

Weakly Supervised Object Recognition and Localization with Invariant High Order Features

Yimeng Zhang
yz457@cornell.edu
Tsuhan Chen
tsuhan@ece.cornell.edu

School of Electronic and Computer Engineering
Cornell University
Ithaca, NY, USA

The "bag of features" (BoF) representation of images [5] has been extremely popular for object categorization task due to its advantages of great invariance and computational simplicity. High order features have been proposed to model the mutual geometrical relationship between the local features [2, 3, 4, 6]. High order features (HOFs) are created to represent a specific number of local features in a particular spatial relationship. According to the number of local features used, the HOFs are called doublet, triplet, or n^{th} order features. This type of method can be learned invariantly with weakly labeled data. However, as the order increases, the number of features increases exponentially to the order.

In this paper, we propose algorithms to perform fast weakly supervised object categorization and localization with high order features. To this end, we first identify translation and scale invariant high order features co-occurring in two images. The co-occurrence is used to calculate a kernel for a SVM. In [6], we proposed an efficient algorithm that identifies translation invariant co-occurring high order features in time linear to the number of local features. In this paper, we extend the work with more invariance and efficiency.

There are two main contributions of this paper. Firstly, we propose an algorithm which can easily add more invariance to the HOF kernel [6]. Our extension is based on the observation that the process for identifying the co-occurring translation invariant high order features of two images in [6] is analogous to the process of the generalized Hough Transform [1]. The Hough transform allows a pair of features from two images to vote for their transformation parameters. Thus we can easily define any transformation invariance we want to introduce, such as translation, scale, or rotation. Secondly, we extend the HOF idea for object localization. A naive way would be to apply the SVM for all possible subwindows. For most detection algorithms that include a SVM with non-linear kernels, the computation can be expensive, since kernel calculations must be processed for all possible subwindows. This requires $O(SM)$ kernel computations per image, where S is the number of support vectors, and M is the number of possible subwindows in an image. The proposed algorithm collects the weights of high order features for the subwindows while calculating kernel value for the entire image, and thus reduces the kernel computations to $O(S)$.

Fig. 1 illustrates the idea of weakly supervised localization with HOFs. A test image is evaluated by calculating the kernel values with each support vector of the SVM learned during training. Because of the invariance of the HOFs, objects that occur in different positions and at different scales in the training images can contribute properly to the scores for localization, although the training images are not aligned, and the scale is not normalized.

We briefly describe the algorithm for calculating the invariant high order feature kernel here. An n^{th} order feature is composed of n visual words in a particular mutual scale and spatial relationship. If an n^{th} order feature can be created as a translation or/and scale transformation of the n words of the other n^{th} order feature, the two high order features are defined as the same feature. Fig. 2 gives an example of two occurrences of a 3rd order feature on two images with position and scale changes. To identify the co-occurring high order features of two images, we use the idea of the Generalized Hough Transform [1]. As illustrated in Fig. 2, For each pair of patches with the same visual word assignments from the two images, we make a vote on the transformation parameter space $(\hat{x}, \hat{y}, \hat{s})$ with the following values: $(x_i - \frac{s_j}{s_i}x'_i, y_i - \frac{s_j}{s_i}y'_i, \log(\frac{s_j}{s_i}))$, where x_i, y_i denote the location, and s_i denotes the region size of a local feature. If we have n votes at a particular point on the parameter space, we have n pair of patches with the same transformation $(\hat{x}, \hat{y}, \hat{s})$. From the definition for high order features, we have a co-occurrence of a certain n^{th} order feature. In the paper, we derived that the n^{th} order kernel value of two images equals to the number of co-occurrences of all n^{th} order features. Therefore, we can calculate the kernel values efficiently using the parameter space. Moreover, we can also obtain the kernel values of the support

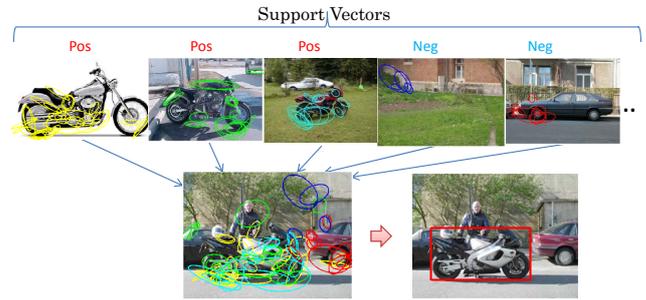


Figure 1: Illustration of the weakly supervised algorithm with invariant high order features. The input image is evaluated on the motorcycle classifier (SVM) learned with weakly labeled data. The images in the top line represent the support vectors with positive or negative weights. Each ellipse corresponds to a local feature. We show the largest order feature co-occurring in the input image and each support vector in different colors for different support vectors.

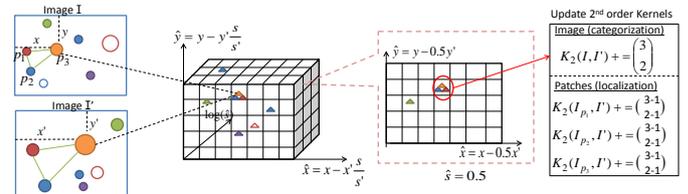


Figure 2: Illustration of the algorithm for finding co-occurrences of invariant high order features. Each circle in the images is a local patch. The different colors represents different visual word assignments. The size of the circle represents the scale of the patch. A triangle corresponds to a vote generated by a pair of patches.

vector and each local patches at the same time. The SVM decision score for each subwindow can be calculated as the summation of the scores of the local patches inside the window. Thus, we can obtain the localization results without performing kernel calculations for all subwindows. Detailed implementation of the algorithm is described in the paper.

We evaluated the proposed approach on several public datasets, including the Pascal VOC 2005, Graz-01, Graz-02, and Caltech-4. The experiment results showed that: 1) Adding more invariance to the high order features significantly improved the recognition performance. 2) The classifier with high order kernels which is learned with only weakly labeled data can be used to localize the objects on new test images. On the Pascal dataset, the detection performance is close to the winner of the challenge, whose model is trained with fully labeled data.

- [1] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Readings in computer vision: issues, problems, principles, and paradigms*, pages 714–725, 1987.
- [2] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *CVPR*, June 2007.
- [3] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.
- [4] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *CVPR*, 2006.
- [5] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73, 2007.
- [6] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009.