left, the aim is to retrieve any left facing horse (intra-*class* variation) which might be walking or running with different feet formations (exemplar deformation).

Recently exemplar SVMs (E-SVM) [33], where an SVM is trained with only a single positive sample, have found applications in the areas of CBIR [3,40] and object detection [33]. Since the E-SVM is trained from a single positive sample (together with many negatives), it is specialized to that given sample. This means that it can be strict (on viewpoint for example), and the negatives give some background suppression. However, the single positive is also a limitation: only so much can be learnt about the foreground of the query (and this can lead to false detections), and more significantly it can lead to lack of generalization. In our context, *generalization* refers to intra-class variation and deformation whilst maintaining the viewpoint. Learning such generalization from a single positive is challenging given the lack of examples of allowable deformations and intra-class variation.

In this work we propose a transfer learning approach for boosting the performance of E-SVMs using part-like patches of previously learned classifiers. The formulation softly constrains the learned template to be constructed from classifiers that have been fully trained (i.e. using many positives). For instance, the neck of a horse can be transferred from the tail of an aeroplane (see Fig. 1), or a jumping bike can borrow part of wheel patches from regular side facing bike or motorbike classifiers (see Fig. 2). The intuitive reason behind borrowing patches from other well trained classifiers is that these classifier patches bring with them a better sense of discriminative features and background suppression, and also bring some generalization properties. The result of the transfer learning is an enhancement of background suppression and tolerance to intra-class variation, hence better coping with occlusions and truncations in the

* Corresponding author.
E-mail address: yusuf@robots.ox.ac.uk, az@robots.ox.ac.uk (Y. Aytar).

http://dx.doi.org/10.1016/j.cviu.2015.04.004
1077-3142/

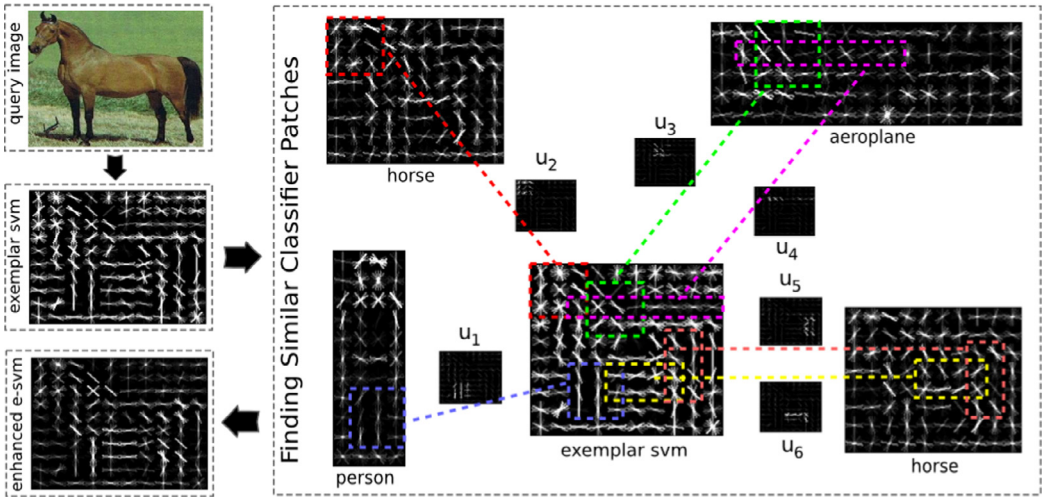

Fig. 1. Overview of the EE-SVM learning procedure. The box on the right shows mining classifier patches from existing classifiers by matching subparts of E-SVM trained from the given query image. Comparing E-SVM and EE-SVM, better suppression of the background can be seen from the visualized classifiers. Note, here and in the rest of the paper we only visualize the positive components of the HOG classifier.

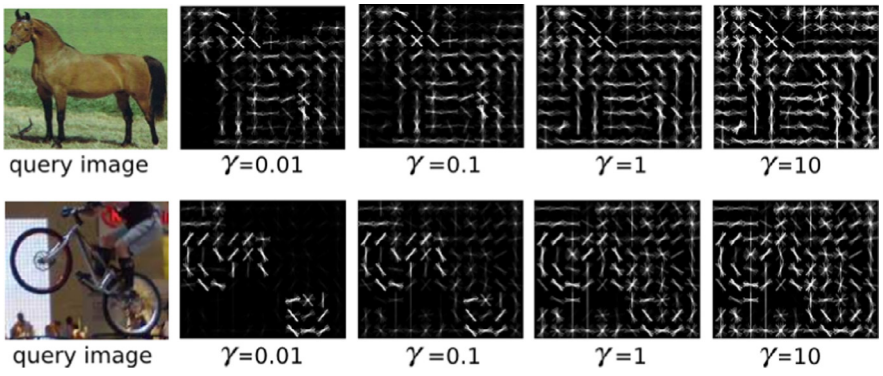

Fig. 2. Two limits of EE-SVM from reconstruction($\gamma = 0.01$) to E-SVM($\gamma = 10$). Learned EE-SVM templates with varying γ values are displayed. λ is fixed to 1.

query image. However, these enhancements incurs no (significant) additional cost in learning and testing. We term the boosted E-SVM, enhanced exemplar SVM (EE-SVM).

Objects and parts don't occur in isolation to each other. They appear with certain correlations in nature. For example, we don't expect to see a zebra in a city scene with road and cars or a bicycle next to a sailing boat in the middle of the sea. Stemming from these observations, co-occurrence statistics are utilized in the computer vision problems such as object detection [12], and semantic segmentation and labeling of objects [25,38] in the scenes. Similarly parts also appear with certain correlations: occurrence of feet supports occurrence of a head, or seeing one wheel increases the chance of seeing another in the close neighborhood. Parts can also have negative correlations, i.e. it is not expected to see spider legs or parts of insects in the close neighborhood of vehicle-like patches. Hence we can transfer not only parts, but also their natural co-occurrence statistics. We include these co-occurrence statistics of the parts, in a convex formulation, for softly enforcing these positive and negative correlations, in the EE-SVM objective.

We describe the relation with the prior work in Section 1, and then introduce the enhanced E-SVM in Section 2, and incorporate part correlations in Section 3. We relate introduced transfer learning methods with feature maps in Section 4. We give implementation details in Section 5. Finally we present a quantitative and qualitative evaluation in Section 6. Although it might be feared that judging the quality of retrieval results will be very subjective, we show that available annotation and measures from the PASCAL VOC [17] can be used for this task.

1. Relation to prior work

Exemplar SVMs are utilized in a variety of problems including object detection[33], face recognition[28], transferring segmentations masks and semantic scene parsing[45,51], cross domain image matching (matching drawings to pictures)[2,40], transferring 3D geometry [1,33], and transferring labels to 3D point clouds [50]. Here we propose a method for improving the quality of E-SVMs, which will potentially boost these performances of all these methods.

Our work uses the notion of parts as patches of classifier templates. Many other studies utilize shared parts across different classes. Torralba et al. [48] introduced a method for sharing small patch oriented templates in a boosting framework and Opelt et al. [34] extended this approach to shared boundary fragments. Fidler et al. [20] explored the shareability of features among object classes in a generative hierarchical framework. Stark et al. [44] proposed a method for transferring part-like shape features through explicit migration of model parameters for each part; however this transfer is manual at the moment. Ott and Everingham [35] introduced part sharing across classes for object detection in the framework of discriminatively trained part-based models [19]. In [22,42] HOG based templates are represented as a sparse reconstruction of shared parts (sparselets) which enable very fast evaluation of

multiple detectors. Dean et al. [9] also use the notion of part sharing for efficient evalution of large number of detectors. In another perspective Aytar and Zisserman [4] use shared parts to reconstruct and evaluate a single detector very fast on large image collections using inverted file indexing. Endres et al. [16] proposes to learn a diverse collection of discriminative parts from object bounding box annotations and utilize it for object category detection. Recently mining mid-level discriminative patches for scene understanding [13,14,23,36,37,41] has attracted considerable attention for their automatic discovery of distinctive parts for scene recognition. Mid-level discriminative patches are also utilized for inferring the 3D surface normals given a single image [7], and aligning object parts in order to discover visual connections in space and time [27] .

The proposed approach also has a strong relation to the line of work that focuses on enriching the image descriptors with the responses of mid-level and high-level classifiers [18,24,26,29,30,43,54]. These approaches either replace or augment the original low-level descriptor with the outputs of higher level classifiers. The proposed method also employs a similar augmentation scheme; however we augment the feature vector with the responses of previously learned classifier patches which are selected and relocated based on the quality of match with an E-SVM template learned from the query image.

Our approach can be viewed both as a transfer regularization approach and a feature augmentation approach. Hence it constructs an equivalence between these two views. We explicitly prove this equivalence and discuss its implications in Section 4.

2. Enhanced exemplar SVM

This section discusses the E-SVM formulation and introduces the enhanced E-SVM objective. The formulation of the E-SVM [33] is:

$$\min_{w,b} \quad \lambda||w||^2 + \sum_{i}^{N} \max\left(0, 1 - y_i(w^\mathsf{T} x_i + b)\right) \tag{1}$$

where λ controls the weight of regularization term, w is the classifier vector, b is the bias term, x_i and y_i are the training samples and their labels, respectively. Note that there is only one positive sample in the training set and its error is weighted more (50 times in [33]) than the negative samples. In order to simplify the formulation, different weightings of positive and negative samples are not explicitly shown.

In enhanced E-SVM, part based transfer regularization is incorporated to the E-SVM formulation. The objective is:

$$\begin{aligned}\min_{w,b,\alpha} \quad & \lambda||w - \sum_{i}^{M} \alpha_i u_i||^2 + \gamma \sum_{i}^{M} \alpha_i^2 \\ & + \sum_{i}^{N} \max\left(0, 1 - y_i(w^\mathsf{T} x_i + b)\right) \\ & st: \alpha_i \geq 0, \qquad \forall i\end{aligned} \tag{2}$$

where λ and γ controls the balance between the two regularization terms as well as the tradeoff between error term and regularization terms. u_i's are the classifier patches cropped from source classifiers and relocated on a w sized template padded with zeros other than the classifier patch (see Fig. 1), and α_i's are transfer weights. As α_i's are the reconstruction weights on u_i's and negative use of a part is not possible, non-negativity constraints are imposed on α_i's. Note that given a fixed set of u_i's the formulation is convex.

The two limits of this formulation are E-SVM and reconstruction from the classifier patches. As $\gamma \to \infty$, since α_i's will be forced to be zero due to infinite penalization, $\sum_i^M \alpha_i u_i$ will be a zero vector and (2) converges to the E-SVM formulation (1). As $\lambda \to \infty$, w will be forced to be equal to $\sum_i^M \alpha_i u_i$ and thus it will be forced to be constructed as a weighted combination of u_i's. Therefore by adjusting λ and γ we can obtain a midway solution between E-SVM and reconstruction from the other classifiers. Fig. 2 shows the smooth transition from reconstruction to E-SVM by changing γ with a fixed λ.

Discussion. Transfer regularization is introduced as an adaptive SVM (A-SVM) [31,52] which transfers information from a single auxiliary classifier. Subsequently A-SVMs were extended to transfer from multiple classes [53]. The proposed formulation is also a transfer regularization objective which transfers from the parts of previously learned classifiers. The main difference to [53] is that we control the weight of transfer with an additional regularization term ($\gamma \sum_i^M \alpha_i^2$) where $\gamma \to \infty$ indicates no transfer and $\gamma \to 0$ indicates maximum transfer. The equivalence of this formulation to a "classical" SVM formulation and advantages will be elaborated in Section 4. Note that this formulation is not specific to E-SVM and this transfer regularization can also be applied to "classical" SVM formulations.

3. Incorporating part correlations to EE-SVM

Parts generally occur with certain correlations. The existence of a part hints about the existence or absence of some other parts. For instance the occurrence of an up-left window-corner-like part might increase the occurrence probability of up-right window-corner-like part and decrease the existence probability of a zebra-face-like part. Consequently, in addition to transferring parts from the well trained source classifiers, we can also transfer statistical correlations between parts. This section will elaborate on the incorporation of such pairwise statistical correlations in the EE-SVM objective.

One common approach for enforcing pairwise statistics is performed through pairwise potential functions using Markov random fields (MRF) or conditional random fields (CRF) frameworks [5]. These potential functions are mostly non-convex and defined over discrete random variables; however, they can be efficiently optimized using graph-cuts or linear programming relaxations. Unfortunately they are not directly applicable for the SVM framework. Here we introduce a pairwise potential function that is convex and can be conveniently applied to convex regularized risk minimization frameworks, particularly SVMs. We are particularly interested in enforcing pairwise statistics in the transfer regularization formulations (i.e. Eq. 3) where the correlation is enforced on the activation of classifier parts u_i's in order to construct the classifier w. Assuming that a positive value of α_i represents the activation of the part u_i and $\alpha_i = 0$ indicates that u_i is not activated, the pairwise statistics can be captured through correlations between α_i, α_j pairs. The energy function to enforce this statistics can be defined as:

$$\min_{\alpha} \quad \phi(\alpha) = \sum_{i,j} C_{ij}||\alpha_i - \alpha_j||^2 \tag{3}$$

where C_{ij} is the pairwise correlation between the variables α_i and α_j. Then, the complete transfer objective is:

$$\begin{aligned}\min_{w,b,\alpha} \quad & \lambda||w - \sum_{i}^{M} \alpha_i u_i||^2 + \gamma \sum_{i}^{M} \alpha_i^2 \\ & + \sum_{i}^{N} \max\left(0, 1 - y_i(w^\mathsf{T} x_i + b)\right) \\ & + \theta \sum_{i,j}^{M,M} C_{ij}||\alpha_i - \alpha_j||^2 \qquad st: \alpha_i \geq 0, \qquad \forall i\end{aligned} \tag{4}$$

where θ is the hyperparameter that controls the weight of enforcing statistical correlations.

Intuitively a positive C_{ij} enforces the values of α_i and α_j to be as close as possible, therefore if one activated the other will be as well, especially when C_{ij} is a very high positive value. Conversely, a negative C_{ij} forces them to be as distinct as possible. However this objective is not necessarily convex, particularly when C_{ij} contains negative values. In order to obtain a convex energy function, we introduce a

slight addition to the pairwise potential and define $\bar{\phi}$ as:

$$\bar{\phi}(\alpha) = \sum_{i,j} C_{ij}||\alpha_i - \alpha_j||^2 - 4\sum_i D^-_{ii}\alpha_i^2 \quad (5)$$

$$= \sum_{i,j} C^+_{ij}||\alpha_i - \alpha_j||^2 - \sum_{i,j} C^-_{ij}||\alpha_i + \alpha_j||^2 \quad (6)$$

where the pairwise correlation matrix C is decomposed into its positive and negative components as $C = C^+ + C^-$, D^+ is a diagonal matrix with entries $D^+_{ii} = \sum_j C^+_{ij}$, and D^- is a diagonal matrix with entries $D^-_{ii} = \sum_j C^-_{ij}$. Intuitively by introducing more penalization (i.e. $-4\sum_i D^-_{ii}\alpha_i^2$) over highly negatively correlated α_i's, which is sensible since these α_i's have smaller probabilities to be activated, we can obtain a convex pairwise potential (6). Finally by plugging $\bar{\phi}(\alpha)$ into (4) as the pairwise potential, we obtain a convex minimization objective that combines transfer regularization and co-occurrence statistics of the parts:

$$\min_{w,b,\alpha} \lambda||w - \sum_i^M \alpha_i u_i||^2 + \gamma\sum_i^M \alpha_i^2 + \sum_i^N \max\left(0, 1 - y_i(w^\top x_i + b)\right) \quad (7)$$

$$+\theta\left(\sum_{i,j}^{M,M} C^+_{ij}||\alpha_i - \alpha_j||^2 - \sum_{i,j}^{M,M} C^-_{ij}||\alpha_i + \alpha_j||^2\right) \qquad st: \alpha_i \geq 0, \quad \forall i$$

This new version of EE-SVM will be referred to as EE-SVM-COR. The part correlations are learnt from the source filters, and the implementation details will be discussed in Section 5.

Discussion. Another way of enforcing part correlations is to use inverse covariance matrix regularization, i.e. $\alpha^\top\Sigma^{-1}\alpha$, assuming that Σ is a valid covariance matrix. However, due to the large number of parameters and the limited samples the estimated Σ may not be positive definite which prevents us from computing the inverse. In [21] Σ is treated as an affinity matrix (i.e. stronger correlation means higher affinity). They directly used it for constructing the precision matrix of a Gaussian and used it in the regularization as $\alpha^\top(I - \lambda\Sigma)\alpha$, where λ is a hyperparameter that ensures the positive definiteness. We also treat the correlation matrix C_{ij} as an affinity matrix and enforce part correlations in a similar manner. The detailed derivations of the formulations are given in Section 4.

4. Relating feature maps, transfer learning and optimization

In this section, initially we will investigate transforming transfer learning formulations to a "classical" SVM formulation, which comes with the benefit of using easy and robust optimization for transfer learning approaches and makes it potentially possible to use for practical purposes without the need of expert knowledge in transfer learning. Several equivalence relations between transfer regularization and feature mapping will be investigated, and their corresponding implications will be discussed.

4.1. Transfer regularization by feature mapping

Transfer regularization [52] is applied for many transfer learning approaches in object classification and detection. We'll start with the general form of the formulation used by [47] for transferring from multiple sources which uses squared loss. Due to the robustness of the hinge loss over the squared loss, we substitute the squared loss term with the hinge loss and obtain:

$$\min_{\alpha,w,b} \quad \lambda||w - \sum_i^M \alpha_i u_i||^2 + \sum_i^N \max\left(0, 1 - y_i(w^\top x_i + b)\right) \quad (8)$$

$$st: ||\alpha|| \leq 1$$

where u_i's are the source classifiers and α_i's are the corresponding weights of transfer for each source classifier. After a slight modification to the transfer regularization formulation (i.e. from (8) to (2) by bringing the constraint into the objective), we will transform the problem to a "classical" SVM formulation through a feature augmentation approach. The derivation below steps through the rearrangements for mapping the transfer regularization objective to an equivalent "classical" SVM formulation where the feature vector is augmented with the responses of source classifier models. Let $w = \hat{w} + \sum_i^M \alpha_i u_i$, and then the derivation is:

$$\lambda||w - \sum_i^M \alpha_i u_i||^2 + \gamma\sum_i^M \alpha_i^2 + \sum_i^N \max\left(0, 1 - y_i(w^\top x_i + b)\right) \quad (9)$$

$$= \lambda||\hat{w}||^2 + \gamma\sum_i^M \alpha_i^2 + \sum_i^N \max\left(0, 1 - y_i\left(\hat{w}^\top x_i + \left(\sum_i^M \alpha_i u_i\right)^\top x_i + b\right)\right) \quad (10)$$

$$= ||\bar{w}||^2 + \sum_i^N \max\left(0, 1 - y_i(\bar{w}^\top \bar{x}_i + b)\right) \quad (11)$$

where

$$\bar{w} = [\sqrt{\lambda}\hat{w}; \sqrt{\gamma}\alpha_1; \sqrt{\gamma}\alpha_2; ...; \sqrt{\gamma}\alpha_M],$$

$$\bar{x}_i = \left[\frac{1}{\sqrt{\lambda}}x_i; \frac{1}{\sqrt{\gamma}}u_1^\top x_i; \frac{1}{\sqrt{\gamma}}u_2^\top x_i; ...; \frac{1}{\sqrt{\gamma}}u_M^\top x_i\right],$$

$\bar{w}$ is the transformed classifier and $\bar{x}_i$ is the augmented feature vector with the responses of u's on x_i. The classifier w, the solution to the original problem (9), can easily be computed from $\bar{w}$ since $w = \hat{w} + \sum_i^M \alpha_i u_i$. As is clear from (11), the transformed problem is a "classical" SVM formulation with feature augmentation, and it can be optimized efficiently using existing powerful SVM solvers. Note that this derivation is applicable to both EE-SVM objective and previously used transfer regularization formulations.

Discussion. The major implication of this derivation is that transfer regularization can also be stated as a classical SVM minimization problem where the feature vector is augmented with the responses of source classifiers. This equivalence constructs a bridge between papers implementing feature augmentation or populating the features with the responses of high-level classifiers [15,18,24,26,32,49,54] and papers performing transfer regularization [46,47,52,53]. Another direct implication is that transfer regularization approaches [46,47,52,53], which requires specialized optimization, can be reformulated to be efficiently optimized with the state-of-the-art SVM solvers.

Another way of converting (9) to a normal SVM formulation would be solving α analytically and substituting it into the original equation. Let the regularization part of (9) be $R = \lambda||w - U\alpha||^2 + \gamma||\alpha||^2$ then the derivation is as follows:

$$\frac{\partial R}{\partial \alpha} = 0 \quad \rightarrow \quad \alpha = \lambda(\gamma I + \lambda U^\top U)^{-1}U^\top w \quad (12)$$

By substituting α in to R we obtain:

$$M = \sqrt{\lambda}(I - \lambda U(\gamma I + \lambda U^\top U)^{-1}U^\top) + \lambda\sqrt{\gamma}(\gamma I + \lambda U^\top U)^{-1}U^\top \quad (13)$$

$$R = ||Mw||^2 = w^\top M^\top M w \quad (14)$$

Similar to the derivation in (11), this regularization can also be implemented as a feature transformation (i.e. $\bar{w} = Mw, \quad \bar{x}_i = (M^\top)^{-1}x_i$) and solved using a normal SVM formulation.

4.2. Enforcing co-occurrence statistics by feature mapping

This section will discuss how to map the objective (7) to a "classical" SVM formulation by defining appropriate feature maps. Initially the formulation will be converted to a graph Laplacian form, and then by incorporation of some new variables, a few pseudo-training samples and appropriate rearrangements the objective will be transformed to a linear SVM objective.

Assuming that C_{ij} are the weights on a graph, $\phi(\alpha)$ in (3) can be re-written in the form below:

$$\phi(\alpha) = \sum_{i,j} C_{ij}||\alpha_i - \alpha_j||^2 = 2\alpha^\top L\alpha \quad \text{where} \quad L = D - C, \quad D_{ii} = \sum_j C_{ij}, \tag{15}$$

where L is the graph Laplacian. Similarly we can rewrite $\bar{\phi}(\alpha)$ in (5) as:

$$\bar{\phi}(\alpha) = \sum_{i,j} C_{ij}||\alpha_i - \alpha_j||^2 - 4\alpha^\top D^- \alpha = 2\alpha^\top \bar{L}\alpha \quad \bar{L} = \bar{D} - C, \quad \bar{D}_{ii} = \sum_j |C_{ij}| = D^+ - D^- \tag{16}$$

Let $U = [u_1, u_2, .., u_M]$ and $w = \hat{w} + \sum_i^M \alpha_i u_i = \hat{w} + U\alpha$, rewriting (7) with appropriate substitutions:

$$\begin{aligned} &\min_{w,b,\alpha} \quad \lambda\hat{w}^\top\hat{w} + \alpha^\top(\gamma I + 2\theta\bar{L})\alpha \\ &+ \sum_i^N \max\left(0, 1 - y_i(\hat{w}^\top x_i + \alpha^\top U^\top x_i + b)\right) \\ &st: \alpha_i \geq 0, \quad \forall i \end{aligned} \tag{17}$$

Considering both $\bar{L}$ and I are positive-semi definite, we can decompose the term $\gamma I + 2\theta\bar{L}$ using the Cholesky decomposition as $RR^\top$ then introduce $\beta = R^\top\alpha$ and rewrite:

$$\alpha^\top(\gamma I + 2\theta\bar{L})\alpha = \alpha^\top RR^\top\alpha = \beta^\top\beta \tag{18}$$

$$\alpha^\top = \beta^\top R^{-1} \tag{19}$$

Plugging in (18) and (19) to (17) we obtain:

$$\begin{aligned} &\min_{\hat{w},b,\beta} \quad \lambda\hat{w}^\top\hat{w} + \beta^\top\beta \\ &+ \sum_i^N \max\left(0, 1 - y_i(\hat{w}^\top x_i + \beta^\top R^{-1}U^\top x_i + b)\right) \\ &st: \left[(R^{-1})^\top\beta\right]_i \geq 0, \quad \forall i \\ &= \min_{\bar{w},b} \quad ||\bar{w}||^2 + \sum_i^N \max\left(0, 1 - y_i(\bar{w}^\top\bar{x}_i + b)\right) \\ &st: \left[(R^{-1})^\top\beta\right]_i \geq 0, \quad \forall i \\ &\text{where} \quad \bar{w} = [\sqrt{\lambda}\hat{w}; \beta], \quad \bar{x}_i = [\frac{1}{\sqrt{\lambda}}x_i; R^{-1}U^\top x_i], \end{aligned} \tag{20}$$

Other than the linear constraints $\left[(R^{-1})^\top\beta\right]_i \geq 0$, the objective becomes a linear SVM objective.

Enforcing linear constraints in SVMs. For any classical SVM formulation:

$$\min_{w,b} \quad ||w||^2 + \sum_i^N \max\left(0, 1 - y_i(w^\top x_i + b)\right) \tag{21}$$

we can implement a linear constraint $a^\top w > 0$ by introducing an additional training sample $x_{N+1} = [\infty \times a]$ with the label $y_{N+1} = +1$. Here $\infty \times a$ is the multiplication of the vector a with the infinity which is approximated with a very high numeric value (i.e. 10^6).

A similar approach will be used for implementing the linear constraints $[(R^{-1})^\top\beta]_i \geq 0, \forall i$ in (20). Another additional training sample will be added for each row of $(R^{-1})^\top$ as each row introduces another constraint on β. Since the constraints only concern the β part of $\bar{w} = [\sqrt{\lambda}\hat{w}; \beta]$, the initial parts of the additional training sample will be filled with a zero vector of the same length as $\hat{w}$. The rest of it will be filled by the appropriate row extracted from $(R^{-1})^\top$ and multiplied by infinity. In a formal definition, linear constraints are implemented by adding M additional samples to our training set as below:

$$\bar{x}_{N+k} = [\mathbf{0}^{|\hat{w}|}; \infty \times (R^{-1})^\top_{k,:}], \quad y_{N+k} = +1, \quad 1 \leq k \leq M \tag{22}$$

where $\bar{x}_{N+k}$ are the additional training samples that implement the constraints, $\mathbf{0}^{|\hat{w}|}$ is the zero vector of the same length as $\hat{w}$, $(R^{-1})^\top_{k,:}$ is the k^{th} row of $(R^{-1})^\top$.

After handling the constraints as well, the transformation of the original objective (7) to a classical SVM formulation is completed. Therefore we can optimize (7) using available robust SVM optimization tools.

5. EE-SVM training and obtaining the part vocabulary

In this section the details of the implementation will be discussed. Initially training source classifiers and E-SVM will be discussed. Then EE-SVM training procedures and obtaining part correlations will be explained.

Source classifiers and part vocabulary. The source classifiers are linear SVM classifiers (templates) over HOG [8,19] features. The source classifiers are trained with three components for each class and without parts, similar to the procedure in [19]. While testing on the PASCAL VOC 2007 test set, the source classifiers are learnt from the 1000 classes of the ImageNet [11] 2012 challenge. In total, the 329 models out of 1000 are selected that have an AP over 30% on a small validation set of 1000 images (50 positive images per class). Together with the mirrored versions of these templates it totals to $329 \times 3 \times 2 = 1974$ source classifiers. To build the vocabulary of classifier patches, all patches with sizes of 5×5 are extracted from the components of the trained DPMs, then a k-means clustering is performed with $k = 10K$ centers. Other ways of composing the vocabulary could be learning the vocabulary via sparse reconstruction as in [42], or learning both the vocabulary items and the classifiers at the same time as in [22]; however this is beyond the scope of our work. While testing on ImageNet, the source classifiers are trained using 20 classes of the PASCAL VOC 2007 training set. All 5×5 classifier patches are used to form a vocabulary of 1948 items.

E-SVM training. Similar to [33], each E-SVM is composed of 100 or slightly less HOG cells where the aspect ratio is chosen according to the query image. The E-SVM is trained with the given query image as the positive sample and randomly selected 2000 negative images from the PASCAL VOC 2007 training set. The training is performed iteratively in a similar fashion to [33] where mined hard negatives are incorporated to the learning after each iteration.

EE-SVM training. The training procedure of EE-SVM, which is briefly visualized in Fig. 1, starts with training an E-SVM classifier from the given query image. After obtaining the E-SVM, for each 5×5 cell classifier patch a *good match* is searched for within the vocabulary. This linear search can be efficiently done using fast matrix multiplication since we have a limited number of vocabulary items. Even though we use 5×5 cell classifier patches for experimental validation, any other varying size and aspect ratio can also be applied. A *good match* is defined by thresholding the cosine similarity (normalized dot product) between the E-SVM patch (a $5 \times 5 \times 32$ dimensional vector) and vocabulary items. This threshold value is fixed to 0.2. After determining where to transfer from, each patch is relocated on a w sized HOG template padded with zeros other than the transferred classifier patch. Finally learning of the EE-SVM is performed

Table 1
MAP (mean average precision) comparisons of EE-SVM variants and E-SVM with changing quality groups. For instance AP $\geq$ 0.01 means the queries which achieved AP = 0.01 or above. Random 10 images are selected for each of the 17 classes and four major poses (i.e. left, right, frontal, rear) from the PASCAL'07 training set as the queries. Tests are performed on PASCAL'07 test set. "EE-SVM w/o learning" defines the classifier as the average of u_i's (i.e. it doesn't learn α_i's). "EE-SVM reconstruction only" learns the classifier as the reconstruction from u_i's only (i.e. $w = \sum_i^M \alpha_i u_i$, this is achieved by solving (2) when $\lambda = \infty$), and the last row is the full EE-SVM (i.e. setting $\lambda = \gamma = 1$ in (2)).

AP $\geq$	0.00	0.01	0.05	0.10	0.15	0.20	0.30	0.50
# of queries	558	216	112	73	48	38	21	8
E-SVM	3.2	8.2	14.1	18.8	23.7	26.7	32.6	41.1
EE-SVM w/o learning	3.3	8.3	14.6	20.3	26.1	29.9	40.8	53.6
EE-SVM reconstruction only	3.5	8.6	15.1	21.0	27.0	30.8	42.0	54.2
EE-SVM	4.2	10.5	18.1	24.2	30.7	34.2	42.8	53.4

using the same set of training samples used for training the E-SVM and no new hard negatives are collected.

Partial matching for occlusion and truncation. In occluded or truncated queries, the parts of the E-SVM classifier might not match well with the vocabulary items due to the distortion caused by occlusions and truncations. Hence instead of using *good mathces* of the E-SVM patch from the vocabulary items as defined above, ignoring a few cells (potentially occluded or truncated regions) while computing similarity would be a better idea. The matches obtained by this partial matching method is referred as *partial good mathces* of the E-SVM patch. Here we only use the most similar (i.e. cosine similarity) β percent of the cells for matching an E-SVM patch to a vocabulary item. For instance if the E-SVM patch and vocabulary items are 5×5 cells and β is 70%, then the top matching 18 (i.e. 70% of 25) cells are used for computing normalized dot product and deciding if the vocabulary item is a partial good match or not. Similar to the *good match*, the same threshold of 0.2 is used for the decision.

Optimization. The optimization of the EE-SVM and EE-SVM-COR objectives are performed using the LIBSVM [6] package through the equivalent feature mapping formulations which are already discussed in Section 4. The only additional cost of EE-SVM and EE-SVM-COR over E-SVM is the transformation of training samples, and training another SVM, which constitutes less than 1% of the training time (i.e. mining hard negatives is costly). The test time complexity of EE-SVM and EE-SVM-COR is exactly the same as that of E-SVM.

Part correlations. The entries of the pairwise part correlation matrix C_{ij} are estimated using the pairwise sample correlation coefficient $\rho_{i,j}$ which is computed from the joint occurrence of parts i and j in the source filters (i.e. the source filters are the samples). The computation of the correlation coefficient is:

$$\rho_{i,j} = \frac{\text{cov}(i,j)}{\sigma_i \sigma_j} = \frac{E[(i-\mu_i)(j-\mu_j)]}{\sigma_i \sigma_j} \tag{23}$$

where cov is the sample covariance, σ_i is the standard deviation of occurrence of part i computed over samples (1 states occurrence and 0 states absence of part i in the given sample), μ_i is the mean occurrence of part i across all the samples, and E is the expectation.

6. Experiments

In this section the experimental results will be described. Initially we give the experimental settings, evaluation metrics and the defaults for the hyperparameters. In the next two sections, we then discuss two sets of experiments performed on the PASCAL VOC 2007 [17] dataset and ImageNet [11]. Average precision (AP), and precision at top K (PR@5, PR@10, PR@50, PR@100) retrievals are used for evaluating the quality of retrieval results. A correct retrieval is defined as the same object class with the same pose as the query image and the retrieved subwindow should have at least 50% overlap with the true bounding box around the object class. The definition of the pose is inherited from the PASCAL VOC metrics [17] where four main poses exist namely *left, right, frontal, rear* (the pose "unspecified" is omitted). In all the experiments the proposed approaches are compared with the E-SVM method. Both λ and γ parameters are fixed to 1, and θ is fixed to 10 in all the experiments unless otherwise stated. The matching similarity threshold, which determines the *good* classifier patch matches based on the normalized dot product of two vectors, is fixed to 0.2.

6.1. Evaluation on PASCAL VOC

The retrievals of PASCAL'07 classes with four main poses are evaluated. The part vocabulary is obtained from ImageNet. The query images are selected as all non-truncated images of the 17 classes (bottle, dining table and potted plant are omitted since they don't have poses) with four main poses from the PASCAL'07 training set. For each query image, an E-SVM and EE-SVM variants are trained and run on the test set. Ground truth is identified as the same object class with the same pose label. Due to the strong visual similarities in the poses (e.g. left–right symmetries for bicycle, car, motorbike, etc.), the detections of the same object class other than the target pose are omitted and not counted towards AP computation. For instance if we are searching for a bicycle facing left, we ignore (i.e neither count as positive or negative) the detections of front, rear, left or unspecified poses of bicycle. In total 1659 queries from 17 classes are evaluated, and the pose distribution is: 452 left, 439 right, 561 frontal, and 207 rear.

For some query images, due to being unusual examples of the pose (e.g. left facing bicycle with front wheel up as in Fig. 2), the AP results can be very low. Conversely for some others, which are canonical examples of the pose, the AP results are much higher. In order to have a better insight on the results and see the boost for different quality of samples, we grouped the queries as being above some AP threshold. The query belongs to the quality group AP > threshold, if the AP of E-SVM, EE-SVM or EE-SVM-COR is above the defined *threshold* (for instance group AP $\geq$ 0 means all the queries).

Table 1 shows the AP improvements of EE-SVM variants over E-SVM on a smaller subset of randomly selected 10 query images for each of the 17 classes and 4 major poses. In this table we demonstrated EE-SVM variants in order to see the effects of different components of EE-SVM. The evaluated models are: (a) "EE-SVM w/o learning" which defines the classifier as the average of u_i's (i.e. $w = \frac{1}{M}\sum_i^M u_i$, it doesn't learn α_i's), (b) "EE-SVM reconstruction only" which learns the classifier as the reconstruction from u_i's only (i.e. $w = \sum_i^M \alpha_i u_i$, this is achieved by solving (2) when $\lambda = \infty$), and (c) full EE-SVM (i.e. setting $\lambda = \gamma = 1$ in formulation (2)). All EE-SVM variants outperform E-SVM. Learning the combination weights (i.e. α_i's) as in "EE-SVM reconstruction only" improves to performance over just averaging the u_i's (i.e. "EE-SVM w/o learning"). Finally the EE-SVM, the midway solution between E-SVM and "EE-SVM reconstruction only", performs significantly better than the others.

Table 2 shows the overall results and the AP improvement of EE-SVM and EE-SVM-COR over E-SVM. This experiment is performed on

Table 2
Relative MAP (mean average precision) improvements of EE-SVM and EE-SVM-COR over E-SVM with changing quality groups. Queries are the images of 17 classes with four major poses (i.e. left, right, frontal, rear) from the PASCAL'07 training set. Tests are performed on PASCAL'07 test set. EE-SVM-COR constantly outperforms EE-SVM which significantly improves over E-SVM.

AP ≥	0.00	0.01	0.05	0.10	0.15	0.20	0.30	0.50
# of queries	1659	650	346	241	169	121	70	14
E-SVM	3.3	8.2	14.0	17.9	21.6	25.3	30.0	44.2
EE-SVM	4.2	10.6	17.9	22.8	27.6	32.4	39.1	55.2
EE-SVM-COR	4.3	10.8	18.3	23.4	28.5	33.4	40.4	57.3
Rel. Imp. of EE-SVM	28.3	28.4	27.9	27.6	28.1	27.9	30.1	24.9
Rel. Imp. of EE-SVM-COR	30.9	31.1	30.9	31.0	32.3	32.0	34.7	29.7

Table 3
MAP results and relative improvements of EE-SVM-COR over E-SVM for individual classes for the quality group (AP ≥ 0.05). Tests are performed on PASCAL'07 test set. Only classes which have more than five queries are shown.

Class	Plane	Bicycle	Bus	Car	Cow	Dog	Horse	M.bike	Sheep	TVmonitor
# of queries	5	59	15	129	13	6	33	28	16	32
E-SVM	5.3	23.3	8.5	16.0	4.9	5.6	9.3	11.3	8.6	10.2
EE-SVM	7.9	30.7	9.0	20.7	7.7	8.9	11.4	13.0	11.9	11.3
EE-SVM-COR	7.1	32.1	8.8	21.0	8.5	10.2	11.6	12.9	12.3	11.3
Rel.Imp.	34.3	38.0	3.4	31.0	74.2	82.8	25.4	13.9	44.3	10.2

Fig. 3. Retrieval results of PASCAL'07 queries. Top 3 positives and negatives are being displayed. Orders in the ranked list is shown left bottom corner of each image.

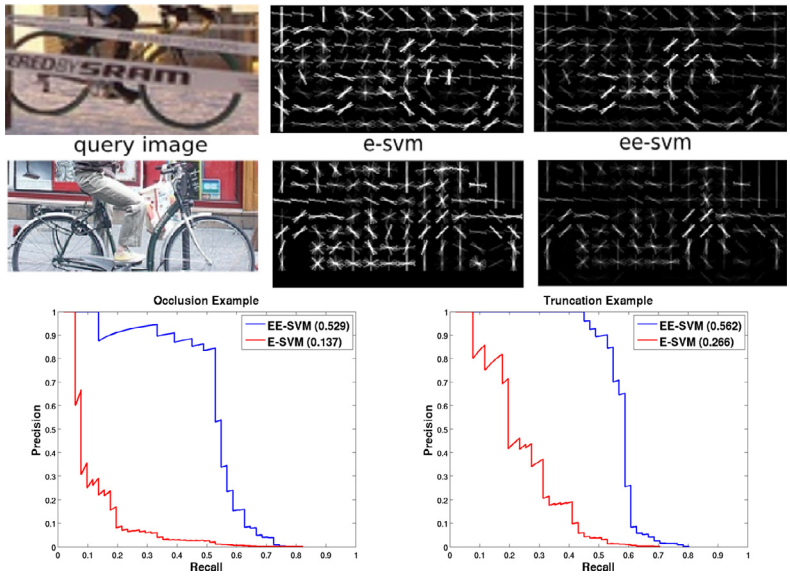

Fig. 4. Occlusion and truncation handling via EE-SVM. The top row shows an occlusion example, note the decreased effect of occlusions around the wheels of the bicycle in the EE-SVM model. The second row shows how the bottom part of the wheels are softly completed by the EE-SVM. It is better visualized with zooming into the document.

Table 4
Precision at top *K* comparison of ImageNet Queries. Three queries with varying poses are evaluated for each class from ImageNet and the mean precisions are presented. Retrieval is performed on the collection composed of PASCAL'07 test images and corresponding ImageNet category.

	Lion		Deer		Tandem		Bulldozer		Ambulance		MEAN	
	e-SVM	ee-SVM	e-SVM	ee-SVM	e-SVM	ee-SVM	e-SVM	ee-SVM	e-SVM	ee-SVM	e-SVM	ee-SVM
PR@5	0.47	**0.60**	0.60	**0.73**	1.00	1.00	0.93	0.93	1.00	1.00	0.80	**0.85**
PR@10	0.40	0.40	0.50	0.50	1.00	1.00	0.90	0.90	1.00	1.00	0.76	0.76
PR@50	0.19	**0.21**	0.30	**0.32**	0.98	**0.99**	0.62	**0.67**	0.85	**0.87**	0.59	**0.61**
PR@100	0.13	**0.14**	0.22	**0.28**	0.93	**0.97**	0.42	**0.49**	0.76	**0.80**	0.49	**0.54**

the full query set. In all the quality groups EE-SVM significantly improves over E-SVM, and EE-SVM-COR improves over EE-SVM. Even though the actual AP improvements are changing across the quality groups of samples, the relative boost of EE-SVM and EE-SVM-COR are consistently similar in different quality groups. In Table 3 the AP results and improvements are shown for individual classes for the quality level ($AP \geq 0.05$). For statistical significance only the classes which have more than 5 queries are shown. For all the classes EE-SVM and EE-SVM-COR significantly outperforms E-SVM.

A few qualitative results can be seen in Fig. 3 where the top three positives and negatives are shown with their ranks in the ordered list of retrieved subwindows. In EE-SVM retrievals the ranks of the top three negatives are much later; this shows that EE-SVM better suppresses the negatives and thus increases the recall.

Handling occlusion and truncation via EE-SVM. It is quite common to come across truncated and occluded query images. Here we'll briefly present a potential use scenario of EE-SVM for handling occlusions and truncations. Considering that certain parts of the query object is not visible in truncated or occluded queries, for each E-SVM classifier patch *partial good matches* (see Section 5) of vocabulary items are used instead of *good matches*. This procedure simply ignores a few cells (potentially the ones which correspond to occluded and truncated regions of the classifier patch) while performing the matching.

Fig. 4 shows two examples of handling occlusion and completing the truncated parts. In these examples 5×5 patches are used and the β parameter of the *partial good match* measure is 70%. γ parameter, which defines the strength of transfer, is set to 1. Particularly in the truncation example (second row on Fig. 4) the truncated parts (i.e. the bottom extension of the query) are partially completed with the support coming from matched vocabulary items, note that there is no visual data coming from the query image for these sections.

6.2. Evaluation on ImageNet

These experiments are conducted on ImageNet and the part vocabulary in obtained from PASCAL'07 dataset. Since PASCAL'07 has only 20 classes, robust computation of the part co-occurrence statistics is not feasible: hence the EE-SVM-COR is omitted in these experiments. For quantitative experiments five ImageNet classes (synsets) are selected: *lion, deer, tandem, bulldozer,* and *ambulance*. For each of these classes three random queries (from one of the main canonical poses) are selected from web images and evaluated on a test set which contains the images of the corresponding synset (~1300 images) and the PASCAL'07 test set (~5000 images). The evaluations are compared using precision at the top *K* retrievals. From the results, displayed in Table 4, we can conclude that the recall of EE-SVM is much better than the recall of E-SVM, particularly for top 50 and top 100 retrievals.

In addition to canonical poses, the method is also qualitatively demonstrated for unusual poses. With the help of part based transfer, since parts can be relocated and migrated across classes, even for quite unusual poses we can obtain significant improvements. The left facing bicycle with the front wheel up (see Figs. 2 and 5) is a nice example where the wheel patches are transferred from motorbike and bicycle classifiers with regular poses. Another example, displayed in Fig. 5, is a sitting lion where the ranks of positives clearly show EE-SVM's ability for better recall.

Recently detecting person–object interactions [10] and compositions of objects [39] gained popularity. Our method can also be

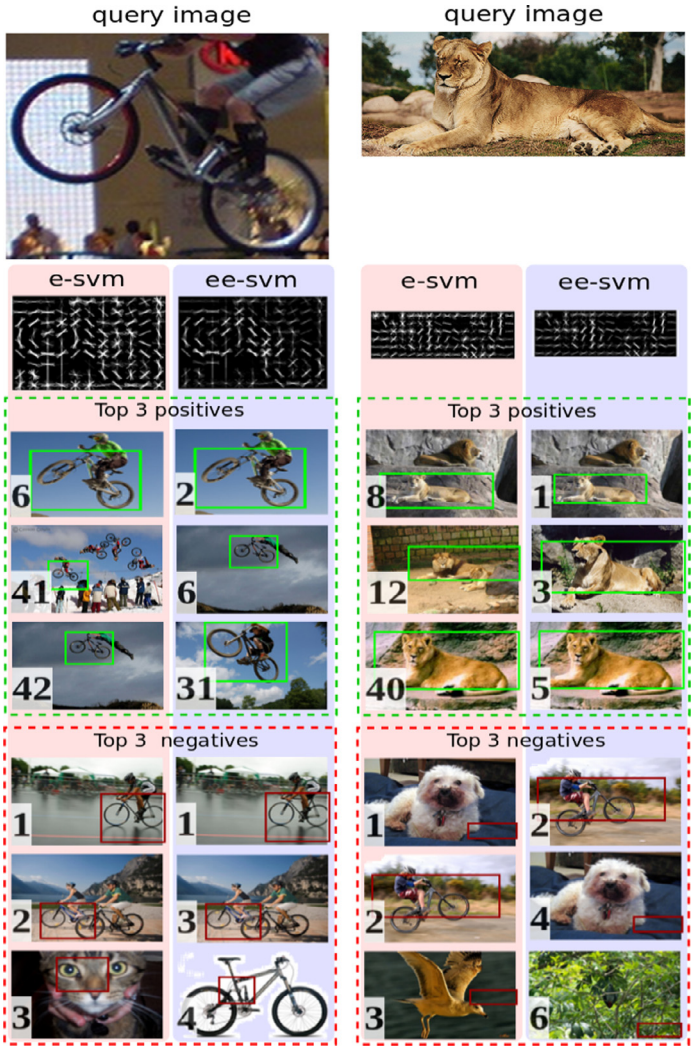

Fig. 5. Retrieval of unusual poses on ImageNet.

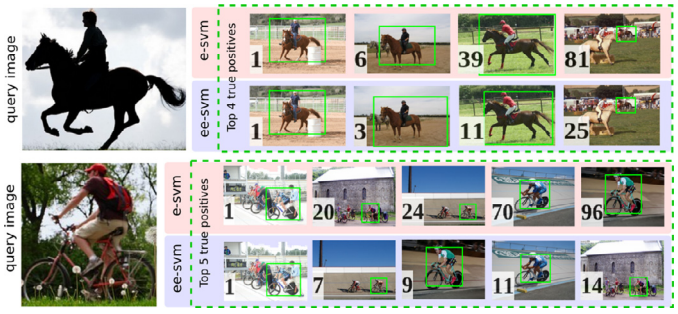

Fig. 6. Retrieval of person-object compositions.

utilized in a similar scenario. A qualitative example for retrieving such object compositions (i.e. person riding horse) is demonstrated in Fig. 6.

7. Conclusion

We introduced a method of part based transfer regularization that boosts the performance of E-SVMs. We demonstrated that EE-SVM suppresses the false detections significantly better than E-SVM. This improvement is shown both quantitatively and qualitatively on PASCAL'07 and ImageNet queries with canonical and unusual poses including compositions of objects. We also discussed the potential advantages of EE-SVM for handling truncation and occlusion.

We also introduced a convex potential function for incorporating the pairwise co-occurrence relations into convex max-margin learning frameworks. We showed that by defining appropriate feature maps, many transfer learning formulations are transformed to a classical SVM formulation, and subsequently solved by much easier and more robust optimization tools developed throughout the past years.

References

[1] M. Aubry, D. Maturana, A. Efros, B. Russell, J. Sivic, Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of CAD models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[2] M. Aubry, B.C. Russell, J. Sivic, Painting-to-3d model alignment via discriminative visual elements, ACM Trans. Graph. 33 (2) (2014) 14.
[3] Y. Aytar, A. Zisserman, Enhancing exemplar SVMS using part level transfer regularization, in: Proceedings of the British Machine Vision Conference, 2012.
[4] Y. Aytar, A. Zisserman, Immediate, scalable object category detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[5] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[6] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.
[7] A. Gupta, D. Fouhey, M. Hebert, Data-driven 3D primitives for single image understanding, in: Proceedings of the International Conference on Computer Vision, 2013.
[8] N. Dalal, B. Triggs, Histogram of Oriented Gradients for Human Detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[9] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, Fast, accurate detection of 100,000 object classes on a single machine, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[10] V. Delaitre, J. Sivic, I. Laptev, Learning person-object interactions for action recognition in still images, in: Advances in Neural Information Processing Systems, 2011.
[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[12] C. Desai, D. Ramanan, C.C. Fowlkes, Discriminative models for multi-class object layout, Int. J. Comput. Vision 95 (1) (2011) 1–12.
[13] C. Doersch, A. Gupta, A.A. Efros, Mid-level visual element discovery as discriminative mode seeking, in: Advances in Neural Information Processing Systems, 2013.
[14] C. Doersch, A. Gupta, J. Sivic, A.A. Efros, What makes paris look like paris? ACM Trans. Graph. 31 (4) (2012) 101.
[15] M. Douze, A. Ramisa, C. Schmid, Combining attributes and Fisher vectors for efficient image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[16] I. Endres, K.J. Shih, J. Jiaa, D. Hoiem, Learning collections of part models for object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[17] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (VOC) challenge, Int. J. Comput. Vision (2010).
[18] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[19] P.F. Felzenszwalb, R.B. Grishick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. (2010).
[20] S. Fidler, M. Boben, A. Leonardis, Evaluating multi-class learning strategies in a generative hierarchical framework for object detection, in: NIPS, 2009.
[21] T. Gao, M. Stark, D. Koller, What makes a good detector? structured priors for learning from few examples, in: Proceedings of the European Conference on Computer Vision, 2012.
[22] R. Girshick, H.O. Song, T. Darrell, Discriminatively activated sparselets, in: Proceedings of the International Conference on Machine Learning, 2013.
[23] M. Juneja, A. Vedaldi, C.V. Jawahar, A. Zisserman, Blocks that shout: Distinctive parts for scene classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[24] M.P. Kumar, P.H.S. Torr, A. Zisserman, Efficient discriminative learning of parts-based models, in: Proceedings of the International Conference on Computer Vision, 2009.
[25] L. Ladicky, C. Russell, P. Kohli, P.H.S. Torr, Graph cut based inference with co-occurrence statistics, in: Proceedings of the European Conference on Computer Vision, 2010.
[26] C.H. Lampert, M.B. Blaschko, Structured prediction by joint kernel support estimation, Mach. Learn. (2009).
[27] Y.J. Lee, A.A. Efros, M. Hebert, Style-aware mid-level representation for discovering visual connections in space and time, in: Proceedings of the International Conference on Computer Vision, 2013.
[28] H. Li, Z. Lin, J. Brandt, X. Shen, G. Hua, Efficient boosted exemplar-based face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[29] L. Li, H. Su, Y. Lim, F.F. Li, Objects as attributes for scene classification, in: ECCV Workshops (1), 2010.
[30] L. Li, H. Su, E.P. Xing, F.F. Li, Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: Advances in Neural Information Processing Systems, 2010.
[31] X. Li, Regularized Adaptation: Theory, Algorithms and Applications (Ph.D. thesis), University of Washington, USA, 2007.
[32] J. Luo, T. Tommasi, B. Caputo, Multiclass transfer learning from unconstrained priors, in: Proceedings of the International Conference on Computer Vision, 2011.
[33] T. Malisiewicz, A. Gupta, A.A. Efros, Ensemble of exemplar-SVMs for object detection and beyond, in: Proceedings of the International Conference on Computer Vision, 2011.
[34] A. Opelt, A. Pinz, A. Zisserman, A boundary-fragment-model for object detection, in: Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 2006.
[35] P. Ott, M. Everingham, Shared parts for deformable part-based models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[36] M. Pandey, S. Lazebnik, Scene recognition and weakly supervised object localization with deformable part-based models, in: Proceedings of the International Conference on Computer Vision, 2011.
[37] S.N. Parizi, J. Oberlin, P. Felzenszwalb, Reconfigurable models for scene recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[38] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: Proceedings of the International Conference on Computer Vision, 2007.
[39] M.A. Sadeghi, A. Farhadi, Recognition using visual phrases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[40] A. Shrivastava, T. Malisiewicz, A. Gupta, A.A. Efros, Data-driven visual similarity for cross-domain image matching, ACM Trans. Graph. 30 (6) (2011) 1–10.
[41] S. Singh, A. Gupta, A.A. Efros, Unsupervised discovery of mid-level discriminative patches, in: Proceedings of the European Conference on Computer Vision, 2012.
[42] H.O. Song, S. Zickler, T. Althoff, R.B. Girshick, M. Fritz, C. Geyer, P.F. Felzenszwalb, T. Darrell, Sparselet models for efficient multiclass object detection, in: Proceedings of the European Conference on Computer Vision, 2012.
[43] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[44] M. Stark, M. Goesele, B. Schiele, A shape-based object class model for knowledge transfer, in: Proceedings of the International Conference on Computer Vision, 2009.
[45] J. Tighe, S. Lazebnik, Finding things: Image parsing with regions and per-exemplar detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[46] T. Tommasi, B. Caputo, The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories, in: Proceedings of the British Machine Vision Conference, 2009.
[47] T. Tommasi, F. Orabona, B. Caputo, Safety in numbers: Learning categories from few examples with multi-model knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[48] A. Torralba, K.P. Murphy, W.T. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, 2004.
[49] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classemes, in: Proceedings of the European Conference on Computer Vision, 2010.
[50] Y. Wang, R. Ji, S.F. Chang, Label propagation from imagenet to 3d point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[51] J. Yang, B. Price, S. Cohen, M.H. Yang, Context driven scene parsing with attention to rare classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[52] J. Yang, R. Yan, A.G. Hauptmann, Adapting SVM classifiers to data with shifted distributions, in: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, 2007.
[53] J. Yang, R. Yan, A.G. Hauptmann, Cross-domain video concept detection using adaptive SVMS, in: Proceedings of the 15th international conference on Multimedia, 2007.
[54] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, F.F. Li, Human action recognition by learning bases of action attributes and parts, in: Proceedings of the International Conference on Computer Vision, 2011.