Three Viewpoints Toward Exemplar SVM

Takumi Kobayashi
National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, Japan
takumi.kobayashi@aist.go.jp

Abstract

In contrast to category-level or cluster-level classifiers, exemplar SVM [17] is successfully applied to classifying (or detecting) a target object as well as transferring instance-level annotations. The method, however, is formulated in a highly biased classification problem where only one positive sample is contrasted with a substantial number of negative samples, which makes it difficult to properly determine the regularization parameters balancing two types of costs derived from positive and negative samples. In this paper, we present two novel viewpoints toward exemplar SVM in addition to the original definition. From these proposed viewpoints, we can give light on an intrinsic structure of exemplar SVM, reducing two parameters into only one as well as providing clear intuition on the parameter, in order to free us from exhaustive parameter tuning. We can also clarify how the classifier geometrically works so as to produce homogeneous classification scores of multiple exemplar SVMs which are comparable to each other without calibration. In addition, we propose a novel feature transformation method based on those viewpoints which contributes to general classification tasks. In the experiments on object detection and image classification, the proposed methods regarding exemplar SVM exhibit favorable performance.

1. Introduction

Classification plays a key role in various pattern recognition problems such as image recognition and object detection. In a real world, there are a lot of categories to be classified, and those multi-class problems are usually treated by dividing multiple categories in a *one-vs-rest* manner [5]. Namely, a standard procedure for training multi-class classifiers is to discriminate a specific object category (car, bike, etc.) from the other categories, resulting in class-specific classifiers of which number is equal to that of categories (Fig. 1a). This approach assumes that samples in a category are well described by using a single parametric model, *e.g.*, linear decision boundary, but it might be actually infeasible.

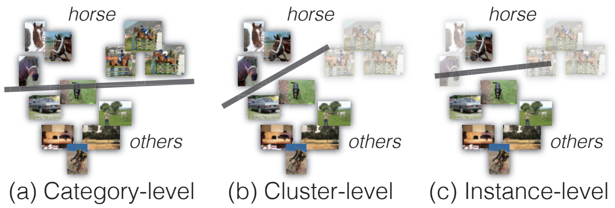

Figure 1. Classification levels. (a) Category-level classification is commonly applied to deal with multi-class problems in a one-vs-rest manner. (b) Cluster-level classifier captures more detailed distribution within the category, and (c) instance-level classification is defined at the finest resolution of classification and provides correspondence to an exemplar.

To alleviate it, samples in a category are further divided into clusters [18] and then the above classification procedure is applied to provide cluster-based representation (Fig. 1b). Clustering metrics are, for example, based on object scale [4] and object view [11]. The clustering-based methods contain a parameter regarding the number of clusters which is generally difficult to be properly determined; a small number of clusters contain large variability within a cluster. In a slightly different approach, poselet method [3] which focuses on detecting people decomposes the category in terms of parts, though requiring exhaustive human effort for labeling parts.

Recently, such direction reaches an extreme classification problem, called *exemplar SVM* [17], in which each instance sample (exemplar) is discriminated from the other samples; that is, a category is divided into samples at the finest resolution (Fig. 1c). In contrast to clustering-based methods, exemplar SVM is a non-parametric method since the number of clusters corresponds to that of samples. The method of exemplar SVM is advantageous in the following two points.

First, exemplar SVM classifier provides correspondence of an input sample to an exemplar, making it possible to directly transfer annotations of the exemplar to the input sample. In contrast, category-based methods can provide only category representative information, usually class label, and

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

even in clustering-based methods, it is hard to convey such instance-level information due to coarse alignment of clusters on samples.

Second, each positive sample is discriminated from plenty of negative samples and thus the information of negative samples is parametrically exploited in a form of classifier without holding samples themselves. Therefore, a substantial number of negatives are effectively utilized for improving discriminative power in exemplar SVM, while in k-NN the number of samples to be kept is limited.

Exemplar SVM is applied in various methods, such as label propagation to point clouds [28], visual similarity learning [1], scene classification [13] and 2D alignment of 3D model [2]. The method, however, has difficulty in determining (tuning) two regularization parameters that balance effects of positive and negative classification costs, due to the highly biased classification problem where only one positive sample is contrasted with a large number of negative samples. Besides, classification scores of multiple exemplar SVMs are required to be calibrated so as to be comparable among an ensemble of heterogeneous classifiers which are individually trained.

In this paper, we present three viewpoints for exemplar SVM; one is original formulation while the other two are novel. From these viewpoints, we can give light on an intrinsic structure of exemplar SVM. Specifically, the regularization parameters, which are hard to be determined in the original exemplar SVM, are reduced into only one parameter whose role is also clearly provided so as to intuitively determine the parameter value. And, we can clarify how the classifier geometrically works, which frees us from carefully calibrating the classification scores. As a result, exemplar SVM can be utilized more simply without careful tuning. In addition, those viewpoints lead to novel feature transformation using exemplar SVM. Although exemplar SVM has been mainly applied to detection tasks so far, the proposed method of feature transformation contributes to improve performance on general classification tasks.

2. Three viewpoints for exemplar-SVM

We begin with reviewing an original definition of exemplar SVM (ex-SVM) [17] and then present two novel viewpoints toward it; these are outlined in Fig. 2. Consequently, we propose a novel formulation of ex-SVM which is more simplified with only one parameter, while original ex-SVM [17] contains two parameters, and gives clearer geometrical intuition to the classifier.

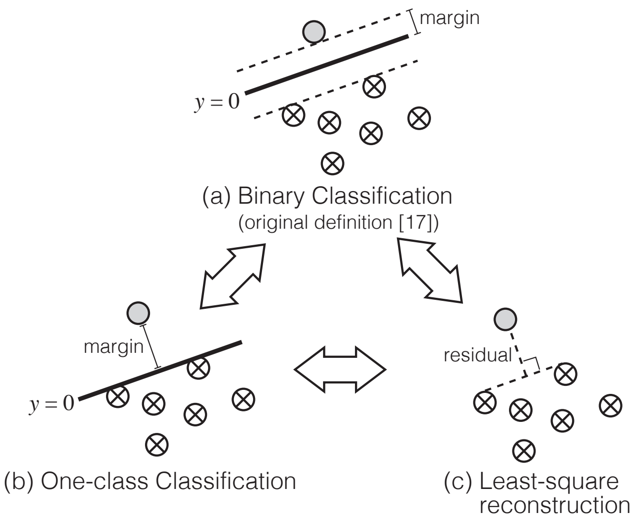
margin
$y = 0$
(a) Binary Classification
(original definition [17])
margin
$y = 0$
residual
(b) One-class Classification
(c) Least-square reconstruction

Figure 2. Three viewpoints for exemplar SVM: (a) Original formulation of large margin binary classification [17], (b) the first novel viewpoint by one-class large margin classifier and (c) the second novel viewpoint by least-square reconstruction. Gray circle indicates the target positive sample, while the others are negative samples.

2.1. Formulation by binary classification

Exemplar SVM (ex-SVM) is originally formulated in [17] as a binary classification problem to discriminate only one target positive sample $\boldsymbol{x} \in \mathbb{R}^d$ from the other negative samples $\{\boldsymbol{\xi}_i\}_{i=1}^N$ (Fig. 2a). It leads to the following large margin formulation in the same manner as standard SVM [25]:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C_p \mathsf{h}(1-\boldsymbol{w}^\top\boldsymbol{x}-b) + C_n \sum_{i=1}^{N} \mathsf{h}(1+\boldsymbol{w}^\top\boldsymbol{\xi}_i+b), \quad (1)$$

where $\mathsf{h}(x) = \max(x,0)$ is a hinge loss function and C_p and C_n are regularization parameters for balancing the positive and negative costs. Ex-SVM is characterized by this highly unbalanced formulation in which only one positive sample is contrasted with plenty of negative samples drawn such as from the categories other than the target one. Therefore, it is required to carefully tune the regularization parameters C_p and C_n.

The above formulation (1) has the following dual.

$$\min_{\hat{\boldsymbol{\alpha}}} \frac{1}{2}\hat{\boldsymbol{\alpha}}^\top \hat{\boldsymbol{K}} \hat{\boldsymbol{\alpha}} - \mathbf{1}^\top \hat{\boldsymbol{\alpha}}, \quad (2)$$

$$s.t.\ \alpha_0 - \sum_{i=1}^{N} \alpha_i = 0,\ 0 \le \alpha_0 \le C_p,\ 0 \le \alpha_i \le C_n, \forall i \ge 1,$$

where $\hat{\boldsymbol{\alpha}} = [\alpha_0, \alpha_1, \cdots, \alpha_N]^\top \in \mathbb{R}^{N+1}$ are Lagrangian multipliers; α_0 is for the positive sample $\boldsymbol{x}$, while α_i $(i \ge 1)$ is for the negative sample $\boldsymbol{\xi}_i$. The matrix $\hat{\boldsymbol{K}} \in \mathbb{R}^{N+1 \times N+1}$ is composed of

$$\hat{\boldsymbol{K}} = \begin{bmatrix} \mathsf{k}(\boldsymbol{x},\boldsymbol{x}) & -\boldsymbol{k}^\top \\ -\boldsymbol{k} & \boldsymbol{K} \end{bmatrix}, \quad (3)$$

where k is a kernel function, $\boldsymbol{k} \in \mathbb{R}^N$ is a kernel vector between $\boldsymbol{x}$ and $\{\boldsymbol{\xi}_i\}_{i=1}^N$, *i.e.*, $k_i = \mathsf{k}(\boldsymbol{x}, \boldsymbol{\xi}_i)$, and $\boldsymbol{K}$ is the kernel Gram matrix of $\{\boldsymbol{\xi}_i\}_{i=1}^N$, *i.e.*, $K_{ij} = \mathsf{k}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$. In the case of linear kernel, an ex-SVM classifier is given by

$$y = \boldsymbol{w}^\top \boldsymbol{z} + b, \tag{4}$$

where b is a bias determined so as to give equal margin for positive and negative boundaries (Fig. 2a), and by using the optimizers $\{\alpha_i^*\}_{i=0}^N$ in (2), the weight vector $\boldsymbol{w}$ is given by

$$\boldsymbol{w} = \alpha_0^* \boldsymbol{x} - \sum_{i=1}^N \alpha_i^* \boldsymbol{\xi}_i. \tag{5}$$

In what follows, we assume linear kernel for simplicity.

2.2. Formulation by one-class SVM

Next, we reformulate ex-SVM from a novel viewpoint of one-class SVM [22] by rewriting the dual problem (2). Since (2) is convex, it is equivalent to

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} - \alpha_0^* \boldsymbol{k}^\top \boldsymbol{\alpha} + \frac{1}{2}\alpha_0^{*2}\mathsf{k}(\boldsymbol{x},\boldsymbol{x}) - 2\alpha_0^*, \tag{6}$$
$$s.t. \sum_{i=1}^N \alpha_i = \alpha_0^*,\ 0 \le \alpha_i \le C_n,$$

where α_0 is *fixed* to its optimizer α_0^*. By ignoring the constant term regarding α_0^*, this is further rewritten to

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^\top \bar{\boldsymbol{K}}\boldsymbol{\alpha},\ \ s.t. \sum_{i=1}^N \alpha_i = \alpha_0^*,\ 0 \le \alpha_i \le C_n, \tag{7}$$

where $\bar{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{1}\boldsymbol{k}^\top - \boldsymbol{k}\boldsymbol{1}^\top + \mathsf{k}(\boldsymbol{x},\boldsymbol{x})\boldsymbol{1}\boldsymbol{1}^\top$ which is the kernel Gram matrix *centered at* $\boldsymbol{x}$; in the case of linear kernel, $\bar{K}_{ij} = (\boldsymbol{\xi}_i - \boldsymbol{x})^\top(\boldsymbol{\xi}_j - \boldsymbol{x})$. For more simplicity, we rescale Lagrangian multipliers $\boldsymbol{\alpha}$ by α_0^*[1] and consequently obtain the final form of

$$\min_{\bar{\boldsymbol{\alpha}}} \frac{1}{2}\bar{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{K}}\bar{\boldsymbol{\alpha}},\ \ s.t. \sum_{i=1}^N \bar{\alpha}_i = 1,\ 0 \le \bar{\alpha}_i \le \frac{C_n}{\alpha_0^*} = C. \tag{8}$$

This is exactly the same as the dual of one-class SVM (oc-SVM) [22] with a regularization parameter C. Thus, we can insist that exemplar SVM is intrinsically reduced to one-class SVM in the feature space centered at the target sample $\boldsymbol{x}$ as shown in Fig. 2b. From this viewpoint, the classifier is optimized so as to maximize a margin from the positive sample $\boldsymbol{x}$ to the negative samples at boundary since the primal problem for (8) is formulated as

$$\min_{\boldsymbol{w},\rho} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \rho + C\sum_i \mathsf{h}(\rho - \boldsymbol{w}^\top(\boldsymbol{\xi}_i - \boldsymbol{x})), \tag{9}$$

where a margin is measured as the distance $\frac{\rho}{\|\boldsymbol{w}\|_2}$ from the origin to the boundary samples $\boldsymbol{\xi}_i - \boldsymbol{x}$ of $\boldsymbol{w}^\top(\boldsymbol{\xi}_i - \boldsymbol{x}) - \rho = 0$.

This formulation for ex-SVM gives a classifier by

$$y = \boldsymbol{w}^\top \boldsymbol{z} - \rho, \tag{10}$$

where ρ is determined such that $\boldsymbol{w}^\top \boldsymbol{\xi}_i - \rho = 0$ for $\{i | 0 < \bar{\alpha}_i^* < C\}$, and

$$\boldsymbol{w} = -\sum_i^N \bar{\alpha}_i^* (\boldsymbol{\xi}_i - \boldsymbol{x}) = \frac{1}{\alpha_0^*}(\alpha_0^* \boldsymbol{x} - \sum_i^N \alpha_i^* \boldsymbol{\xi}_i), \tag{11}$$

where we use $\bar{\alpha}_i^* = \frac{\alpha_i^*}{\alpha_0^*}$ and $\sum_i^N \bar{\alpha}_i^* = 1$. Note that the direction of $\boldsymbol{w}$ is opposite to ordinary oc-SVM for consistency with ex-SVM that discriminates $\boldsymbol{x}$ from $\{\xi_i\}_{i=1}^N$. This form corresponds to that of ex-SVM (5) except for the scaling by $\frac{1}{\alpha_0^*}$ which does not affect classification. Thus, we can say that ex-SVM (1) and oc-SVM (9) produce the same classifier except for the scaling. Due to clear geometrical intuition that a margin is measured from the boundary negative samples, we employ (10) as an ex-SVM classifier.

This insight into ex-SVM is also practically useful in the following three points regarding parameter issues.

1. Two parameters C_p and C_n required in the original ex-SVM formulation are reduced to only one parameter C.
2. The paramter C is limited in $\frac{1}{N} \le C \le 1$ according to the constraints $\sum_{i=1}^N \bar{\alpha}_i = 1, \bar{\alpha}_i \ge 0$.
3. We can clarify an intuitive role of the parameter C; the number of support vectors (outliers) is controlled by C [22] and is greater than $\frac{1}{C}$[2]. Thus, in the case that negative samples are all drawn from the categories other than the target one, we can simply set $C = 1$ assuming minimum outliers.

2.3. Formulation by least squares

Finally, we show the third viewpoint for exemplar SVM in the framework of least squares. We consider a problem to reconstruct $\boldsymbol{x}$ by using (restricted) convex combination of $\{\boldsymbol{\xi}_i\}_{i=1}^N$;

$$\boldsymbol{x} \approx \sum_{i=1}^N \bar{\alpha}_i \boldsymbol{\xi}_i,\ \ s.t. \sum_{i=1}^N \bar{\alpha}_i = 1,\ 0 \le \bar{\alpha}_i \le C. \tag{12}$$

[1] Note that $\alpha_0^* > 0$ since $\alpha_0^* = 0$ leads to $\alpha_i^* = 0, \forall i$, which obviously violate KKT conditions.

[2] In oc-SVM [22], the regularization parameter is givecn by $C = \frac{1}{N\nu}$ where ν is the ratio of support vectors among N samples.

The optimum coefficients $\bar{\boldsymbol{\alpha}}^*$ are obtained by applying the following least-square method;

$$\min_{0\leq\bar{\boldsymbol{\alpha}}\leq C} \frac{1}{2}\|\boldsymbol{x} - \sum_{i=1}^N \bar{\alpha}_i\boldsymbol{\xi}_i\|_2^2, \ \ s.t. \ \sum_{i=1}^N \bar{\alpha}_i = 1, \tag{13}$$

$$\Leftrightarrow \min_{0\leq\bar{\boldsymbol{\alpha}}\leq C} \frac{1}{2}\bar{\boldsymbol{\alpha}}^\top \boldsymbol{K}\bar{\boldsymbol{\alpha}} - \boldsymbol{k}^\top\bar{\boldsymbol{\alpha}} + \mathsf{k}(\boldsymbol{x},\boldsymbol{x}), \ s.t. \ \sum_{i=1}^N \bar{\alpha}_i = 1, \tag{14}$$

$$\Leftrightarrow \min_{0\leq\bar{\boldsymbol{\alpha}}\leq C} \frac{1}{2}\bar{\boldsymbol{\alpha}}^\top \bar{\boldsymbol{K}}\bar{\boldsymbol{\alpha}}, \ \ s.t. \ \sum_{i=1}^N \bar{\alpha}_i = 1. \tag{15}$$

In the case of $C = 1$, $\sum_{i=1}^N \bar{\alpha}_i\boldsymbol{\xi}_i$ covers whole convex hull of $\{\boldsymbol{\xi}_i\}_{i=1}^N$ and (15) provides orthogonal projection onto the convex hull (Fig. 2c). The primal problem (15) is the same quadratic programming (QP) as (8) and the residual vector $\boldsymbol{x} - \sum_i \bar{\alpha}_i^*\boldsymbol{\xi}_i$ is equivalent to the classifier vector $\boldsymbol{w}$ of oc-SVM (11) and ex-SVM (5) except for the scaling as mentioned above. Thus, note that this least-square formulation is connected to ex-SVM via the dual problem in oc-SVM.

The least-square formulation is also found in (unsupervised) similarity metric learning methods [27, 15]. Based on a locally linear assumption as in LLE [20], those methods employ the model to approximate a sample vector by non-negative convex construction of the other samples. They are different from the least-square formulation for ex-SVM in that the optimized non-negative coefficients $\bar{\boldsymbol{\alpha}}_i^*$ are utilized to construct similarity measure between $\boldsymbol{x}$ and $\boldsymbol{\xi}_i$, while we make use of the residual vector for an ex-SVM classifier. In that sense, the similarity learning methods and ex-SVM are complementary to each other, through the identical QP formulation.

2.4. Summary

We have shown that in exemplar SVM, maximizing standard SVM margin (Sec. 2.1) corresponds to maximizing a margin from the target positive sample (Sec. 2.2), and even to *minimizing* reconstruction residual for the positive sample in a least-square manner (Sec. 2.3). From the oc-SVM viewpoint, parameters are reduced to only one, C, which is inherently restricted into $C \in [\frac{1}{N}, 1]$ (we usually set $C = 1$) and controls the amount of support vectors (outliers). This significantly reduces exhaustive parameter tuning. From the least-square viewpoint, we can give geometrical interpretation to an ex-SVM classifier as well as to the Lagrangian multipliers (as *similarity*), which leads to propose novel feature transformation in Sec. 4. These three viewpoints are shown in Fig. 2.

At the last, we refer to an extreme case of $C = \frac{1}{N}$. In that case, the Lagrangian multipliers simply result in $\bar{\alpha}_i^* = \frac{1}{N}$, $\forall i$, and the classifier vector $\boldsymbol{w}$ is described as

$$\boldsymbol{w} = \boldsymbol{x} - \frac{1}{N}\sum_{i=1}^N \boldsymbol{\xi}_i, \tag{16}$$

which is regarded just as subtracting the mean of negative samples. On the other hand, LDA version of exemplar SVM is also proposed for computational efficiency [2] by applying two-class linear discriminant analysis (LDA) [7];

$$\boldsymbol{w} = \boldsymbol{S}_W^{-1}(\boldsymbol{x} - \frac{1}{N}\sum_{i=1}^N \boldsymbol{\xi}_i), \tag{17}$$

where $\boldsymbol{S}_W$ is a within-class scatter matrix. This is the same form as (16) except for the whitening projection via $\boldsymbol{S}_W^{-1}$. In other words, exemplar SVM in such a whitened space is exactly the same as LDA version. Thus, the original and LDA version of exemplar SVM are viewed in a unified manner in the proposed framework.

In the followings, we apply exemplar SVM to detection and classification tasks based on these novel viewpoints.

3. Object detection by ensemble of exemplar SVMs

We first apply an ensemble of exemplar SVM classifiers to object detection as presented in [17] where exemplar SVMs operate on respective object instances to form an ensemble of classifiers in total. Each exemplar SVM is trained by using only one positive object image (bounding box) and plenty of negative images belonging to the other object categories or background. Thus, as stated in Sec. 2.2, the parameter C is simply set to $C = 1$.

3.1. Exemplar SVM for L_2-normalized feature vectors

In this task, we assume that feature vectors are normalized in a unit L_2 norm, which does not so lack generality since such normalization is commonly applied in various features such as HOG features [6].

We also normalize $\boldsymbol{w}$ in a unit L_2 norm with accordingly rescaling the bias ρ;

$$y = \frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_2}\boldsymbol{z} - \frac{\rho}{\|\boldsymbol{w}\|_2} = \hat{\boldsymbol{w}}^\top\boldsymbol{z} - \hat{\rho}, \tag{18}$$

in which classification score y corresponds to distance from the classification hyperplane. The L_2-normalized feature vectors are located on a unit hypersphere, and thus the distance y is bounded in $-2 \leq y \leq 2$. Such distance has geometrically homogeneous meaning across any exemplar SVM classifiers and can be directly comparable; the larger distance (higher classification score) insists more confidently that the sample is far from negative samples

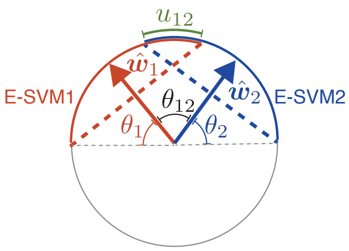

Figure 3. Overlap between two exemplar SVMs.

(classes), that is, belonging to the positive class. As a result, it is not required to calibrate classification scores of ensemble ex-SVMs [17].

3.2. Similarity measure for integrating multiple ex-SVMs

To perform detection task, it is necessary to integrate multiple ex-SVM classifiers pointing the identical object category since ex-SVM classifiers of visually similar object instances would co-occur at certain object regions in an image. For that purpose, we define similarity measure between ex-SVM classifiers. The ex-SVM classifier (18) cuts a unit hypersphere and we measure overlap between those cut sub-region on a hypersphere. Cutting planes by two ex-SVM classifiers are shown in Fig. 3. Along the arc between two classifier vectors $\hat{\boldsymbol{w}}_1$ and $\hat{\boldsymbol{w}}_2$, the overlap (length) u_{12} is given by

$$u_{12} = \max[\min(\theta_1, \theta_{12} + \theta_2) - \max(-\theta_1, \theta_{12} - \theta_2), 0], \tag{19}$$

where as shown in Fig. 3, θ_1 and θ_2 indicate the spread angle of respective ex-SVM classifiers (computed by $\arccos(\hat{\rho})$) and $\theta_{12} = \arccos(\hat{\boldsymbol{w}}_1^\top \hat{\boldsymbol{w}}_2)$ is the angle between $\hat{\boldsymbol{w}}_1$ and $\hat{\boldsymbol{w}}_2$. Similarity measure is defined by the ratio of the overlap (intersection) compared to the union of two classifier spreads;

$$S_{ij} = \frac{u_{12}}{2\theta_1 + 2\theta_2 - u_{12}}. \tag{20}$$

Note that, in the case that respective exemplar SVMs are constructed from HOG features of different size (dimensionality), the angle θ_{12} can not be directly computed by the inner product due to inconsistency of dimensionality. In this study, we roughly compute it as the maximum value of the convolution of $\hat{\boldsymbol{w}}_1$ and $\hat{\boldsymbol{w}}_2$[3].

The candidate windows for the target object are detected via thresholding classification scores by 0. At a window detected by the i-th ex-SVM classifier, we can accumulate the scores of overlapped windows detected by the other ex-SVM classifiers, resulting in the context feature [3] denoted by $\boldsymbol{v}_i \in \mathbb{R}^M$ where M denotes the number of exemplars in a category of interest. The final score is computed by using similarity matrix $\boldsymbol{S}$ (20);

$$y_{fin} = (\boldsymbol{s}_i + \epsilon)^\top \boldsymbol{v}_i, \tag{21}$$

where $\boldsymbol{s}_i$ is the i-th column vector of $\boldsymbol{S}$ and ϵ is a bias, set to $\epsilon = 1$ in this experiment.

3.3. Experiments on VOC2007 detection dataset

We evaluated performance of exemplar SVM detectors in the proposed formulation on PASCAL VOC2007 object detection dataset [8]. The task is to detect objects of 20 categories (car, bike, etc) in 4,952 images, while the detectors are trained on 5,011 images containing 12,608 object instances (exemplars), each of which works as a positive sample to produce an ex-SVM detector.

Experimental setup. Note again that the exemplar SVM classifier in the proposed formulation takes a decision boundary ($y = 0$) at the negative support vectors (Fig. 2b) with rescaling in (18). We employ two types of features, HOG features [6] and deep-learning (DL) based features [9] using `CAFFE` network (`fc6` layer) [12] trained on ImageNet dataset. Detectors work in manners of sliding windows for HOG features and selective search [23] for deep-learning based features; for more details, refer to [9]. For each exemplar SVM detector, the detected windows that exhibit positive score $y > 0$ are fed to construct context feature $\boldsymbol{v}$ mapped into the final score y_{fin} via similarity-based integration (Sec. 3.2). Finally, non-maximum suppression is applied to those detected windows with the score y_{fin}.

Performance results. Table 1 shows detection performance measured by mean average precision (mAP) rate according to VOC evaluation protocol.

As to HOG features, we show at the first row the results reported in [17] using the original exemplar SVM detection which are compared with those of the proposed method at the second row. For reference, we exclude the process to integrate multiple ex-SVM scores (Sec. 3.2) from our proposed method; that is, the classification score y of each ex-SVM is directly passed to the final score, $y_{fin} = y$. The performance of the degraded method is shown at the third row. The proposed method produces favorable performance by effectively integrating ex-SVM scores via the similarity measure which is computed based only on ex-SVM classifiers, and it is slightly superior to the original one. It should be noted that for this detection task, the proposed method regarding classification does not have any parameters to be tuned; the only one regularization parameter is set to $C = 1$ from the oc-SVM viewpoint, and neither calibrating classification score nor empirically constructing similarity matrix $\boldsymbol{S}$ is required unlike the original ex-SVM detection [17].

As to deep-learning (DL) features, the proposed method (at the fifth row) again improves performance of the method

[3] The HOG features, consequently classifier vectors, are formulated in the array of [*feature dimension* × *height* × *width*], on which convolution can be applied.

feature	methods	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mean
HOG	original ex-SVM[17]	.208	.480	.077	.143	.131	.397	.411	.052	.116	.186	.111	.031	.447	.394	.169	.112	.226	.170	.369	.300	.227
	Ours	.290	.513	.097	.165	.103	.356	.461	.059	.127	.214	.104	.100	.478	.398	.185	.100	.262	.145	.182	.377	**.236**
	Ours w/o integration	.192	.408	.094	.112	.096	.268	.387	.098	.100	.142	.049	.097	.404	.323	.137	.049	.150	.108	.153	.259	.181
DL	SVM	.508	.600	.413	.316	.294	.569	.563	.447	.298	.515	.475	.454	.519	.559	.375	.259	.453	.411	.528	.552	.455
	Ours	.552	.669	.533	.386	.356	.633	.655	.474	.285	.536	.523	.508	.622	.635	.476	.271	.529	.443	.597	.606	**.514**
	Ours w/o integration	.494	.617	.390	.292	.261	.501	.612	.332	.180	.397	.282	.375	.561	.551	.342	.187	.428	.275	.513	.557	.407

Table 1. Detection performance on VOC 2007 dataset.

without integrating scores (the sixth row). The proposed method also significantly outperforms a linear SVM detector (the fourth row).

4. Feature transformation for classification

We then apply exemplar SVM to transform feature vectors for enhancing discriminative power, which is novel application of exemplar SVM to our best knowledge.

4.1. Exemplar SVM as feature transformation

Exemplar SVM provides classifiers specific to respective samples. Each classifier is designed to maximize the difference of the target (positive) sample from negatives as described in Sec. 2.2. Thus, by transforming the sample so that the classification score is increased, we can obtain a feature vector of more discriminative power. Such transformation should be regularized based on the manifold on which the feature vectors are actually distributed. In the case of L_2 normalized features forming a manifold of a unit hypersphere, we can simply transform a feature vector $\boldsymbol{x}$ by maximizing the score on the manifold as shown in Fig. 4;

$$\arg \max_{\hat{\boldsymbol{x}} |\ \|\hat{\boldsymbol{x}}\|_2 = 1} \boldsymbol{w}^\top \hat{\boldsymbol{x}} - b \ \Rightarrow \boldsymbol{x} \mapsto \hat{\boldsymbol{x}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \qquad (22)$$

where $(\boldsymbol{w}, b)$ is an exemplar SVM classifier model for $\boldsymbol{x}$. The samples that have similar ex-SVM classifiers are transformed into closer points (Fig. 4).

We can also give another interpretation to this transformation. From the least-square viewpoint in Sec. 2.3, exemplar SVM approximates the target positive sample $\boldsymbol{x}$ by using the negative samples $\{\boldsymbol{\xi}_i\}_{i=1}^N$. Thus, the reconstructed vector $\sum_{i=1}^N \bar{\alpha}_i \boldsymbol{\xi}_i$ shares some sort of patterns with $\boldsymbol{x}$; for example, the positive sample and the negatives may share some background patterns. By subtracting such common pattern from the target, discriminative feature patterns are enhanced; the subtracted feature vector, $\boldsymbol{x} - \sum_{i=1}^N \bar{\alpha}_i \boldsymbol{\xi}_i = \boldsymbol{w}$, corresponds to (22) by L_2 normalization.

4.2. Mining negative samples

The above transformation requires *negative* samples of which categories are different from that of the target sample.

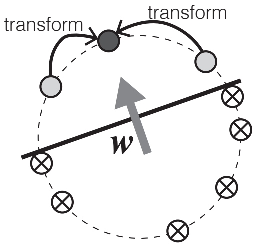

Figure 4. Feature transformation by exemplar SVM.

Since we deal with unknown test samples whose categories are not given, the problem is how to draw the negative samples. We consider three approaches as follows.

`Irrelevant dataset`. The first approach is to prepare another dataset that is *irrelevant* to the classification target. For example, in the case of *object* classification, we can utilize *scene* dataset as a source of negative samples. For any samples to be classified, negative samples are drawn from such another dataset.

`Smaller C`. In this approach, training dataset is regarded as a source of negative samples. The feature transformation (22) actually demands support vectors to contain negative samples of different categories from that of the target sample. While larger C produces a small number of support vectors which would be dominated by the samples of the same category, we can enforce C to be sufficiently small so that negative samples of different categories likely contribute as support vectors. Let m be the number of samples in the same category, and $C < \frac{1}{m}$ works for that purpose.

`Two-pass`. This approach mines negative samples from training dataset more aggressively by excluding similar categories to that of the target sample. For evaluating such similarity, we can also utilize exemplar SVM as similarity learning described in Sec. 2.3. First of all, exemplar SVM with moderately small C is applied to the input sample and the whole training dataset which we regard as negatives. Then, the class categories that (most of) support vectors belong to are excluded. The input feature vector is subsequently transformed by again applying exemplar SVM with $C = 1$ to the training samples of the remaining categories as negatives. The algorithm is shown in Algorithm 1;

Algorithm 1 : Feature transformation by `two-pass` ex-SVM

```
Input: $\boldsymbol{x}$: input feature vector ($\|\boldsymbol{x}\|_2 = 1$) to be transformed,
    $\{\boldsymbol{\xi}_i, c_i\}_{i=1}^N$: training samples with category label $c \in \{1, \cdots, M\}$,
    $C_1, \tau$: parameters
[1st pass] Apply ex-SVM with $C = C_1$ in (8) to $\boldsymbol{x}$ and $\{\boldsymbol{\xi}_i\}_{i=1}^N$ for producing Lagrangian multipliers $\{\bar{\alpha}_i\}_{i=1}^N$.
Accumulate $\bar{\alpha}$ over categories by $h_c = \sum_{c_i = c} \bar{\alpha}_i$, $\forall c$.
Sort $\{h_c\}_{c=1}^M$ in descending order and let $\{\Phi_c\}_{c=1}^M$ be the sorted category indexes, $\Phi_c \in \{1, \cdots, M\}$, $\forall c$.
Pick up dominant categories $\{\Phi_c\}_{c=1}^{\bar{M}}$
    where $\bar{M} = \min c', \ s.t. \ \frac{\sum_{c=1}^{c'} h_{\Phi_c}}{\sum_{c=1}^{M} h_c} \geq \tau$.
Exclude the categories $\{\Phi_c\}_{c=1}^{\bar{M}}$ from $\{1, \cdots, M\}$ and let $\mathcal{P}$ be the index set of the remaining categories.
[2nd pass] Apply ex-SVM with $C = 1$ in (8) to $\boldsymbol{x}$ and $\{\boldsymbol{\xi}_i\}_{c_i \in \mathcal{P}}$ for producing classifier vector $\boldsymbol{w}$.
Output: $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \in \mathbb{R}^d$: transformed feature vector in (22).
```

parameters C_1, τ are empirically set to $C_1 = 0.2, \tau = 0.8$ in this study. The first pass exemplar SVM works on computing the similarity from the target sample (Sec. 2.3). This first screening excludes the categories which are close to $\boldsymbol{x}$, possibly containing the correct category of $\boldsymbol{x}$. Then, the second exemplar SVM actually transforms the feature vector on the basis of the remaining (negative) categories.

Note that the latter two approaches operate within the training dataset, while the first one relies on other datasets.

4.3. Experiments on image classification

We tested the proposed method on image classification tasks using Caltech-256 [10] for object recognition, CUB-200-2011 [26] for fine-grained bird classification, MIT-67 [19] for indoor scene classification.

Experimental setup. We employ Fisher kernel features [21] applied to SIFT descriptors [16] transformed by the method in [14]. Gray-scale SIFT is used in Caltech-256 and MIT-67, while color (hsv) SIFT [24] is applied in CUB-200-2011. The proposed method transforms the Fisher kernel features extracted from an image; for more details in feature extraction, refer to [14]. Linear SVM is applied to the transformed feature vectors for classification. In Caltech-256, we randomly draw 60 training samples per category and it is repeated three times, while in CUB-200-2011 and MIT-67 we use the given training/test split.

Experimental results. We analyze three proposed methods described in Sec. 4.2 basically in comparison to the original features without transformation.

`Irrelevant dataset`. Scene dataset (MIT-67) is utilized as a source of negative samples for object recognition in Caltech-256 and CUB-200-2011, and vice versa. As shown in Table 2, performance is not so improved, compared to the original features. This is because characteristics of negative samples would be far away from the target categories to be classified and the discriminative power of features would not be enhanced; in this case, *object* images might be totally different from *scene* images. This result suggests that for effective feature transform, it is required to draw negative samples from a similar domain to the target categories.

`Smaller C`. The regularization parameter C is varied on the basis of averaged number of samples per category which is denoted by $\bar{m}$: $C \in \{\frac{2}{\bar{m}}, \frac{1}{\bar{m}}, \cdots, \frac{1}{32\bar{m}}\}$. Note that the number of support vectors is greater than $\frac{1}{C}$. Thus, at smaller C, the support vectors contain negative samples of different categories. Table 4a shows that as expected, the smaller C improves performance; $C = \frac{1}{8\bar{m}}$ produces favorable performance in these datasets. In contrast, the larger C degrades performance which is inferior even to the original features since the support vectors are dominated by the positive samples of the same category. Discriminative power is enhanced by comparing to counter-category samples (so-called negative samples). This method, however, has an issue regarding computation time; for smaller C producing larger number of support vectors, the computation time[4] to solve QP (8) becomes larger as shown in Table 4b.

`Two-pass`. In this method, the first-pass ex-SVM mines similar samples and categories to the target sample which are subsequently excluded in the second pass. So detected categories are expected to contain the true category of the target sample for successfully enhancing discriminativity power by the second ex-SVM. Fig. 5 shows quality of the detected categories by first-pass ex-SVM in terms of precision and recall on the training set by varying a threshold parameter τ on Caltech-256. As expected, performance is improved at higher recall rate where the true category of the target positive sample is likely to be excluded in the second pass, contributing to improve discriminative power. Thus, we can say that higher recall is preferable compared to higher precision for this feature transformation. Since the first-pass ex-SVM provides similarities from the target sample to the categories, we can also perform classification based on the similarity in a manner similar to k-NN. The performance results in Table 3a show that only the first pass is not enough for achieving higher classification performance; performance by similarity-based classification is inferior even to the original ones. The second-pass ex-SVM significantly improves it, producing favorable performance. In addition, the computation time for `two-pass` ex-SVM is quite small as shown in Table 3b; it is faster than `smaller C` method of $C = \frac{1}{8\bar{m}}$.

As a result, for feature transformation, the method of `two pass` ex-SVM is superior to the other two methods in terms of classification accuracy and computation time.

[4] It is measured on Xeon 3.4GHz PC with Matlab and Mex-C.

	original feature [14]	irrelevant dataset	
Caltech-256	57.4±0.4	57.3±0.2	(MIT-67)
CUB-200-2011	28.7	28.9	(MIT-67)
MIT-67	63.4	64.0	(Caltech-256)

Table 2. Classification performance by irrelevant dataset approach. The dataset from which negative samples are drawn is shown in the parentheses.

	two pass	Similarity based k-NN
Caltech-256	58.3±0.4	50.2±0.4
CUB-200-2011	30.0	16.6
MIT-67	64.8	58.0

(a) Accuracy (%)

	two pass
Caltech-256	4.83
CUB-200-2011	1.39
MIT-67	1.92

(b) Time per sample (msec)

Table 3. Classification performance and computation time by two-pass approach.

	$C = \frac{1}{\bar{m}} \times$	2	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
Caltech-256	$(\bar{m} = 60)$	49.9±0.2	53.8±0.4	56.2±0.4	57.3±0.4	57.7±0.4	57.5±0.4	57.6±0.4
CUB-200-2011	$(\bar{m} = 30)$	26.7	27.9	29.1	29.4	29.9	29.8	29.3
MIT-67	$(\bar{m} = 80)$	54.8	59.3	62.1	64.4	64.3	64.2	64.2

(a) Accuracy (%)

	$C = \frac{1}{\bar{m}} \times$	2	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
Caltech-256	$(\bar{m} = 60)$	1.56	2.11	3.34	5.98	11.75	42.82	65.77
CUB-200-2011	$(\bar{m} = 30)$	0.49	0.57	0.74	1.18	2.10	3.95	7.78
MIT-67	$(\bar{m} = 80)$	0.76	0.96	1.50	2.75	5.15	9.43	40.26

(b) Time per sample (msec)

Table 4. Classification performance and computation time by smaller C approach. $\bar{m}$ indicates the averaged number of sample per category.

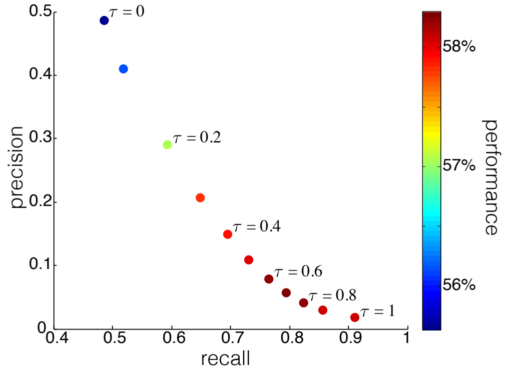

Figure 5. Precision-recall curve for measuring the quality of the first-pass exemplar SVM on Caltech-256. Performance is shown in pseudo colors. This figure is best viewed in color.

5. Conclusion

We have presented two novel viewpoints toward exemplar SVM [17] in addition to the original definition; one is related to one-class SVM and the other is based on least-square reconstruction. From these novel viewpoints, regularization parameters of exemplar SVM are reduced into only one with clear intuition, which frees us from exhaustively tuning parameters. Geometrical interpretation is given not only to ex-SVM classifier but also to Lagrangian multipliers which correspond to support vector coefficients. These viewpoints also lead to a novel framework of ex-SVM detection as well as to a novel feature transformation method. In the experiments on VOC2007 detection and image classification using three datasets, the proposed methods exhibit favorable performance.

In particular, the proposed feature transformation method is so general that our future work includes to apply it to various types of features.

References

[1] O. Aghazadeh, H. Azizpour, J. Sullivan, and S. Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision (ECCV)*, pages 115–128, 2012.

[2] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs : exemplar part-based 2d-3d alignment using a large dataset of cad models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3769, 2014.

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, pages 168–181, 2010.

[4] Y. Chen, X. Zhou, and T. S. Huang. One-class svm for learning in image retrieval. In *International Conference on Image Processing (ICIP)*, pages 34–37, 2001.

[5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2001.

[8] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[11] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *European Conference on Computer Vision (ECCV)*, pages 408–421, 2010.

[12] Y. Jia, E. Shelhamer, J. Donahuef, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 923–930, 2013.

[14] T. Kobayashi. Dirichlet-based histogram feature transform for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3278–3285, 2014.

[15] T. Kobayashi and N. Otsu. Efficient similarity derived from kernel-based transition probability. In *European Conference on Computer Vision (ECCV)*, pages 371–385, 2012.

[16] D. Lowe. Distinctive image features from scale invariant features. *International Journal of Computer Vision*, 60:91–110, 2004.

[17] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *International Conference on Computer Vision (ICCV)*, pages 89–96, 2011.

[18] E. Ohn-Bar and M. Trivedi. Fast and robust object detection using visual subcategories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 179–184, 2014.

[19] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 413–420, 2009.

[20] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[21] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[22] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[23] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[24] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[25] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[26] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[27] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):55–67, 2008.

[28] Y. Wang, R. Ji, and S. Chang. Label propagation from imagenet to 3d point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3135–3142, 2013.