This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

Early burst detection for memory-efficient image retrieval

Miaojing Shi[1], Yannis Avrithis[2] Hervé Jégou[3]
[1]Peking University. [2]University of Athens, NTUA. [3]INRIA, Rennes.

Visual burstiness [3] tends to dominate the similarity measure in image retrieval and classification, which degrades the quality of the comparison, as other non-bursty yet possibly distinctive features have a comparatively lower contribution. Various strategies have been proposed to discount the contribution of bursts on the similarity measure. Some are inspired by text like the power-law normalization [5] for bag-of-words or the Polya or Dirichlet models [1]. These strategies have been pragmatically extended to and improved for more complex image vector representations such as VLAD [2]. They are also standardly used in matching approaches like Hamming Embedding [4] or selective match kernels (SMK/ASMK) [6].

Fig. 1 illustrates a representative example of bursts in an image. Despite the similarity in appearance, one might expect a high density of points around each burst in the descriptor space, which, however, is far from being true. *Isolated* descriptors (not belonging to any group) appear to have the same density as *bursty* ones (belonging to some group), while bursts have arbitrary shape and large extent. One cannot hope that bursts will fit within the cells of a codebook. Therefore, we propose an *early burst detection*, before quantizing descriptors. We compare pairwise distances of all patches and join pairs whose distance is below a certain threshold. We then find the connected components and color them as bursty groups in Fig. 1.

Given two local image features f, g, we define *feature kernel* function

$$k(f,g) = k_u(u_f,u_g)k_s(s_f,s_g)k_\theta(\theta_f,\theta_g), \tag{1}$$

consisting of three factors, namely the *descriptor kernel* k_u, the *scale kernel* k_s and the *orientation kernel* k_θ. Here, u_f, s_f, θ_f are the descriptor, scale and orientation of feature f. Intuitively, two patches belong in the same burst if they are similar in appearance and have similar scales and orientations. *Descriptor kernel* k_u measures the similarity of a pair of descriptors $x, y \in \mathbb{R}^d$ and is a function of the inner product $z = \langle x, y \rangle$. In particular, we adopt a generative model for a binary classifier: if $\mathcal{B}$ is the class of descriptor pairs that belong to the same burst and $\overline{\mathcal{B}}$ is its complement, we define

$$k_u(x,y) = p(\mathcal{B}|\langle x,y\rangle) = p(\mathcal{B}|z). \tag{2}$$

Here, $p(\mathcal{B}|z)$ is the posterior probability of $\mathcal{B}$ given z and can be formulated via class-conditional densities, $p(z|\mathcal{B}), p(z|\overline{\mathcal{B}})$. We train a classifier from a dataset of matching/non-matching patch pairs [7], where these densities are modeled as normal densities, fitted to data samples according to maximum likelihood. *Scale and orientation kernels* employ a Gaussian and a von Mises kernel respectively.

Given an image, we construct its affinity matrix K including all pairwise feature similarities, $K_{ij} = k(f_i, f_j)$ where kernel k is given by (1). The affinity matrix is the only input for burst detection. We examine a number of kernel methods to detect the bursts from K. Preliminary qualitative evaluation shows that *connected components* is the fastest and most effective one, so this is adopted in most quantitative experiments.

The result of burst detection is a partition of its features into groups. To *aggregate*, we simply take the average of the descriptors in each group and ℓ_2-normalize. Discarding geometry, this yields a set of descriptors to represent the image, so any encoding/retrieval model applies. We follow VLAD [2] and SMK/ASMK [6] in particular. We also follow two aggregation strategies: *symmetric* and *asymmetric*, depending on whether query descriptors are aggregated or not. We apply different thresholds to the affinity matrix to vary the number of bursts such that *aggregation%* — the proportion of aggregated to original descriptors — varies in the range of 10-100%. It turns out that asymmetric is superior for low aggregation%.

Fig. 2 compares three different initial feature sets (*-S, -M, -L*) on *Holidays* and measures mAP *vs*. absolute number of descriptors per image,

Figure 1: An image along with the features of the six most populated bursts detected. A dot is shown at the position of each feature, colored according to the burst it belongs to.

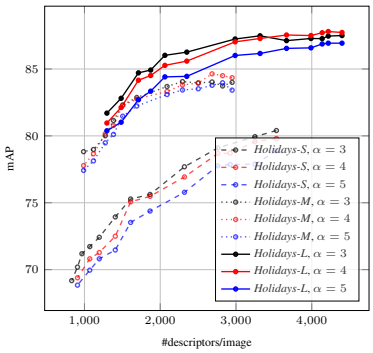

Figure 2: ASMK* mAP *vs*. average number of aggregated descriptors/image on *Holidays* for three different initial feature sets and values of selectivity exponent α. Vocabulary size $k = 65$k; asymmetric aggregation.

which directly reflects memory. The largest set maintains a gain of over 10% over the smallest one. This is a key aspect of the trade-off and suggests a way to improve performance: *augment the initial features, aggregate, and gain in mAP at the same memory*. Our conclusion is that, by fusing the descriptors before feeding them to the indexing or search system, we reduce the computational cost in both quantization and retrieval, typically by a factor of two. We also reduce the memory footprint in the same proportion for search engines employing an inverted file; or, performance may be increased at the same memory.

[1] RG Cinbis, J. Verbeek, and C. Schmid. Image categorization using Fisher kernels of non-iid image models. In *CVPR*, Jun. 2012.

[2] J. Delhumeau, PH. Gosselin, H. Jegou, and P. Perez. Revisiting the VLAD image representation. In *ACM Multimedia*, Oct 2013.

[3] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

[4] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3), 2010.

[5] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, (99): 1–1, 2011.

[6] G. Tolias, Y. Avrithis, and H. Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, 2013.

[7] S. Winder and G. Hua. Picking the best daisy. In *CVPR*, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.