A Survey of Extended Methods to the Bag of Visual Words for Image Categorization and Retrieval

Mouna Dammak, Mahmoud Mejdoub and Chokri Ben Amar

REGIM-REsearch Groups on Intelligent Machines, National Engineering School of Sfax,
University of Sfax, BP 1173, 3038 Sfax, Tunisia

Keywords: Image Representation, Spatial Neighboring Relation, Bag of Visual Words, Encoding and Pooling, Graph Representation, Image Categorization.

Abstract: The semantic gap is a crucial issue in the enhancement of computer vision. The user longs for retrieving images on a semantic level, but the image characterizations can only give a low-level similarity. As a result, recording a stage medium between high-level semantic concepts and low-level visual features is a stimulating task. A recent work, called Bag of visual Words (BoW) have arisen to resolve this difficulty in greater generality through the conception of techniques genius relevantly learning semantic vocabularies. In spite of its clarity and effectiveness, the building of a codebook is a critical step which is ordinarily performed by coding and pooling step. Yet, it is still difficult to build a compact codebook with shortened calculation cost. For that, several approaches try to overcome these difficulties and to improve image representation. In this paper, we introduce a survey investigates to cover the inadequacy of a full description of the most important public approaches for image categorization and retrieval.

1 INTRODUCTION

The Bag of visual Words (BoW) is a method which offers a Mid-Level Descriptors (MLD) which facilitates the reduction of the semantic gap (Smeulders et al., 2000) between the Low-Level Descriptors (LLD), withdrawn from an image, and the High-Level Descriptors (HLD) concepts to be classified. The building of the BoW model can be fractured into chained stages of encoding and pooling . The encoding step assigns the local descriptors onto the codebook components while the pooling step aggregates the assigned words into a vector. We can distinguish three problems in the standard visual word, which may be the core factors of their restricted descriptive competence:

1. K-means method based visual vocabulary building can not conduct to very efficient and compact visual word assembly;
2. An individual visual word includes restricted details. So, it is not effective in describing the features of objects and scenes;
3. Ignore spatial information.

There exist some challenges that have been advanced to improve the performance of the conventional BoW paradigm and to integrate spatial information. We can classify these methods into two main categories: the first category attempts to improve the generation of the visual vocabulary (Farquhar et al., 2005; Avrithis and Kalantidis, 2012); the second category contains techniques that add spatial information over the BoW which have been proven to enhance the performance of scene classification and retrieval (Xie et al., 2012; Jiang et al., 2012). Therefore, the aim of this paper is to review the most developed work of BoW for image categorization and retrieval.

The paper starts with describing the general process of building the bag of visual words. Subsequent sections discuss the advanced approaches for bag of visual word model. Step one of the review presents recent approaches based on coding step. Step two of the review presents recent approaches based on pooling step. In the last section, we present a conclusion.

2 THE BASELINE SYSTEMS OF BAG OF WORDS REPRESENTATION

The standard pipeline to obtain the Bag of visual

Dammak M., Mejdoub M. and Ben Amar C..
A Survey of Extended Methods to the Bag of Visual Words for Image Categorization and Retrieval.
DOI: 10.5220/0004750506760683
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 676-683
ISBN: 978-989-758-004-8

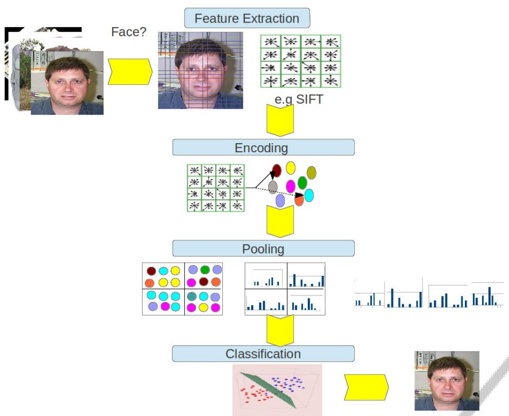

Figure 1: General schematic overview of the Bag of visual Words framework

Words, consists firstly to group a large prototype of low level descriptor from a collection of training images. The second stage is based on clustering these descriptors using the K-means clustering. The K cluster centers draws the visual words. The K value is a user-supplied parameter and represent the size of the vocabulary. When the codebook is builded, a new image is calculated in the following way: extraction of low level descriptor, attribution of the descriptors on the codebook established on a training collection, and computation of a histogram that counts the number of times of occurrence of the codebook visual words (see figure 1) . In this section, we will detail these stages specifying the different approaches.

2.0.1 Point/Region Detection

In the literature, we can distinguish two different types of patch based image representations: Interest Points (IP) and dense sampling. On the one hand, IP concentrate on interesting positions in the image and contain diverse ranks of viewpoint and illumination invariance, bringing about improved repeatability outcome, such as corners or blobs position, whose scale and shape are determined by an algorithm of feature detection. Dense sampling, on the other hand, which is composed of patches of adjusted size and shape are located on a constant grid and can be repeated on various scales, provides improved coverage of the image, a prevailing number of features per image, and simple spatial relations between features. By integrating both criteria, the authors (Tuytelaars, 2010) have proposed a hybrid scheme called dense interest points which they have started from densely sampled patches even enhance their location and scale parameters positionally.

2.1 Feature Extraction

2.1.1 Local Descriptor

Various and recent feature descriptors have been greatly drawn in the general visual recognitions such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004), Speed Up Robust Features (SURF) (Bay et al., 2006), Binary Robust Independent Elementary Features (BRIEF) (Su and Jurie, 2011; Calonder et al., 2012), etc. Due to the achievement of SIFT, image local features have been greatly applied in a variety of computer vision and image processing applications.

The SIFT descriptor proposed by Lowe describes the local shape of a region using edge orientation histograms. The gradient of an image is shift-invariant: taking the derivative cancels out offsets. Under light intensity changes, i.e . a scaling of the intensity channel, the gradient direction and the relative gradient magnitude remain the same. Because the SIFT descriptor is normalized, the gradient magnitude changes have no effect on the final descriptor. Particularly, various recent works have taken advantage of SIFT to develop advanced object classifiers. The SIFT descriptor is not invariant to light color changes, because the intensity channel is a combination of the R , G and B channels. Color features provide powerful information for object and scene classification, indexing and retrieval. Due to these two important causes, several descriptor color extensions of SIFT are proposed including HSV-SIFT (Bosch et al., 2008), OpponentSIFT (van de Sande et al., 2010), RGB-SIFT (van de Sande et al., 2010). Furthermore, SIFT is not a flip invariant. As a consequence, the descriptors extracted from two identical but flipped local patches could be completely different in feature space. For that, several invariant descriptors are based on improvement partitioning scheme of local region including Mirror and Invert invariant SIFT (MI-SIFT) (Ma et al., 2010) and Neat Flip Invariant Descriptor (FIND)(Guo and Cao, 2010).

2.2 Codebook Generation

The authors (Sivic and Zisserman, 2003; Csurka et al., 2004) have creatively proposed to cluster the low-level features with the K-means clustering, which is the most dominant method, to get the Bag of visual Words. Given a set $x_1,x_2,\ldots,x_N \in \mathbb{R}^D$ of N training descriptors. K-means searchs K vectors $\mu_1,\mu_2,\ldots,\mu_K \in \mathbb{R}^{\mathbb{D}}$ and a data-to-means assignments $q_1,q_2,\ldots,q_N \in \{1,2,\ldots,K\}$ such that the additive approximation error $\sum_{i=1}^{N} \|x_i - \mu_{q_i}\|^2$ is minimized. An extended clustering based on a generative model, called the Gaussian Mixture Model (GMM) have been proposed by (Farquhar et al., 2005). This model is characterized by a continuous histogram representation contrast to a discrete histogram representation caused by features which are assigned to all words probabilistically. GMM utilizes a mixture of gaussians gathering a linear combination of gaussian densities. Each gaussian density has its own mean and covariance. The number of clusters is proportional to the number of gaussians. So, clustering of data can be accomplished by means of estimating the parameters connected to the latent variable of the Gaussian mixture. Expectation–Maximization (EM) algorithm (G. Mclachlan, 2000) can be utilized to determine maximum likelihood estimators in gaussian mixtures with latent variables. GMM method clusters population and shape, but it considers pairwise interaction of all data with all clusters and it is tardier to converge. To overcome this limitation, the authors (Avrithis and Kalantidis, 2012) have proposed an approximate version, called Approximate Gaussian Mixtures (AGM), to large scale visual vocabulary learning. In this case, opposite to the usage of model GMM, descriptors of indexed images are adequate only to their nearest visual word to retain enough index sparse. They suggest a variant of EM that can converge rapidly while dynamically estimating the number of components. They employed approximate nearest neighbor search to speed-up the Expectation step and exploit its iterative nature to make search incremental, boosting both speed and precision.

2.3 Encoding and Pooling Phase

Having the keypoints detected, the features extracted and the visual words generated, the final step of extracting the representation from images is based on two successive stages of coding and pooling, maintaining the discriminating potential of the local descriptors. The coding step assigns the local descriptors onto the bag of visual words while the pooling step aggregates the assigned words into a vector. Let X be a set of D-dimensional local descriptors extracted from an image, i.e. $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{\mathbb{D}\times\mathbb{N}}$. Given a visual dictionary with K visual words, i.e. $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{\mathbb{D}\times\mathbb{N}}$. We use α_n to denote the code vector. The dimension of α_n is the same as the size of D except Fisher kernel representation.

The coding step can be modeled by an activation function for the codebook, stimulating each of the visual words corresponding to the local descriptor. In the BoW model, the coding function stimulates only the closest codeword to the descriptor.

$$\alpha_{n,k} = 1 \; iff \; k = \arg\min_{k\in\{1\cdots K\}} \| x_n - d_k \|^2 \qquad (1)$$

where $\alpha_{n,k}$ is the k^{th} element of the encoded vector α_n. This scheme corresponds to a hard coding or Vector Quantization (VQ) over the dictionary. The generating binary vector is very sparse, but it undergoes sensitivities when the descriptor is coded on the boundary of proximity of diverse bag of visual words (Gemert et al., 2010).

As a result, new methods to that approaches have been recently emerged. Sparse coding (Yang et al., 2009) modifies the optimization function by jointly considering reconstruction error and sparsity of the vector, using the famous attribute that regularization with the $l1-norm$, for a large enough regularization parameter λ, yields sparsity:

$$\alpha_n = \arg\min_{\alpha\in\mathbb{R}^{\mathbb{K}}} \| x_n - D\alpha \|_2^2 + \lambda\|\alpha\|_1$$

where λ penalizes the $l1-norm$ regularizes, which controls the sparsity of α . Powerful tools have been suggested to get tractable settlings (Mairal et al., 2010). Another way based on a soft assignment to each visual word, called soft coding (Gemert et al., 2010). It gives weight according to similarities between descriptors and codewords. In the pooling step, soft coding leads to ambiguities because of the superposition of the elements. This results in dense code vectors, which is unfavorable. So, diverse intermediate approaches, called semi-soft coding (Liu et al., 2011), have been suggested, often performing the soft assignment just to the K nearest neighbors of the input feature. Contrary to the sparse coding, Locality-constrained Linear Coding (LLC), proposed by Wang et al. (Wang et al., 2010), enforces locality instead of sparsity and this leads to smaller coefficient for the basis vectors far away from the local feature x_n. The coding coefficients are obtained by solving the following optimization

$$\alpha_n = \arg\min_{\alpha\in\mathbb{R}^{\mathbb{K}}} \| x_n - D\alpha \| + \lambda \| \beta_n \odot \alpha \|^2 \qquad (2)$$

where $\odot$ denotes the element-wise multiplication and β_n is the locality adaptor that gives weights for each basis vector proportional to its similarity to the input descriptor x_n. The distance metric used :

$$\beta_n = \exp\left(\frac{dist\,(x_n, D)}{\sigma}\right)$$

where $dist\,(x_n, D) = [dist\,(x_n, d_1)\,, dist\,(x_n, d_2)\,, \ldots, dist\,(x_n, d_K)]^{\mathsf{T}}$ and $dist\,(x_n, d_k)$, is the Euclidean distance between x_n and d_k. σ is used for adjusting the weighted decay speed for the locality adaptor.

The distance regularization of LLC effectively performs feature selection, and in practice only those bases close to x_n in feature space have non-zero coefficients. This suggests to develop a fast approximation of LLC by removing the regularization completely and instead using the K-nearest neighbours $(\tilde{K})$ of x_n $\left(\tilde{K} < D < K\right)$ as a set of local bases D_i.

$$\alpha_n = \arg\min_{\alpha\in\mathbb{R}^{\mathbb{K}}} \| x_n - D_i\tilde{\alpha}_n \| \quad st. = 1\forall i \tag{3}$$

This reduces the computation complexity from $\theta\left(K^2\right)$ to $\left(K+\tilde{K}^2\right)$ and the nearest neighbours can be found using ANN methods such as kd-trees. Perronnin et al. (Csurka and Perronnin, 2010) have presented the Fisher Vector (FV) extending the BOW. They have hypothesized that the descriptors can be modeled by a probability density function. A BoW has been learnt by a GMM model. They have captured the average first and second order differences between the image descriptors and the centres of a GMM. They have concatenated the mean and the second order for all K Gaussian components, giving an encoding of size $2DK$ where D represents local descriptor size. They infer that FV describes how the set of descriptors deviates from an average distribution of descriptors, modeled by a parametric generative model. In (Zhou et al., 2010), the authors have proposed a Super Vector (SV) approach to extend VQ by a function approximation scheme and it is similar to the fisher encoding. There are two variants of SV, based on hard assignment to the nearest codeword or soft assignment to several near neighbours. For the hard SV, a feature is assigned to the nearest visual word μ_k, which obtained K-means clustering algorithm, and for the soft SV, this method eventually result in aggregating the difference vectors $x_n - d_k$ around the visual word d_k. This provides an encoding of size $K(D+1)$. Compared to the fisher encoding, the super vector encoding: (1) investigates only the first order distinctions d_k between features and cluster centres; (2) accumulates the elements s_k which represent the weight of each cluster; (3) normalizes each cluster by the square root of the posterior probability instead of the prior probability.

The BoW approach involves a large codebook of several thousands of visual words. However, clustering high dimensional feature spaces and large scale is not an easy task. To tackle this problem, Jgou et al. (H.Jeou et al., 2012) have suggested vector representations called Vector of Locally Aggregated Descriptors (VLAD) that use smaller codebooks. Instead of using GMM to model the feature distribution, VLAD uses a K-means clustering algorithm to build a codebook. Then, the descriptors are voted on their nearest codewords. A vector for each visual word is the accumulation of the differences between the nearest descriptors and itself resulting a vector size D. the final image representation concatenates the K vectors on the codebook and generates a vector of size KD.

The authors (Picard and Gosselin, 2013) have extended the VLAD approach, called Vectors of Locally Aggregated Tensors (VLAT), by adding an aggregation of the tensor product of descriptors. As in VLAD, the visual codebook has been built by clustering the descriptor using k-means. This representation is included several formulas. The first formula, as in VLAD, is the aggregate of differences between the nearest descriptors and a visual word. The second formula is the total of self tensor product of the nearest descriptors assigned to a visual word. in general, a higher order p of tensor products on centred descriptors can be calculated to add tensor formulas to the signature. In practice, the number of formulas is limited to the second order. The last image representation is the vector concatenating all the formulas in vectors for all visual words. The images having different numbers of descriptors to be compared, a further $l-2$ normalization of the signatures is achieved. Notice that DK-dimensional vector representation.

3 INVESTIGATIONS OF INTEGRATION SPATIAL INFORMATION INTO BAG OF VISUAL WORDS METHODS

3.1 Approaches based on Pairwise Features

Recent works (Zhang and Mayo, 2010; Morioka and Satoh, 2010b; Wang et al., 2012; Khan et al., 2012; Morioka and Satoh, 2010a; Herve and Boujemaa, 2009; Morioka and Satoh, 2011) have been founded on pairs of visual words. A codebook of size K $\{c_k\}_K$, has been learned using unsupervised learning, from a haphazardly sampled collection relevantly to the descriptors. Then, every descriptor d_i is assigned to the closest cluster c_k in the feature space. To integrate the spatial information, their approaches are differtents: In (Herve and Boujemaa, 2009), firstly, Herve et al. have constructed a base vocabulary containing K words, then, they built a $K\left(\frac{K+1}{2}\right)$ word pair vocabulary, called Quadratic Pairwise Codebook (QPC), to capture spatial information between words. They have considered the pairs which the distance between the two patches centers is below the given radius.

They have simply accumulated the pairs in an histogram. To overcome quadratic number of possible pairs of visual words, Morioka and Satoh (Morioka and Satoh, 2010a) have proposed a compact codebook called a Local Pairwise Codebook (LPC). Contrust to previous appraoch based on quantize the descriptors to learn a set of visual words, they have started by joint feature space. After that, they have applied a clustering algorithm to build a compact codebook. Then, they have computed a histogram. Then, they have combined it with spatial pyramid matching kernel to demonstrate that local and global spatial information complement each other.

These authors (Morioka and Satoh, 2010b) have extended LPC by adding directional information to the representation to produce two new approaches called Directional Local Pairwise Bases (DLPB) and Directional Local Pairwise Codebook (DLPC). In DLPB, they have used a sparse coding to learn a compact collection of bases appropriating interconnection between descriptors. Moreover, these bases have been learned for each quantized direction. Thus, it adds to the representation explicit directional information. For DLPC, which is a variant of DLPB, a K-means is used to replace sparse coding to build specific directional codebooks. For every directional codebooks, the authors have computed spatial pyramid matching kernel to extract the average of the kernels.

LPC achieves a compact codebook of pairs of spatially close local descriptors. As a result, it is not considered as scale invariant and it is also appropriate only for densely sampled local features. Contrary to that, the Proximity Distribution Kernel (PDK) method is characterized by a scale invariant and robust representation. It, then, captures rich spatial proximity information between local features, but the number of visual words increases quadratically. Inspired by the two above mentioned techniques, the authors (Morioka and Satoh, 2011) have unified of the LPC and the PDK to represent a new method called the Compact Correlation Coding (CCC) to combine the powers of both techniques. Compared to the PDK, CCC performs a more general and compact codebook. Yet, it captures robust spatial proximity distribution of local features and scale invariant which cannot be achieved under the properties of LPC.

3.2 Approaches based on High Orders Features (Visual Phrases)

Although the studies of second order features are intensive, to capture more the spatial information, some works (Zheng et al., 2008; Zhang and Chen, 2009; Bingbing et al., 2013) are particularly interested in how to model high-order local features. In (Zheng et al., 2008), the local spatial neighborhood have been extracted for each local region. The FP-growth algorithm has been applied to perform the Frequent Itemset Mining (FIM) task to present the visual words. In order to integrate both the local proximity of visual words and co-occurrence information, the authors have defined the visual synset as a probabilistic concept of visual words, in which the latter has been learned through supervised learning. In (Zhang et al., 2011a), the authors have suggested the descriptive visual words and descriptive visual phrases as the visual analogical to text words and phrases, when visual phrases attribute to the repeatedly occurring visual word pairs. The co-occurring is computed between two visual words inside a short distance. For each image category, they have defined the Descriptive Visual Words (DVW) candidates as the comprised visual words and they have defined the Descriptive Visual Phrases (DVP) candidate generation by agreeing it with the rotation invariant spatial histogram. The co-occurrence frequency can be computed by counting the frequency of co-occurrence within the spatial distance between two visual words in the same category. In (Xie et al., 2012), Xie et al. have extracted SIFT and Edge-SIFT descriptors and they have combined them to build codebook. Then, they have generated a geometric visual phrases by taking a phrase as a set disordered neighbors of visual words. A max-pooling step is performed on the whole phrase and they have applied spatial weighting based on a smoothed edgemap. The authors (Cao et al., 2010) have proposed an ordered bag of features based on projecting features onto certain lines or circles which are able to capture basic geometric information in images. These representations are the basis of the spatial bag of words. They have treated the same operations for histogram features, i.e. calibration, equalization and decomposition to capture more and more typical transformations of image including translation, rotation and scaling. Then, they have adopted the RankBoost algorithm (Freund et al., 2003) to select the most effective configuration. In (Zhang et al., 2011b), the authors have integrated the algorithm proposed in (Zhang and Chen, 2009) to identify the co-occurring the geometry-preserving visual phrases (GVP) in two images. Added to co-occurrences, the GVP method captures the local and long-range spatial layouts of the visual words. To measure the GVP similarity value of two images, they have calculated the offset $\triangle_{(x,y)}$ for each pair of the same word in these images. Then, a vote has been yielded on the offset space at $\triangle_{(x,y)}$. On the offset space, K votes locating at the same place corresponding to a co-occurring GVP of length K. After obtaining the dot product, the similarity of the two images is the dot product dividing the L2-norms.

The authors (Jiang et al., 2012) have proposed a new visual phrase selection approach based on random partition of images. After extracting local invariant features, they have randomly split the image multiple times to form a pool of overlapping image patches. Each patch groups the local features inside it and is described by a visual phrase. primarily, for each local descriptor, they have yielded a number of Randomized Visual Phrases (RVP) varying shapes and sizes according to its spatial contexts. For each RVP, they have independently computed matching score between the test image and the query image, and they have dealed with it as the voting weight of the appropriate patch. The final reliability score of each pixel has been computed as the expectation of the voting weights of all patches that comprise this pixel. By determining the pixel-wise voting map, the similarity image can eventually be recognized. In (Bingbing et al., 2013), the authors have proposed a model of high-order local spatial context called Spatialized Random Forest (SRF). SRF can explore much more complicated and informative local spatial patterns randomly, applying spatially random neighbor selection and random histogram-bin partition during the tree construction. A set of informative high-order local spatial patterns are drifted, because of the discriminative capability test for the random partition in each tree node's division procedure. Consequently, new images have been encoded by calculting the repetitions of these discriminative local spatial patterns.

3.3 Approaches based on Graph and Graph Matching

To take into account the spatial constraints in images, several authors (Quack et al., 2007; Bowen et al., 2012; Kisku et al., 2010; Jaechul Kim, 2010; Duchenne et al., 2011) have proposed the graph matching technique to establish the correspondences between images. Visual graphs provide powerful structural models but their use in image classification has been limited due to the difficulties of matching between graphs. In (Jaechul Kim, 2010), Kim et al. proposed a dense feature matching. To match two images, they segment one and unsegment another. Then, they find correspondences between points within each region of the segmented image and some subsets of those within the unsegmented image. Layout consistency is meanwhile efficiently enforced in each of the region-to-image match group via an objective solvable with dynamic programming. This method was extended by Duchenne et al. (Duchenne et al., 2011). They formulated image graph matching as an energy optimization problem. The graph nodes and edges represent the regions associated with a coarse image grid and their adjacency relationships. Visual graphs supply competent compositional patterns, however their application in image classification is restricted ensuing to the complexities of matching between graphs which is known to be NP-complete. In (Pham et al., 2012), an image is represented by SIFT descriptors for each keypoint extracted, color histograms and edge descriptor where a region is defined by grid partition , and HSV color value where a region is defined by sampling pixel. The second step represents each image as a graph generated by a set of weighted concepts and a set of weighted relations. For that, they build a visual vocabulary for each type of image representation by k-means clustering. Each region is defined by a visual word and two relation sets left_of and top_of are extracted from the two connected region for integrate the relationships between the regions. The third step is related to the fact that we want to retrieve relevant images to a given query. Therefore, they take into account the different types of image representations and spatial relations during matching by computing likelihood of two graphs using a language model framework. Visual graphs supply competent compositional patterns, however their application in image classification is restricted ensuing to the complexities of matching between graphs which is known to be NP-complete. In order to relax the graph matching condition, (Quack et al., 2007), (Wu et al., 2013) proposed to identify the similarity between two image graphs comparing subgraphs extracted from them rather than using graph matching. Quak et. al. (Quack et al., 2007) have used FIM to discover a set of distinctive spatial configurations of visual words to learn different object categories. In (Wu et al., 2013), the authors divide an image into a sets of spatial grids on several levels. Then, they defined a directed graph to describe the relationship between these grids which the grids are represented by the nodes, and the relation of grids is represented by the edges. After that, they construct a histogram on node reflects the occurrence of features in a block, and a histogram on edge reflects the occurrence of features which lie in one block and tend to shift into another.

4 CONCLUSIONS

The Bag of Visual Words has successfully been applied to various computer vision applications include image categorization and retrieval. The construction of BoW starts by the construction of local features. After that, two steps are necessary: Encoding and pooling. Despite their simplicity, the spatial information is ignored. In this paper, we introduces the methods that improve the construction of BoW such as LLC, Fisher vector, VLAD, VLAT and the methods that integrate the spatial information such as approches based on pairwise features (LPC, DLPC, CCC, ...), approaches based on visual phrases (Phreselet), and approaches based on graph.

REFERENCES

Avrithis, S. and Kalantidis, Y. (2012). Approximate gaussian mixtures for large scale vocabularies. In *European Conference on Computer Vision*, volume 7574, pages 15–28. Springer.

Bay, H., Tuytelaars, T., and Gool, L. (2006). Surf speeded up robust features. In *European Conference on Computer Vision*.

Bingbing, N., Shuicheng, Y., Meng, W., Kassim, A., and Qi, T. (2013). High order local spatial context modeling by spatialized random forest. *IEEE Transactions on Image Processing*, 22(2).

Bosch, A., Zisserman, A., and Muoz., X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions Pattern Analysis and Machine Intelligents*, 30:712–727.

Bowen, F., Du, E. Y., and Hu, J. (2012). A novel graph-based invariant region descriptor for image matching. In *EIT*.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., and Fua, P. (2012). Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298.

Cao, Y., Wang, C., Li, Z., Zhang, L., and Zhang, L. (2010). Spatial bag of features. In *CVPR*.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

Csurka, G. and Perronnin, F. (2010). Fisher vectors: Beyond bag-of-visual-words image representations. In *VISIGRAPP*, pages 28–42.

Duchenne, O., Joulin, A., and Ponce, J. (2011). A graph-matching kernel for object categorization. In *ICCV*.

Farquhar, J., Szedmak, S., Meng, H., and Shawe-Taylor, J. (2005). Improving bag-of keypoints image categorisation. Technical report, University of Southampton.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969.

G. Mclachlan, D. P. (2000). Finite mixture models.

Gemert, J., Veenman, C., and Geusebroek, J. (2010). Visualword ambiguity. *TPAMI*.

Guo, X. and Cao, X. (2010). Find: A neat flip invariant descriptor. In *20th International Conference on Pattern Recognition*, pages 515–518.

Herve, N. and Boujemaa, N. (2009). Visual word pairs for automatic image annotation. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME 09.

H.Jeou, Perronnin, F., Douze, M., Sanchez, J., Perez, P., and Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9).

Jaechul Kim, K. G. (2010). Asymmetric region to image matching for comparing images with generic object categories. In *CVPR*.

Jiang, Y., Meng, J., and Yuan, J. (2012). Randomized visual phrases for object search. In *CVPR*.

Khan, R., Barat, C., Muselet, D., and Ducottet, C. (2012). Spatial orientations of visual word pairs to improve bag-of-visual-words model. In *BMVC*.

Kisku, D. R., Rattani, A., Grosso, E., and Tistarelli, M. (2010). Face identification by sift-based complete graph topology. *CoRR*.

Leordeanu, M. and Hebert, M. (2005). A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on Computer Vision*, pages 1482–1489.

Liu, L., Wang, L., and Liu, X. (2011). In defense of soft-assignment coding. In *International Conference on Computer Vision*, ICCV '11, pages 2486–2493.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2).

Ma, R., Chen, J., and Su, Z. (2010). Mi-sift: mirror and inversion invariant generalization for sift descriptor. *International Conf. on Image and Video Retrieval*, pages 228–236.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60.

Morioka, N. and Satoh, S. (2010a). Building compact local pairwise codebook with joint feature space clustering. In *11th European conference on Computer vision*, ECCV10.

Morioka, N. and Satoh, S. (2010b). Learning directional local pairwise bases with sparse coding. In *BMVC*.

Morioka, N. and Satoh, S. (2011). Compact correlation coding for visual object categorization. In *ICCV*.

Pham, T., Mulhem, P., Maisonnasse, L., Gaussier, E., and Lim, J. (2012). Visual graph modeling for scene recognition and mobile robot localization. *Multimedia Tools Appl.*, 60(2).

Picard, D. and Gosselin, P. (2013). Efficient image signatures and similarities using tensor products of local descriptors. *Computer Vision and Image Understanding*, 117(6):680–687.

Quack, T., Ferrari, V., Leibe, B., and Gool, L. V. (2007). Efficient mining of frequent and distinctive feature configurations. In *International Conference on Computer Vision)*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, pages 1470–1477.

Smeulders, A., M.Worring, Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Su, Y. and Jurie, F. (2011). Semantic contexts and fisher vectors for the imageclef 2011 photo annotation task. In *CLEF (Notebook Papers/Labs/Workshop)*.

Tuytelaars, T. (2010). Dense interest points. In *Computer Vision and Pattern Recognition*, pages 2281–2288.

van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367.

Wang, L., Song, D., and Elyan, E. (2012). Improving bag-of-visual-words model with spatial-temporal correlation for video retrieval. In *21st ACM international conference on Information and knowledge management*, CIKM 12.

Wu, Z., Huang, Y., Wang, L., and Tan, T. (2013). Spatial graph for image classification. In *11th Asian conference on Computer Vision*, ACCV 12.

Xie, L., Tian, Q., and Zhang, B. (2012). Spatial pooling of heterogeneous features for image applications. In *20th ACM international conference on Multimedia*, MM 12.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition*.

Zhang, E. and Mayo, M. (2010). Improving bag-of-words model with spatial information. In *25th International Conference of Image and Vision Computing New Zealand*.

Zhang, S., Tian, Q., Hua, G., Huang, Q., and Gao, W. (2011a). Generating descriptive visual words and visual phrases for large scale image applications. *IEE Transacton on Imgage Processing*, 20(9).

Zhang, Y. and Chen, T. (2009). Efficient kernels for identifying unbounded-order spatial features. In *CVPR*.

Zhang, Y., Jia, Z., and Chen, T. (2011b). Image retrieval with geometry-preserving visual phrases. In *CVPR*.

Zheng, Y., Zhao, M., Neo, S., Chua, T., and Tian, Q. (2008). Visual synset: Towards a higher level visual representation. In *Computer Vision and Pattern Recognition*.

Zhou, X., Yu, K., Zhang, T., and Huang, T. (2010). Image classification using super-vector coding of local image descriptors. In *11th European conference on Computer vision*, ECCV'10, pages 141–154.