Efficient Kernels Couple Visual Words Through Categorical Opponency

Ioannis Alexiou

Anil Anthony Bharath
http://www.bg.ic.ac.uk/research/a.bharath/

Biologically Inspired Computer Vision
Department of Bioengineering
Imperial College London, UK

Vision systems, designed for tasks such as object recognition and categorization, are based on a series of well-defined processing stages. Typically, these stages will include keypoint detection schemes, which sample a dense, transform-domain representation of an image. Commonly used keypoints detectors are Harris, Hessian, MSER and DoG, which are able to produce a sparsified representation of an image, although recent work has shown that dense sampling can drastically increase the recognition performance. A "selection" mechanism may also be employed to sparsify the dense descriptor representation where combined sparse coding and max-pooling 'distil' the features over defined image regions.

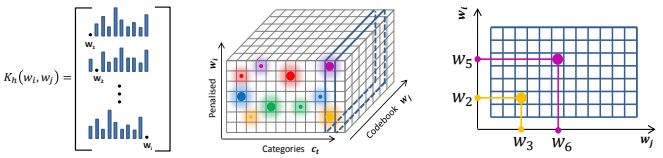
There are numerous methods that can bind those mid-level features into a compact representation, whilst ensuring that spatial information is ecoded: spatial pyramids, neural networks and other, high-order of spatial features. Spatial pyramids incorporate coarse spatial relationships producing good recognition rates. Yet another efficient approach binds descriptors into multiplets of features yielding much improved recognition performance compared to a bag of features approach. The latter approach attempts to encode spatial relationships using relative distances (*Correspondence transform*). Co-occurences of these high-order features are mapped onto an offset space where occurence counts are assigned to the features which satisfy predefined distance criteria.

Inspired by the observation that binding descriptors may lead to quasi-contour construction, we may assume that parts of contours can be a flexible tool to improve recognition rates and pose invariance. Specifically, constellation methods have shown that fixed numbers of parts, which come from standard descriptors, is a powerful representation tool. Typical disadvantages of such approaches involve a predetermined, fixed number of parts where computationally expensive search assigns parts to locations and set a hypothesised center for the model. In this research we focus on whether a selection mechanism can reduce the dictionary size of high-order features. Specifically, we focus on the transition from visual words to 2_{nd}, 3_{rd}, $\ldots$, k_{th} visual word order, which increases the computational complexity by a factor of N^k where N is the size of the dictionary. This work proposes efficient kernels to create small dictionaries of paired words utilizing *categorical opponency* to unveil such pairs. Secondly, we examine how proximity of such pairs can be scalable, and how the proximity of pairing affects performance. Thirdly, coupling kernels are applied per image using decision functions to detect co-occurrences using an approach that is compatible with fast indexing.

Dictionaries of higher-order (than simple visual words) can raise the size of the dictionary exponentially. Even paired words can yield numerous combinations depending on the size of the initial codebook. Among these numerous pairs a subset can exist that makes the problem tractable. In addition it is assumed that specific pairs are very unique to each class and others non-informative. Considering the aforementioned, a data-driven method is described to mine those pairs from the train set given the ground truth of the object classes.

$$K_h(w_i, w_j) = P(w_j|w_i) \cdot G_i \tag{1}$$

The kernelized expression (1) provides a statistical medium to monitor specific visual words down-weighted by the "G_i" term. We look at the probability that a keyword occurs conditional on another specified word, and estimate a function similar to a joint probability of occurrence. Specifically, a histogram of visual words is constructed by removing the candidate w_t. By removing the w_t we monitor the co-occurrence of other(non-identical) words in the image. Often the word w_t can be assigned to several keypoints then the votes are accumulated to the correspondent row $P(w_j|w_t)$ in the kernel $K_h(w_t, w_j)$. This procedure is done per object then the kernels are L1 normalized and arranged according to object classes. The creation of these kernels is not much slower than computing histograms of visual words. The words are assigned once for every image, then the look up is achieved by indexed retrieval

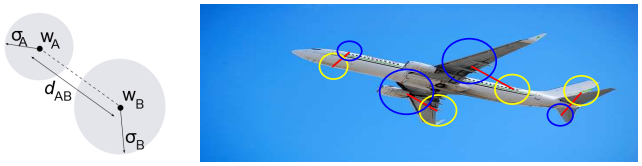


(a) (b) (c)

Figure 1: Starting from left to right: (a) illustrates how a kernel may be constructed using conditional histograms. (b) This figure shows the concept of first and second maxima. Each colour represents a different pair, where the size of the coloured/shaded blobs illustrates the 1^{st} and 2^{nd} maxima. (c) Once a candidate category is selected, we focus on the maximum votes that belong to this category. Finally, these pairs are selected in order from highest to lowest ratios.

$$K_{Op}(w_i, w_j) = \frac{\max\limits_{c_t} K_h(w_i, w_j, c_t)}{\max\limits_{c_t' \neq c_t} \left[K_h(w_i, w_j, c_t') \leqslant \max\limits_{c_t} K_h(w_i, w_j, c_t) \right]} \tag{2}$$

The parameter c_t characterises an object label assigned to a normalized kernel. Presumably if a histogram of words can provide a rough discrimination among categories then pair might must exist enhancing this behaviour. Expression (2) is a analytic approach to detect which pairs have high dominance over a specific class. Searching along the categories $c_t \in \{1, \ldots, Total\,Categories\}$ the maximum of kernel entry is divided by the second maximum found in another category. This forms a ratio where the higher the ratio the higher the opponency of this class against the others. This means that for a given combination of visual words this tends to be unique to the examined class.



(a) (b)

Figure 2: (a) illustrates a word pair comprised of two words w_A and w_B. The associated keypoint information captures the effective descriptor (spatial) radius (σ_A and σ_B), and relative Euclidean distance (d_{AB}), which is used to derive the paired words. (b) illustrates the single word pair occurring 4 times; multiple occurrences such as this are used in Equation (3).

$$\mathcal{B}^{(p)} = \sum_{m|n=1}^{Q} \max_{n|m=1}^{Q} K^{(p)}_{N\times M}(\phi^{(n)}_{w_j} = w_j, \phi^{(m)}_{w_i} = w_i) \tag{3}$$

where Q is taken to be the smaller dimension of the $K_{N\times M}$ matrix, i.e. either M or N.

Experiments in Caltech 101 database using 30 training examples per class are summarised in the next table to show the amount of improvement using pairs of visual words.

	Pyramids (10500)	Pairs (5000)
Keypoint-Based	$52.30_{\pm 1.28}$	$71.00_{\pm 2.00}$
Grid-Based	$60.16_{\pm 1.56}$	$73.88_{\pm 2.10}$

Table 1: Average accuracies per method for Caltech 101.

In conclusion, pairing up visual words can increase the classification rates comparing to spatial pyramids with hard word assignments. Future aims are to explore alternative pairing kernels, merge spatial pyramids and sparse coding approaches with pairs of visual words.