Content Based Image Retrieval using Bag-Of-Regions: an Efficient Approach

Remi Vieux, Jenny Benois-Pineau, Jean-Philippe Domenger

Content Based Image Retrieval Using Bag-Of-Regions

Rémi Vieux, Jenny Benois-Pineau, and Jean-Philippe Domenger

LaBRI - CNRS UMR 5800 - Université de Bordeaux firstname.name@labri.fr

Abstract. In this work we introduce the Bag-Of-Regions model, inspired from the Bag-Of-Visual-Words. Instead of clustering local image patches represented by SIFT or related descriptors, low level descriptors are extracted and clustered from image regions, as given by a segmentation algorithm. The Bag-Of-Region model allows to define visual dictionaries that capture extra information with respect to Bag-Of-Visual-Words. Combined description schemes and ad-hoc incremental clustering for visual dictionnaries are proposed. The results on public datasets are promising.

Keywords: Content Based Image Retrieval, Bag-Of-Regions, Incremental Clustering, Meta-Search

1 Introduction

Image retrieval is a challenging topic that has been a research challenge for several decades. The task is difficult for several reasons. Smeulders *et. al.* [24] introduced the concept of *semantic gap*, that is the discrepancy between the low level descriptors that can be computed from the images and the interpretation of the image done by humans. A query to an image retrieval system is ill-defined by nature. Such a query could take several forms. One of the earliest successful system, QBIC[8], accepted queries as a user defined color palette, that images should matches. A query can be formulated using an example image (Query-By-Example – QBE– paradigm). The system must retrieve the most similar images to the query. In this case, the notion of similarity is implicit for the user, and the system must approximate this notion into a computable quantity. In the best case, it can be related to several measurements in terms of low level descriptors.

In the last decade, a breakthrough in image retrieval and object recognition have been achieved using the Bag-Of-Visual-Words (BOVW) model based on interest-point descriptors such as SIFT[16]. In the mean time, methods based on region-based properties of the image have known a decrease of popularity for CBIR and classification tasks, since the fundamental work of Duygulu *et. al.* [5]. Few examples include Souvannavong *et. al.* [25] for video content indexing and retrieval and Gokalp and Askoy [10] for scene classification. However, current state-of-the-art for accurate object class image segmentation rehabilitates image segmentation and region-based visual description of the image content [12,26,27].

In this paper, we study an image retrieval system that extends the traditional notion of BOVW vocabulary not only to keypoint-based descriptors, but to region based descriptors, as obtained by a segmentation algorithm. Region-based descriptors open the way to exploit a vast amount of different visual cues, such as colour, texture and shape, that are not captured by keypoint descriptors. With this extension arise several challenges: first, in section 2, we clarify the definition of a region-based visual dictionary. The computation of visual dictionary has been considered traditionally as an offline, time independent process. This point becomes more problematic when several such dictionaries must be built. In section 3, we propose to rely on an incremental clustering method that has lower memory and computational complexity than k-means, the reference algorithm. Able to express the image content through several visual dictionaries, we combine them to improve the single-modality retrieval results. We introduce in section 4 the topic of *meta-search* and its most famous strategies. In section 5, we perform deep experiments of the method on three public datasets. We conclude the paper in section 6 and give research perspectives for the future of this work.

2 Bag-Of-Regions Model

The BOVW model has been inspired by the Bag-Of-Words (BOW) model for text document representation. In the BOW model, a text document is represented by the number of occurrences of the words in the document. Despite the simplicity of the model, which neither takes into account the order of the words, nor the relationships between them, this model is very efficient for document classification tasks [11]. Sivic and Zisserman proposed to compute a visual dictionary by clustering similar visual entities inspired by BOW model. Hence, to build a visual dictionary, we must define two key concepts: what entity defines the *spatial support* for a visual word and which *descriptor* underpins the notion of similarity in the clustering process. In the BOVW model, local interest points are used as salient image patches and SIFT or related [18] descriptors used to describe the patches. We propose to define as the basic entities supporting the visual words the *image regions* obtained by a segmentation algorithm. An extremely rich collection of descriptors can be extracted from image regions providing new kind of visual dictionaries. We call the Bag-Of-Regions dictionary BOR. Figure 1 shows the different steps for BOR extraction. If these are similar to BOVW, BOR is trickier to compute due to the number of parameters in the process. The parameters are represented by the colored box in figure 1: there can be several segmentations, many visual features and different quantization of the visual space to compute a single BOR dictionary. Hence, the number of single BOR models that can be computed explode with the number of parameters. The time spent in the vocabulary computation is usually considered as irrelevant, since clustering is performed offline. Time does matter with such an amount of dictionaries to compute. There is a need for an efficient clustering algorithm able to produce these results in a reasonable time.

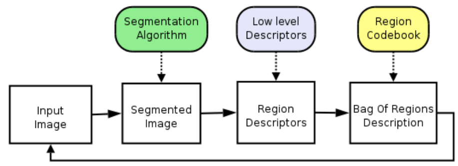

Fig. 1. Bag-Of-Regions extraction pipeline

3 Incremental Clustering for Visual Dictionnaries

Clustering for codebook construction is a difficult problem because of the large number of data samples as well as clusters. K-means clustering has been the reference and most popular method so far [23]. Alternative for to k-means have been proposed by Nister and Stewenius, for handling large amount of data [19]. Yeh *et. al.* proposed the dynamic computation of visual vocabulary using adaptive vocabulary forests [28]. We propose to replace the traditional k-means algorithm with the incremental vector quantization of Lughofer [17]. The principles of vector quantization are the following:

1. Choose initial values for the k cluster centers, $\mathbf{c}_k, k = 1, \ldots, K$.
2. Fetch out the next data sample $\mathbf{x}$ of the data set $\mathcal{D}$
3. Calculate the distance of the selected data point to all cluster centers.
4. Elicit the cluster center which is closest to the data point as

$$\mathbf{c}_{win} = \arg\min_k d(\mathbf{x}, \mathbf{c}_k) \tag{1}$$

5. If $d(\mathbf{x}, \mathbf{c}_{win}) \geq \rho$ then the current sample $\mathbf{x}$ becomes the center of a new cluster. Otherwise move the cluster center towards the new point:

$$\mathbf{c}_{win}^{new} = \mathbf{c}_{win}^{old} + \eta(\mathbf{x} - \mathbf{c}_{win}^{old}), \eta \in [0, 1] \tag{2}$$

The main advantage of incremental clustering with respect to k-means is the lower computational complexity. All the data are processed in a single pass. The computational complexity of the incremental clustering is $O(KNd)$ with K the number of clusters, N the number of vectors and d their dimension. K-means is $O(IKNd)$ with I the number of iterations. Extensions to vector quantization to fit a real incremental clustering task with unknown number of clusters are given in [17] that we do not detail here.

However, we have seen that those extensions are not directly suitable for our task. Indeed, clustering model can differ significantly while processing the same data in different order. We propose to choose the initial cluster centers according to the k-means++ initialisation [1]. In this way, we ensure that a minimal number of clusters (visual words) is reached. Moreover, as the position of the centroid is crucial during the incremental clustering process, it is natural to chose centers that reflect well the organisation of the data as k-means++

does, leading to a more robust model. In section 5, in order to compare the incremental clustering with k-means, we ensured that no clusters were created incrementally (*i.e.* setting up a high threshold ρ).

4 Fusion of Multiple Retrieval Systems

The BOR model allows a profusion of different representations of image content based on the nature of the low level descriptor, the granularity of the segmentation or the quantization method used for vocabulary construction. Retrieval systems based on these vocabularies are likely to return different sets of images. The optimal combination of this set is the problem of *meta-search* [2]. We assume that the retrieval systems return results in a decreasing order of similarity to the query. Two major types of error can occur for any such system[9]: 1) giving a high rank to non relevant documents and 2) giving a low rank to relevant ones. In table 1, we present the most widely known strategies for combination [9,13]. $S(q,d)$ is the similarity value of document d to query q. CombMIN minimizes probability of 1), while CombMax minimizes probability of 2). CombMED tries to handle 1) and 2). The other three methods consider the relative similarity values given by each method, instead of selecting a value from the set of runs. CombSUM gives the numerical mean of similarity values, CombANZ ignore effects of single runs failing to retrieve relevant documents and CombMNZ provides higher weights to documents retrieved by multiple retrieval methods. Experiments have shown that CombSUM and CombMNZ usually offer the best increase in performances [2,9,13]. We considered these methods as they are very simple and are almost a standard in information retrieval, despite the existence of more advanced literature on this subject.

Name	Formula
CombMIN	$S(q,d) = \min_i(S_i(q,d))$
CombMAX	$S(q,d) = \max_i(S_i(q,d))$
CombMED	$S(q,d) = median(S_i(q,d))$
CombSUM	$S(q,d) = \sum_i S_i(q,d)$
CombANZ	$S(q,d) = CombSUM / \sum_{i \mid S_i(q,d) \neq 0} 1$
CombMNZ	$S(d) = CombSUM \times \sum_{i \mid S_i(d) \neq 0} 1$

Table 1. Classical combination strategies for multiple retrieval system results

5 Experiments

The goal of the experiments our threefold:

1. Test that the proposed incremental clustering approach for visual dictionary computation does not affect the performances of the retrieval systems.

2. Show that BOR is a suitable approach that can be as efficient as traditional BOVW.
3. Effectively combine the results of multiple systems to build a meta-search engine with increased performances.

The three points are addressed in the following subsections. We used three publicly available datasets, namely WANG, SIVAL and CALTECH101. WANG [15] is a subset of Corel dataset containing 1000 images classified in 10 different categories. We chose WANG to compare our results with the in-depth evaluation of features for image retrieval of Deselaers *et. al.* [4]. SIVAL is a more challenging dataset which has been specifically built for *localized* CBIR, *i.e.* where the user is interested in retrieving images of a specific object[21]. 25 objects have been pictured at different locations on the same set of complex backgrounds. There are 1500 images in this dataset. Using this dataset, we will show that the BOR representation is suitable for performing local queries as is the BOVW. Finally, we used CALTECH101 dataset [6] to provide larger scale experiments. CALTECH101 contains approximately 9000 images grouped into 101 categories. The number of images per category differs from one another.

We fixed the parameters to compute BOR visual dictionaries for all the datasets to the following:

- 7 different segmentations per image. 5 segmentations were computed with the algorithm of Felzenszwalb and Huttenlocher [7], tuning the parameters to produce different levels of region granularity. 2 segmentations were computed using Turbopixels [14]. We the number of regions to k and $2k$ with $k = 50$ for WANG, $k = 1000$ for SIVAL and $k = 100$ for CALTECH101. We used different k because images in SIVAL have much larger resolution than images in WANG. Examples are given in figure 2.
- 2 low level descriptors were computed from the regions: HSV color histogram and the histogram of Local Binary Patterns (LBP) [20] as a texture descriptor.
- 5 different size for visual vocabularies were used: $\{500, 1000, 2000, 5000, 10000\}$ words.

Hence, we computed $5 \times 2 \times 7 = 70$ BOR vocabularies for each dataset. We also computed the BOVW by the clustering of SURF points [3] using the same dictionary sizes. When computing the actual image signature for the BOR representation, we can weight the contribution of each word either by the area of the regions or by the number of regions in the image. We tried both approaches. Thus we have $70 \times 2 + 5 = 145$ visual vocabularies for each dataset. In the case of SIVAL dataset, we will consider 2 cases: global queries, where the BOR and BOVW signature are computed using the full image, and local queries, where they are computed considering only the regions or keypoints that are inside the object bounding box, that have been manually annotated.

For all datasets, we computed the visual vocabularies using the whole dataset, as no supervised learning is employed which would require the definition of a training and test set. We evaluate the Mean Average Precision (MAP) to asses

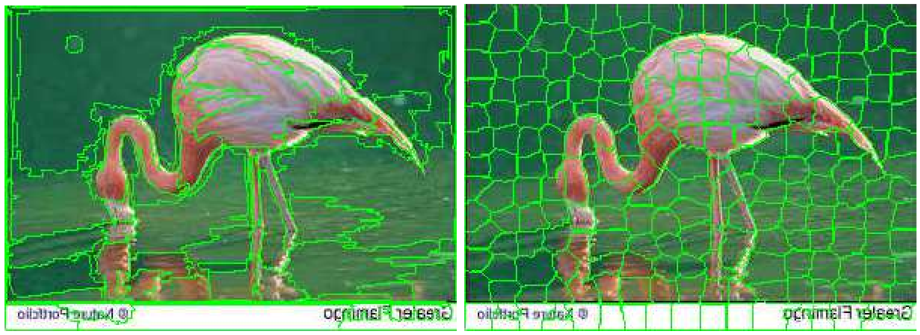

Fig. 2. Example of segmentation with Felzenszwalb [7] (left) and TurboPixels [14] (right). Image from CALTECH101 flamingo category.

the performances of the systems. The MAP is evaluated using every image of the datasets as query.

5.1 Incremental Clustering

We compared k-means clustering and the proposed incremental clustering on WANG and SIVAL, as it was not possible to compute k-means on CALTECH101 due to memory requirements. For incremental clustering and k-means, we used k-means++ initialisation of centroids. We set up the incremental clustering to ensure that no clusters were added incrementally, in order to have a fair comparison using vocabularies of the same size. Figure 3 shows the results obtained with the two methods on the 145 systems for each. The curve of performances of systems built with k-means and incremental clustering always have the same shape. This shows that the incremental clustering does not affect the retrieval performances, despite the lower computational and memory requirements. The incremental clustering curve seems even slightly higher than the k-means curves in all cases (*i.e.* WANG, SIVAL global and SIVAL local). Note that the increase is not really significant and we are far from claiming that incremental clustering should be favored to build retrieval systems with improved retrieval efficiency, but use it for computational efficiency.

5.2 Retrieval Efficiency Using Bag-Of-Regions Vocabulary

In figure 4 we compare the results obtained by the HSV, LBP and original BOVW with SURF descriptors. The results are presented in the same way as in figure 3, *b*y increasing MAP scores. There are only 5 systems based on SURF and 70 systems based on HSV and LBP, which is why the SURF curve is shorter.

For WANG dataset, the results obtained with SURF descriptors is outperformed by HSV and LBP. More than half of the HSV and LBP systems are better than SURF. The best results are achieved by HSV descriptors, which is in accordance with the experiments of [4], where global color histograms were the best

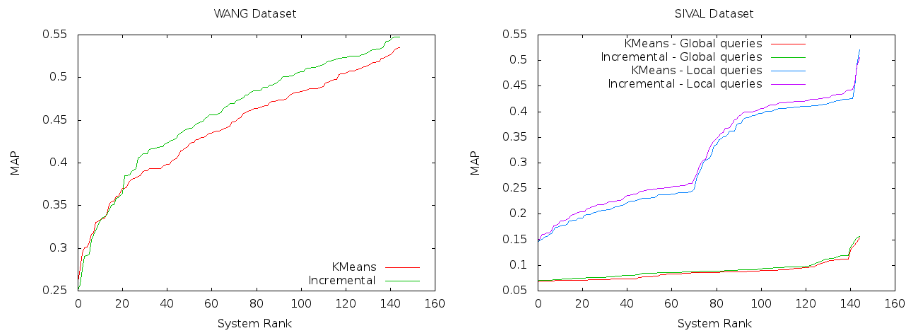

Fig. 3. Comparative performances of retrieval systems built with k-means clustering and the proposed incremental clustering scheme. x-axis denotes the system rank (from worst to best), y-axis the system MAP score

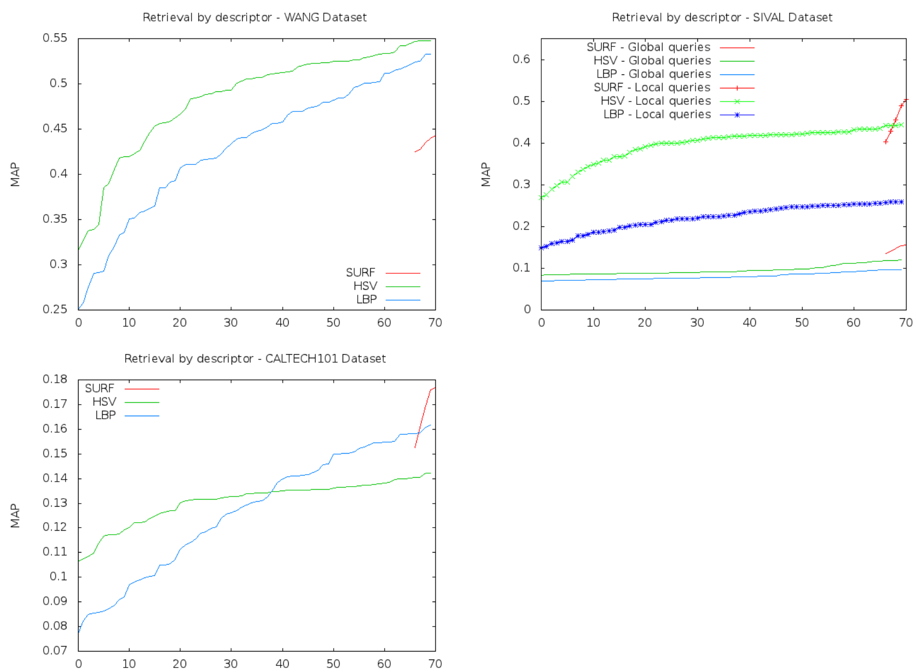

Fig. 4. Performances of retrieval systems using HSV, LBP and SURF descriptors.

features for this dataset. Note that the best MAP reported in [4] was 0.505, while our best system obtains 0.548 MAP, and many systems outperform 0.505. The best run based on SURF is 0.443. For the SIVAL and CALTECH101 datasets, the best overall systems are obtained with SURF (0.157 on SIVAL global, 0.505 on SIVAL local, 0.177 CALTECH). However, SURF does not clearly outperforms color and texture systems. In figure 5, we detail the MAP per categories of the best performing SURF, HSV and LBP systems. The retrieval performances of the descriptors is linked with the queries. The most explicit example can be seen in CALTECH101 dataset, where there are 2 blue peaks corresponding to a high MAP obtained with HSV. While HSV is overall the worst descriptor(figure 4), it is particularly well suited for **car sides** and **leopards** categories. This is surprising for the first category, but is actually due to an artifact of the dataset. All **car sides** images are black and white, while other are color images.

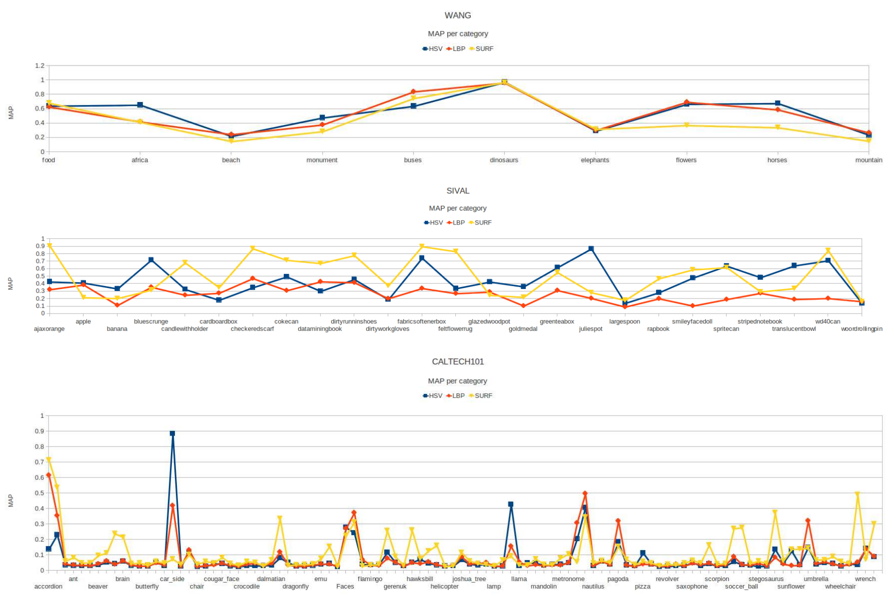

Fig. 5. MAP per category for the best performing SURF, HSV and LBP systems

5.3 Combining Multiple Retrieval Systems

The results obtained in the previous section naturally let us think that the combining the runs could greatly improve the efficiency of the method. In this section, we present the results of combined systems using the strategies shown in table 1. Since we returned for each query the ranked list of all images in the database, CombANZ and CombMNZ are not applicable as they are equivalent to CombSUM. We defined two sets of systems to combine. In the first set, we combine the best HSV, LBP and SURF runs. In the second set, we combine the top 5 systems overall, independently of the descriptors. Results are shown in table 2. It is clear from table 2 that combining the different systems is beneficial to the overall retrieval efficiency. The best performing system, shown in bold,

Single runs			Combined runs							
SURF	HSV	LBP	(SURF,HSV,LBP)				(Top5)			
			MIN	MAX	MED	SUM	MIN	MAX	MED	SUM
WANG										
.443	.548	.533	.539	.551	. 522	**.639**	.553	.551	.556	.563
SIVAL Global										
.157	.120	.097	.131	.120	.122	.132	.135	**.157**	.148	.149
SIVAL Local										
.505	.443	.260	.461	.437	.508	.555	.541	.475	.579	**.603**
CALTECH										
.177	.142	.162	.173	.178	.178	.197	.172	.178	.186	**.223**

Table 2. MAP results for different combination strategies

is always the result of a meta-search. As reported in [9,13], CombSUM seems to be the best choice. In these experiments, we have fixed a set of systems to combine. The choice of this set is still an interesting research challenge. The greatest increase in result is achieved when selecting high performing single systems with complementary results. This is the case when combining SURF, HSV and LBP for WANG, while the 5 best systems on this dataset are HSV-based and the improvement is small. Results are greatly improved for SIVAL-Local and CALTECH101 when combining the top 5, which are composed of 2 HSV and 3 SURF, 4 SURF and 1 LBP respectively. To our knowledge, there is no comparable CBIR results on SIVAL dataset. Ramanathan *et. al.* reported 0.0978 MAP on CALTECH101 using a quadtree extended vector space model [22], but they queried the system using only 10 images per category. This corresponds to a 19.2% of increase compared to the regular BOVW in their experiments (8.3% increase using spatial pyramid matching). We increase the BOVW results by 26%. An example for a SIVAL-local query is shown in figure 6.

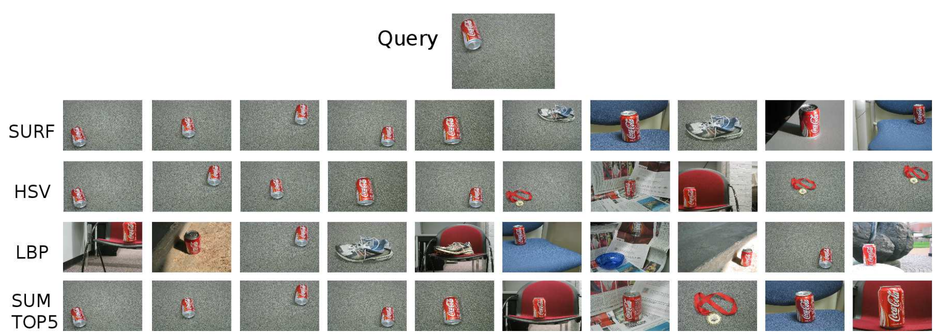

Fig. 6. Example query from SIVAL Dataset (local). First row: best SURF run. Second row: best HSV run. Third row: best LBP run. Fourth row: CombSUM of the top 5 ranking system results, showing a higher average precision for the query.

6 Conclusion

In this work, we proposed a new kind of image signatures based on the clustering of image regions, BOR. We have demonstrated that the BOR signatures are well adapted to build efficient CBIR systems, that can outperform traditional BOVW signatures on some datasets. BOR signatures are a good counterpart of BOVW, and can be more appropriate depending on the specific queries. Upgrading the traditional BOVW framework to BOR leads to a multiplication of possible visual dictionaries. With this increase, the computational time to build the dictionaries becomes problematic. We proposed to rely on a single pass, incremental clustering algorithm with appropriate k-means++ bootstrapping method. Experiments have shown that visual dictionaries built upon the incremental clustering lead to results as efficient as k-means. Finally, we proposed to combine the results of the different systems using well known meta-search techniques from text information retrieval. In all the cases, the combined results have favorably impacted the performances of the single retrieval systems.

This work has demonstrated that a system based on BOVW and BOR signatures is already very effective. Yet, it opens the way for further research on at least two different aspects. The impact of the segmentation algorithms on the retrieval performances would be interesting to study. Is an accurate segmentation really necessary? Should we over-segment, under-segment, or both? Such a study could allow to fix a few set of segmentation parameters that enable to capture complementary information from the BOR signatures. The second research topic that we will investigate is concerning the meta-search. In our experiments, we just fixed the set to combine under what seemed to be a reasonable choice. However, we know for sure that those choices are far from optimal. Moreover, the choice of an optimal combination is likely to differ between queries. In the future we will investigate meta-models that allow the weighted combination of each single system (*e.g.* weighted Borda-Fuse of Aslam and Montague [2]), where the weights are interactively computed using relevance feedback.

References

1. D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *SODA07*, pages 1027–1035, 2007.
2. J. A. Aslam and M. Montague. Models for metasearch. In *ACM SIGIR*, 2001.
3. H. Bay, Andreas Ess, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110:346–359, 2008.
4. T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: An experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.
5. P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV'02*, pages 97 – 112, 2002.
6. L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR'04*, 2004.

7. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004.
8. M. Flickner, H. Sawhney, and W. Niblack et al. Query by image and video content: the qbic system. *IEEE Computer*, 28(9):23–32, 1995.
9. E.A. Fox and J. A. Shaw. Combination of multiple searches. In *Third Text Retrieval Conference (TREC-1994)*, 1994.
10. D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *CVPR'07*, pages 1–8, 2007.
11. T. Hofmann. Learning the similarity of documents : an information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, 2000.
12. L. Ladicky, C. Russel, and P. Kohliwu. Associative hierarchical crfs for object class image segmentation. In *ICCV09*, 2009.
13. Joon Ho Lee. Analyses of multiple evidence combination. In *ACM SIGIR'97*, 1997.
14. A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi. Turbopixels: Fast superpixels using geometric flows. *IEEE PAMI*, 31:1–9, 2009.
15. J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE PAMI*, 25:1075–1088, 2003.
16. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.
17. Edwin Lughofer. Extensions of vector quantization for incremental clustering. *Pattern Recognition*, 41:995–1011, 2008.
18. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 27:1615–1630, 2005.
19. D. Nister and H. Stewenius. Scalable recognition witha vocabulary tree. In *CVPR2006*, 2006.
20. T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, 24:971–987, 2002.
21. R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts. Localized content based image retrieval. *IEEE PAMI*, 30:1902–1912, 2008.
22. V. Ramanathan, S S. Mishra, and P. Mitra. Quadtree decomposition based extended vector space model for image retrieval. In *IEEE WACV*, 2011.
23. J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV'03*, volume 2, pages 1470–1477, 2003.
24. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE PAMI*, 22(12):1349–1380, 2000.
25. F. Souvannavong, B. Merialdo, and B. Huer. Region-based video content indexing and retrieval. In *CBMI05*, 2005.
26. J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV2010*, 2010.
27. Remi Vieux, Jenny Benois-Pineau, Jean-Philippe Domenger, and Achille Braquelaire. Segmentation-based multi-class semantic object detection. *Multimedia Tools and Applications*, Available Online Oct. 2010:1–22, 2010.
28. T. Yeh, J. Lee, and T. Darrell. Adaptive vocabulary forests br dynamic indexing and category learning. In *ICCV'07*, pages 1 –8, oct. 2007.