Visual object retrieval by graph features

Yi Ren, Aurélie Bugeau, Jenny Benois-Pineau

Visual object retrieval by graph features

Yi REN, Aurélie BUGEAU, Jenny BENOIS-PINEAU

Univ. Bordeaux, LaBRI, UMR 5800 - F-33400 Talence, France
{yi.ren, aurelie.bugeau, jenny.benois-pineau}@labri.fr

Thème – Thème principal: T1 Représentations et modèles - T1.4 Graphes en signal et image, modèles graphiques, réseaux de capteurs, graphes de terrains - T4.6 Indexation

Problème traité – Object and Image retrieval

Originalité – Visual object retrieval using graph features

Résultats – Our approach addresses the well-known problem of adding spatial information to the standard bag-of-visual-words (BoVW) approach. To that end, a set of graphs descriptors are constructed for the entire image or for local image regions. These descriptors are then incorporated into a bag-of-visual-words approach, leading to a method called bag-of-bag-of-visual-words (BBoVW).

1 Introduction

This paper presents an object retrieval system based on the bag-of-visual-words (BoVW) approach [8]. It is well known that this method does not embed any spatial information in the image representation. In this paper, we propose to incorporate such information through the representation of an image by a set of graphs. The proposed approach works as follows : first, the feature points of an image are partitioned into several graphs by minimizing an energy via graph-cuts. Using a codebook computed from all the feature points in a database, a histogram is built for each graph by assigning each of its features to the closest word in the codebook. Hence, an image being composed of k graphs is characterized by a set of k histograms. We refer to this technique as bag-of-bag-of-visual-words. In order to compare two images, a similarity measure that relies on the computation of distances between graphs is finally designed. Our initial experiments show improvements over the original BoVW for certain object categories in which feature points are more stable.

2 Related works

Indexing methods Recent methods in the domain of image indexing all rely on the BoVW model [8]. The idea, borrowed from document processing, is to compute a visual dictionary or codebook from all the feature points in a database, and then represent each descriptor by an index referring the most similar visual word in the vocabulary. Since the final image representation is made by computing a histogram of visual words occurrences, image features are therefore considered as *independent* and *orderless*, thus ignoring any spatial relationship between them. However, spatial information has shown to be very useful in tasks like image retrieval, image classification, and video indexing. For that reason, several works can be found in the literature that have tackled this issue [6, 7].

Graph-based image representation Graphs are powerful and versatile tools that are useful for representing patterns in computer vision and pattern recognition applications. When graphs are used to represent an image, measuring the similarities between images becomes equivalent to finding similar patterns inside a series of graphs that represent those images. In this case, representing kinds of patterns by attributed graphs is extremely convenient and have been vastly investigated [4, 5].

3 Representing an image by a set of graphs

Feature points selection First, feature points are extracted on the whole image or on the mask of the object being retrieved. We use SURF descriptors [2] in our experiments. In some cases, the extracted points may be very close to each other, what will lead to too small triangles preventing from suitable triangulations. To avoid that, we filter the feature points by preserving only the most pertinent keypoint over any 10-pixels diametral region. After filtration, we obtain a set of filtered keypoints $\mathcal{P} = \{p_1, ..., p_n\}$ (see figure 1(b)).

(a) (b) (c) (d) (e)

FIGURE 1 – Graph construction. (a) Initial SURF keypoints. (b) Filtered keypoints. (c) Seeds selection. (d) Delaunay triangulation over filtered feature points. (e) Graph-cuts result.

Keypoints Triangulation Next, a Delaunay triangulation is performed over the set $\mathcal{P}$. A connected graph is then generated for each image (figure 1(d)). It is composed of triangles built according the Delaunay constraint, *i.e.* by maximizing the minimal angle of the triangulation. Using this process, the spatial relationship between local features is invariant to image translation, scaling, and rotation.

Constructing the set of graphs Let us denote $G = (\mathcal{V}, \mathcal{E})$ as a graph generated by the Delaunay triangulation. The vertices set $\mathcal{V}$ contains all the feature points : $\mathcal{V} = \mathcal{P}$. The edges set $\mathcal{E}$ contains all unordered pairs of points $\{p_i, p_j\}$ that are neighbours in the Delaunay graph. We want to separate graph G into k smaller graphs using graph-cuts. This can be formulated as a labeling problem : given a points set $\mathcal{P}$ in image I, and a label set $\mathcal{L} = \{l_1, l_2, ..., l_k\}$, for each $p \in \mathcal{P}$, we are looking for its label $l(p) = l_p \in \mathcal{L}$. In order to be consistent with the image content, we construct a set of sub-graphs $\{g_I^1, ..., g_I^k\}$, according to the following requirements : each sub-graph, *i*) should be as compact as possible ; *ii*) should have, as much as possible, a uniform color. To solve this labeling problem, we minimize, via graph-cuts, an energy function containing a data term and a smoothness term :

$$E(L) = E_{data}(L) + E_{smoothness}(L) = \sum_{p \in \mathcal{S}} D(p, \ell_p) + \lambda \cdot \sum_{(p,q) \in \mathcal{E}} V_{p,q}(\ell_p, \ell_q) \tag{1}$$

The first term is called the *data term*. It is only applied to seed points $p \in \mathcal{S}$. Seed points $\mathcal{S} \subset \mathcal{P}$, are user-defined hard constraints imposing on graph cuts initialization. The keypoints that are more prominent in $\mathcal{P}$ will be prior to be selected. Moreover, to ensure that seeds with different labels will fall into heterogeneous regions in the image, we set both color threshold and distance threshold empirically to differentiate between chosen seeds (see figure 1(c)). For the i^{th} seed point $p_i \in \mathcal{S}$, the data term is defined as : $D(p_i, \ell) = 0$ if $\ell = \ell_i$ and ∞ otherwise.

The second term $V_{p,q}$ is the *smoothness term*. To follow the pre-stated requirements, we will encourage that any two neighbouring nodes $(p, q) \in \mathcal{E}$ are *i*) spatially close to each other ; *ii*) if node p and q have similar colors, they tend to have the same label. Hence, $V_{p,q}$ is composed of a color term $f_c(p, q)$ and a distance term $f_d(p, q)$:

$$V_{p,q}(\ell_p, \ell_q) = f_c(p, q) f_d(p, q)(1 - \delta(\ell_p, \ell_q)) \tag{2}$$

where $\delta(\ell_p, \ell_q)$ is the Kronecker's delta : $\delta(\ell_p, \ell_q) = 0$ if $\ell_p \neq \ell_q$ and 1 if $\ell_p = \ell_q$. The color term is given by

$$f_c(p, q) = \exp\left(-\lambda_1 (\bar{I}_p - \bar{I}_q) \Sigma_c^{-1} (\bar{I}_p - \bar{I}_q)^T\right) \tag{3}$$

where $\bar{I}_p = (\bar{Y}_p, \bar{U}_p, \bar{V}_p)$ is a mean color vector in YUV color space, computed over a 4-connected $5\times$ 5 region centred on p. We use the YUV color space in order to have independent color channels. The covariance matrix Σ_c is therefore considered diagonal. It is computed as, for all channels, the mean of the mean color vector of all nodes.

Denoting $p = (x_p, y_p)$ as the coordinates of node p, the distance term is defined as

$$f_d(p, q) = \exp\left(-\lambda_2 (p - q) \Sigma_d^{-1} (p - q)^T\right). \tag{4}$$

The covariance matrix Σ_d is also diagonal. It is computed as the mean distance of all neighbouring points in the graph G.

The labeling of each node is finally obtained by minimizing (1) with α-expansion [3]. A result is shown on (figure 1(e)).

4 The "Bag-of-bag-of-visual-words" Strategy

This section overviews our bag-of-bag-of-words object retrieval engine. The codebook is first computed using k-means clustering method over all descriptors of the filtered keypoints from the whole images in a learning database. The codewords C then correspond to the centres of the learned clusters. A bag of visual word representation, *i.e.* a histogram of word frequencies, is then assigned to each subgraph.

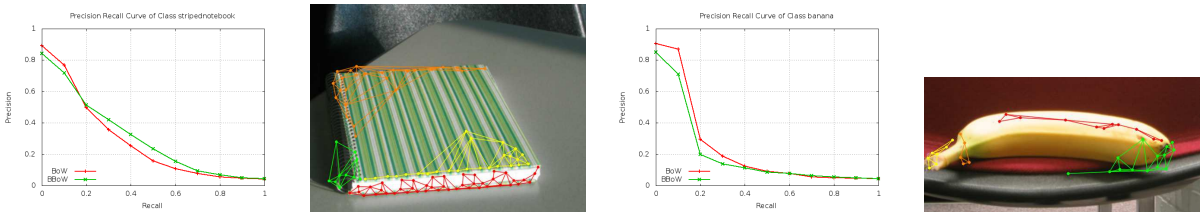
Therefore, given an image I, its sub-graph g_I^i has its own signature H_I^i computed using only the SURF descriptors of its nodes. The histograms set of I is $H = \{H_I^1, \cdots, H_I^k\}$. A (dis)similarity measure between two graphs can be easily computed as we just need to compare two histograms. We here adopt the idea of RootSIFT [1] by first L1 normalizing the histograms and then compute the histogram intersection of them. Nevertheless, we still need to compute the similarity between two images, knowing that they both contain k graphs. We have experimented different strategies to combine the $k \times k$ distances between a pair of query image I_1, and database image I_2 from their corresponding histogram sets H_{I_1} and H_{I_2} : maximum, resp. minimum, resp. sum, etc. of all distances. Our experiments have shown a better behavior with : $d(I_1, I_2) = \sum_{i=1}^{k} \min_{j=\{1,\ldots,k\}}\{1 - \sum_{b=1}^{C} min(H_1^i(b), H_2^j(b))\}$.

5 Experimental Results

Database The SIVAL benchmark is used for the experiments. It includes 25 different image categories with 60 images per category. This benchmark emphasizes the task of localized Content-Based Image Retrieval. It allows to clearly identify the content of images for which the method has good performance, and for which it does not.

Performance is measured by calculating precision, recall, and the Mean Average Precision (MAP) over each category, since that mixing of all categories has no sense for disparity of the object appearance across categories. Our results are compared with the classical BoVW representation built on the same set of feature points that the initial graphs were built on.

For three categories : 'stripednotebook', 'goldmedal', 'woodrollingpin' among 25 categories, the proposed BBoVW method performs better than BoVW. The example of results for a "good" category and a "bad" category are presented in figure 2. The mean improvement of MAP on the three relevant categories is of 8%. As for "bad " categories, the discrepancy of results is too high to make a meaningful statistics. The method is sensitive to the stability of feature points and also needs further development for robust energy terms.



(a) (b) (c) (d)

FIGURE 2 – Comparision between BoW and BBoW on SIVAL benchmark. (a) "good" category : 'stripednotebook'. (b) graph descriptors from 'stripednotebook'. (c) "bad" category : 'checkeredscarf'. (d) graph descriptors from 'checkeredscarf'.

6 Conclusion

In this paper, we have presented a novel method to improve the orderless bag-of-words model. We combine the bag-of-words histograms with graph features at decision level, thus spatial information is injected by graph descriptors. The experimental results on SIVAL dataset showed that the method is applicable for the classes of images with stable feature points and that further research is needed for identification of more stable energy potentials. This is the focus of our actual work.

Références

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[2] H. Bay, T. Tuytelaars, and L. J. Van Gool. Surf : Speeded up robust features. In *ECCV*, 2006.

[3] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9) :1124–1137, 2004.

[4] H. Bunke and K.Riesen. Towards the unification of structural and statistical pattern recognition. *Patt. Rec. Letters*, 33(7) :811 – 825, 2012.

[5] J. Gibert, E. Valveny, and H. Bunke. Graph embedding in vector spaces by node attribute statistics. *Patt. Rec.*, 45(9) :3072 – 3083, 2012.

[6] S. Karaman. *Indexation de la vidéo portée : application à l`étude épidémiologique des maladies liées à l`âge*. PhD thesis, Université Bordeaux 1, 2011.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[8] J. Sivic and A. Zisserman. Video google : a text retrieval approach to object matching in videos. In *ICCV*, 2003.