Classification with Global, Local and Shared Features

Hakan Bilen* Vinay P. Namboodiri † Luc Van Gool*
*ESAT-PSI-VISICS/IBBT,
Katholieke Universiteit Leuven, Belgium
firstname.lastname@esat.kuleuven.be
† Alcatel-Lucent Bell Labs
Antwerp, Belgium
vinay.namboodiri@alcatel-lucent.com

1. Introduction

In this paper, we consider the classification problem of deciding whether one of a number of pre-specified categories is present in an image. We show that jointly learning and localizing pairwise relations between classes improves such classification: when having to tell whether or not a specific *target class* is present, sharing knowledge about other, *auxiliary classes* supports this decision. In particular, we propose a framework that combines target class-specific global and local information with information learnt for pairs of the target class and each of a number of auxiliary classes. Adding such pairwise information helps to learn the common part and context for a class pair and discriminate against other classes. For instance, in the case of flower species recognition, a flower type with pink and yellow petals, and another one with white and yellow petals can share the common yellow petals. The target class-specific models rather focus on the specific pink and yellow parts that are needed to discriminate between the pair. Our approach is especially suitable for the fine-grained classification domains where classes are closely similar to one other, e.g. bird, flower species recognition, and classification is hard. This abstract is based on our paper [2] where the method is shown in more detail with additional results.

In summary, our approach combines information about:

1. global image appearance, using a spatial pyramid over the image, thereby providing context information;

2. local appearance, based on a target class-specific window, loosely corresponding to a bounding box;

3. shared appearances, based on a series of windows, each jointly defined for the target class and one of the auxiliary classes that shares visual commonalities.

We show that all components of this combined representation can be learnt jointly, with as only supervision the class label for the training images (i.e. which target class appears in the images without any information on its location).

We have evaluated our approach for flower species and human interaction classification tasks using standard benchmarks. We have experimentally evaluated each of these components individually and jointly for solving these various problems. The results show that adding the shared component is beneficial in all cases.

The central contribution of this paper is the use of local appearance properties that are shared by pairs of classes. Shared representation and transfer learning have been showed to enhance classification accuracy [6, 3]. [6] and [3], similarly to our work, learn to localize and classify jointly. While [6] and [3] require annotation of attribute labels and user feedback to discover meaningful attributes resp, our method assumes only image-level class labels and automatically learns discriminative shared appearance without requiring any semantic attribute.

2. Model Definition

To build our classifiers, we make use of the latent structural SVM (LSSVM) formulation with latent parameters [8]. In our model, input $x \in \mathcal{X}$, output $y \in \mathcal{Y} = \{c_1, \cdots, c_k\}$ and latent parameters $h \in \mathcal{H}$ correspond to the image, its label, and a set of image windows, resp. We use discriminant functions of the form $f_\theta : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathcal{R}$ which scores triplets of (x, y, h) for a learnt parameter vector θ of the LSSVM model as

$$f_\theta(x, y, h) = \theta^y \cdot \Psi^y(x, y, h) \tag{1}$$

where $\Psi^y(x, y, h)$ is a joint feature vector that describes the relation among x, y and h. In our model, each $\Psi^y(x, y, h)$ concatenates histograms which are obtained from multiple rectangular windows with the bag of words (BoW) representation with a spatial pyramid (SP) of [7]. We use different windows to encode the 3 information channels, *i.e.* global, local, and shared. We can write our feature vector for class y as $\Psi^y(x, y, h) = \left(\Psi^y_{gl}, \Psi^y_{loc}, \Psi^y_{sh,c_1}, \cdots, \Psi^y_{sh,c_k}\right)$.

A graphical illustration of our model for a toy object classification task is shown in Fig.1. The images x_1, x_2

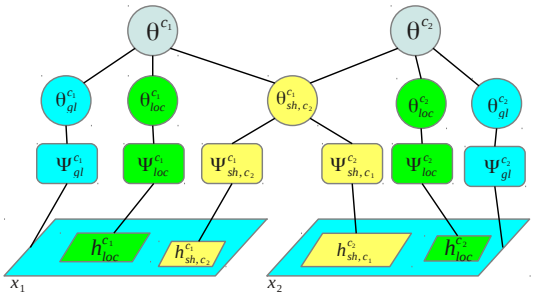

Figure 1. Illustration of our model for two images containing one target class each. Different features are denoted in different colors.

are labeled as c_1, c_2 resp. While there are separate class-specific parameter vectors for the global $\theta^{c_1}_{gl}, \theta^{c_2}_{gl}$ and local $\theta^{c_1}_{loc}, \theta^{c_2}_{loc}$ channels, an identical parameter vector $\theta^{c_1}_{sh,c_2}$ is shared between the labels c_1 and c_2.

Global Features: $\Psi^y_{gl} = \phi(x)$, shown in cyan color in Fig.1, is the SP representation over the whole image x.

Local Features: $\Psi^y_{loc} = \phi(x, h^y_{loc})$, shown in green color in Fig.1, is a SP over an image part selected with window h^y_{loc}, which roughly corresponds to a bounding box h^y_{loc} around the instance of the target class.

Shared Features: $\Psi^y_{sh,\hat{y}} = K_{\mathcal{S}}(y,\hat{y})\phi(x, h^y_{sh,\hat{y}})$, shown in yellow color in Fig.1, is a SP over a window $h^y_{sh,\hat{y}}$. $\mathcal{S}$ is the set of all class pairs of the target class (y) and each one of the auxiliary classes $(\hat{y})$ with which the target class is supposed to share information. $K_{\mathcal{S}}(y,\hat{y})$ is an indicator function that outputs 1, if the label pair $(y,\hat{y}) \in \mathcal{S}$, and else is 0. Note that $K_{\mathcal{S}}(y,\hat{y}) = K_{\mathcal{S}}(\hat{y},y)$. We obtain the set $\mathcal{S}$ from the confusion table of the validation sets.

We can now rewrite the discriminant function (1) by including these feature vectors:

$$f_\theta(x,y,h) = \theta^y_{gl} \cdot \phi(x) + \theta^y_{loc} \cdot \phi(x, h^y_{loc}) + \sum_{\hat{y} \in \mathcal{Y}} K_{\mathcal{S}}(y,\hat{y})\, \theta^y_{sh,\hat{y}} \cdot \phi\left(x, h^y_{sh,\hat{y}}\right) \quad (2)$$

where θ^y_{gl} ,θ^y_{loc}, $\theta^y_{sh,\hat{y}}$ denote the parts of θ^y that correspond to the global, local, and shared classifier parameters resp, *i.e.* we define $\theta^y = \left(\theta^y_{gl}, \theta^y_{loc}, \theta^y_{sh,c_1}, \cdots, \theta^y_{sh,c_k}\right)$ and $\theta = \left(\theta^{c_1}, \cdots, \theta^{c_k}\right)^{\mathrm{T}}$. The set of latent parameters can similarly be written as $h^y = \left(h^y_{loc}, h^y_{sh,c_1}, \cdots, h^y_{sh,c_k}\right)$ and $h = \left(h^{c_1}, \cdots, h^{c_k}\right)^{\mathrm{T}}$.

We use a common or *shared* parameter vector $\theta^y_{sh,\hat{y}}$ to encode the similarity between the labels y and $\hat{y}$. The equality $\theta^y_{sh,\hat{y}} = \theta^{\hat{y}}_{sh,y}$ means that the classes y and $\hat{y}$ share a common parameter vector. The latent parameters are used to learn instance specific shared, rectangular windows $h^{c_1}_{sh,c_2}$ and $h^{c_2}_{sh,c_1}$ as well as the target class-specific rectangular windows $h^{c_1}_{loc}$ and $h^{c_2}_{loc}$. Yet, as the window labels h are actually not available, we treat them as latent parameters and follow the LSSVM formulation of [8] to train the classifier parameters (θ).

		Flowers17 [4]	Interactions [5]
Baselines	gl[7]	65.6±4.3	34.4
	loc [1]	63.1±4.0	35.2
Ours	gl+loc	68.7±3.2	37.2
	loc+sh	65.2±5.1	37.6
	gl+sh	66.1±4.0	**40.0**
	gl+loc+sh	**71.1**±0.7	**40.0**

Table 1. Classification results are given as the classification accuracy averaged over the different target classes, in percentages.

3. Experiments

We evaluate our method on the Oxford Flowers17 [4] and Interactions [5]. We report the classification results for each of the feature types individually, and also their combinations, *i.e.* gl+loc, gl+sh, loc+sh and gl+loc+sh. We refer to gl and loc as the baselines, corresponding to the work by [7] and [1], resp. The results for the baselines and the proposed methods are depicted in Table 1.

Flowers17: The dataset contains 17 flower categories and 80 images from each flower species. We use densely sampled Lab color values and quantize them using an 800 words dictionary. We obtain an improvement of 5.5% using the combined configuration of the 'gl+loc+sh' model. Adding the shared part of the model always came out to be beneficial and enhanced the classification performance.

Interactions: This dataset contains video sequences containing four human interaction types: hand shakes, high fives, hugs, kisses and an additional background class. In this dataset we obtain an improvement of 4.8% over the baseline method. Again, the accuracy for classifying actions in these videos was improved by adding shared features. This is interesting as the nature of the dataset is quite different from the image classification datasets. The localization here is purely temporal.

Acknowledgments: This work was supported by the EU Project FP7 AXES ICT-269980.

References

[1] H. Bilen, V. P. Namboodiri, and L. Van Gool. Object and Action Classification with Latent Variables. In *BMVC*, 2011.

[2] H. Bilen, V. P. Namboodiri, and L. Van Gool. Classification with global, local and shared features. In *DAGM*, 2012.

[3] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, pages 3474–3481, 2012.

[4] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, pages 1447–1454, 2006.

[5] A. Patron, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, pages 1–11, 2010.

[6] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *CVPR*, pages 537–544, 2009.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[8] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, 2009.