Object and Action Classification with Latent Variables

Hakan Bilen[1]
hakan.bilen@esat.kuleuven.be
Vinay P. Namboodiri[1]
vinay.namboodiri@esat.kuleuven.be
Luc J. Van Gool[1,2]
luc.vangool@esat.kuleuven.be

[1] ESAT-PSI
VISICS/K.U. Leuven
Leuven, Belgium

[2] Computer Vision Laboratory
BIWI/ETH Zürich,
Zürich, Switzerland

In this paper we address the problem of classifying objects (e.g. person or car) and actions (e.g. hugging or eating) [2]. The more successful methods are based on a uniform pyramidal representation (SPM) built on a visual word vocabulary [1]. In this paper, we augment the classification by adding more flexible spatial information. This will be formulated more generally as inferring additional unobserved or 'latent' dependent parameters. In particular, we focus on two such types of parameters:

- The first type specifies a cropping operation. This determines a bounding box in the image. This box serves to eliminate non-representative object parts and background.
- The second type specifies a splitting operation. It corresponds to a *non-uniform* image decomposition into 4 quadrants or temporal decomposition of a spatio-temporal volume into 2 video sequences.

Apart from using these operations separately, we also study the effect of applying and jointly learning both these types of latent parameters, resulting in a bounding box which is also split. In any case, uniform grid subdivisions are replaced by more flexible operations.

Suppose we are given a training set $\mathscr{S} = \{(x_1, y_1), ..., (x_n, y_n)\}$ where $x_i \in \mathscr{X}$ are the input images/videos and $y_i \in \mathscr{Y}$ are their class labels. We want to learn a discriminant function $g : \mathscr{X} \rightarrow \mathscr{Y}$ which predicts the class label of unseen examples. In our applications input-output pairs also depend on unobserved latent variables $h \in \mathscr{H}$. Therefore we learn the mapping in the structured learning framework of [4],

$$g(x) = argmax_{(y,h)\in\mathscr{Y}\times\mathscr{H}} f(x,y,h). \quad (1)$$

where $f(x,y,h)$ is a discriminative function that measures the matching quality between input x and output y.

For training the discriminant function, we follow the generalized support vector machine in margin rescaling formulation [4],

$$\begin{aligned} &\min_{\omega,\xi_i\geq 0} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i, \\ &\text{subject to } \max_{h_i\in\mathscr{H}} \omega\cdot\left(\Phi(x_i,y_i,h_i) - \Phi\left(x_i,\hat{y}_i,\hat{h}_i\right)\right) \geq \Delta(y_i,\hat{y}_i) - \xi_i, \\ &\forall \hat{y}_i \in \mathscr{Y}, \forall \hat{h}_i \in \mathscr{H}, i = 1,\ldots,n \end{aligned} \quad (2)$$

where $f(x_i,\hat{y}_i,\hat{h}_i) = \omega\cdot\Phi\left(x_i,\hat{y}_i,\hat{h}_i\right)$, ω is a parameter vector and $\Phi\left(x_i,\hat{y}_i,\hat{h}_i\right)$ is a joint feature vector. $\Delta(y_i,\hat{y})$ is the loss function that penalizes misclassification. Since our applications require multiclass classification, we design our loss function as

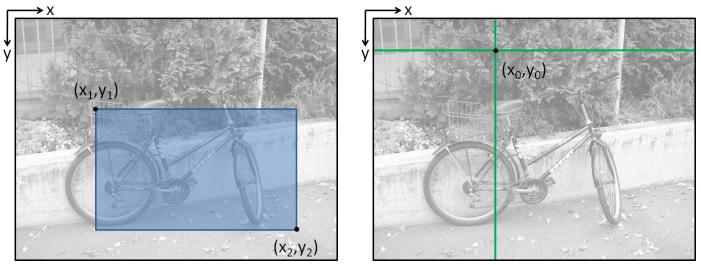
$\Delta(y_i,\hat{y}) = 100\,[y \neq \hat{y}]$, with $[\,]$ are the Iverson brackets and our feature vector as

$$\Phi_{multi}(x,y,h) = \begin{pmatrix} 0 & \ldots & 0 & \Phi(x,y,h) & 0 & \ldots & 0 \end{pmatrix}^T \quad (3)$$

where the feature vector $\Phi(x,y,h)$ is concatenated into position y. It should be noted that the problem reduces to the Standard Structural SVM formulation [3] in the absence of latent variables. It is used as the learning tool for the baseline approach.

The crop latent model is represented using a rectangular bounding box to separate the image into two parts, relevant to the class or not. The bounding box is represented by two points for both spatial and temporal cropping. We denote the latent parameter set with $h_{crop} = \{x_1, y_1, x_2, y_2\}$ and $h_{crop} = \{t_1, t_2\}$ for images and video sequences resp. An illustrative figure for each latent model is shown in Fig.1.(a).

The split operation results in a pyramidal representation that divides an image into unequal quadrants. In the same vein, we allow a video fragment to be temporally be split into two parts, which are not halves. Indeed, a naive split would probably not keep all object or action evidence within the same subdivision. Note that in this paper we only consider a single layer of subdivision of the pyramid, and the extension to full pyramids is not covered yet. Hence, such split is fully characterised by one point. We denote the latent variable set with $h_{split} = \{x_0, y_0\}$ (Fig.1.(b)) for images.



(a) crop (b) split

Figure 1: Illustrative Figure for Latent Models - Images

Dataset	BoF	SPM	Latent-model
Graz-02	80.40	81.47	**86.89** (crop-uni-split)
Caltech 101	80.77	85.42	**88.25** (crop-split)
VOC2006	55.07	57.89	**59.38** (crop-uni-split)
Everyday	82.67	84.67	**86.67** (crop-uni-split)

Table 1: Average Classification Accuracy for the Graz-02, VOC2006 and Everyday datasets

We evaluate our system on four publicly available datasets, the Graz-02, the modified PASCAL VOC 2006 and the reduced Caltech 101 datasets for object classification, and the activities of daily living life dataset for action classification. The features are obtained by the standard bag of visual words (*B*ag *o*f *F*eatures) approach for objects and actions. We compare the performance of the proposed latent models, 'crop', 'split', 'crop-uni-split', 'crop-split' to the standard BoF and one level SPM representations. We provide the results for the best performing latent model. Further results are provided in detail in the paper.

As can be seen from table 1, the use of latent models results in an improvement of performance. Additionally, we have also explored a method to improve the learning of the model with latent variables by iteratively growing the latent parameter space to avoid local optima. To conclude, we have proposed a method for classifying objects and actions with latent variables. We have specifically shown that learning latent variables for flexible spatial operations like 'crop' and 'split' are useful for inferring the class label. We have adopted the latent SVM method to jointly learn the latent variables and the class label. The evaluation of our principled approach yielded consistently good results on several standard object and action classification datasets. In future, we are interested in extending the set of operations that aid classification and improving the learning of multiple parameters.

[1] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[2] Axel Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4), 2005.

[3] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proc. Int. Conf. on Machine Learning (ICML)*, page 104, 2004.

[4] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 1169–1176, 2009.