International Scholarly Research Network
ISRN Artificial Intelligence
Volume 2012, Article ID 376804, 19 pages
doi:10.5402/2012/376804

Review Article

Bag-of-Words Representation in Image Annotation: A Review

Chih-Fong Tsai

Department of Information Management, National Central University, Jhongli 32001, Taiwan

Correspondence should be addressed to Chih-Fong Tsai, cftsai@mgt.ncu.edu.tw

Received 26 August 2012; Accepted 19 September 2012

Academic Editors: F. Camastra, J. A. Hernandez, P. Kokol, J. Wang, and S. Zhu

Content-based image retrieval (CBIR) systems require users to query images by their low-level visual content; this not only makes it hard for users to formulate queries, but also can lead to unsatisfied retrieval results. To this end, image annotation was proposed. The aim of image annotation is to automatically assign keywords to images, so image retrieval users are able to query images by keywords. Image annotation can be regarded as the image classification problem: that images are represented by some low-level features and some supervised learning techniques are used to learn the mapping between low-level features and high-level concepts (i.e., class labels). One of the most widely used feature representation methods is bag-of-words (BoW). This paper reviews related works based on the issues of improving and/or applying BoW for image annotation. Moreover, many recent works (from 2006 to 2012) are compared in terms of the methodology of BoW feature generation and experimental design. In addition, several different issues in using BoW are discussed, and some important issues for future research are discussed.

1. Introduction

Advances in computer and multimedia technologies allow for the production of digital images and large repositories for image storage with little cost. This has led to the rapid increase in the size of image collections, including digital libraries, medical imaging, art and museum, journalism, advertising and home photo archives, and so forth. As a result, it is necessary to design image retrieval systems which can operate on a large scale. The main goal is to create, manage, and query image databases in an efficient and effective, that is, accurate manner.

Content-based image retrieval (CBIR), which was proposed in the early 1990s, is a technique to automatically index images by extracting their (low-level) visual features, such as color, texture, and shape, and the retrieval of images is based solely upon the indexed image features [1–3]. Therefore, it is hypothesized that relevant images can be retrieved by calculating the similarity between the low-level image contents through browsing, navigation, query-by-example, and so forth. Typically, images are represented as points in a high dimensional feature space. Then, a metric is used to measure similarity or dissimilarity between images on this space. Thus, images close to the query are similar to the query and retrieved. Although CBIR introduced automated image feature extraction and indexation, it does not overcome the so-called semantic gap described below.

The semantic gap is the gap between the extracted and indexed low-level features by computers and the high-level concepts (or semantics) of user's queries. That is, the automated CBIR systems cannot be readily matched to the users' requests. The notation of similarity in the user's mind is typically based on high-level abstractions, such as activities, entities/objects, events, or some evoked emotions, among others. Therefore, retrieval by similarity using low-level features like color or shape will not be very effective. In other words, human similarity judgments do not obey the requirements of the similarity metric used in CBIR systems. In addition, general users usually find it difficult to search or query images by using color, texture, and/or shape features directly. They usually prefer textual or keyword-based queries, since they are easier and more intuitive for representing their information needs [4–6]. However, it is very challenging to make computers capable of understanding or extracting high-level concepts from images as humans do.

Consequently, the semantic gap problem has been approached by automatic image annotation. In automatic image annotation, computers are able to learn which low-level features correspond to which high-level concepts.

Specifically, the aim of image annotation is to make the computers extract meanings from the low-level features by a learning process based on a given set of training data which includes pairs of low-level features and their corresponding concepts. Then, the computers can assign the learned keywords to images automatically. For the review of image annotation, please refer to Tsai and Hung [7], Hanbury [8], and Zhang et al. [9].

Image annotation can be defined as the process of automatically assigning keywords to images. It can be regarded as an automatic classification of images by labeling images into one of a number of predefined classes or categories, where classes have assigned keywords or labels which can describe the conceptual content of images in that class. Therefore, the image annotation problem can be thought of as image classification or categorization.

More specifically, image classification can be divided into object categorization [10] and scene classification. For example, object categorization focuses on classifying images into "concrete" categories, such as "agate", "car", "dog", and so on. On the other hand, scene classification can be regarded as abstract keyword based image annotation [11, 12], where scene categories are such as "harbor", "building", and "sunset", which can be regarded as an assemblage of multiple physical or entity objects as a single entity. The difference between object recognition/categorization and scene classification was defined by Quelhas et al. [13].

However, image annotation performance is heavily dependent on image feature representation. Recently, the bag-of-words (BoW) or bag-of-visual-words model, a well-known and popular feature representation method for document representation in information retrieval, was first applied to the field of image and video retrieval by Sivic and Zisserman [14]. Moreover, BoW has generally shown promising performance for image annotation and retrieval tasks [15–22].

The BoW feature is usually based on tokenizing keypoint-based features, for example, scale-invariant feature transform (SIFT) [23], to generate a visual-word vocabulary (or codebook). Then, the visual-word vector of an image contains the presence or absence information of each visual word in the image, for example, the number of keypoints in the corresponding cluster, that is, visual word.

Since 2003, BoW has been used extensively in image annotation, but there has not as yet been any comprehensive review of this topic. Therefore, the aim of this paper is to review the work of using BoW for image annotation from 2006 to 2012.

The rest of this paper is organized as follows. Section 2 describes the process of extracting the BoW feature for image representation and annotation. Section 3 discusses some important extension studies of BoW, including the improvement of BoW per se and its application to other related research problems. Section 4 provides some comparisons of related work in terms of the methodology of constructing the BoW feature, including the detection method, the clustering algorithm, the number of visual words, and so forth and the experimental set up including the datasets used, the number of object or scene categories, and so forth. Finally, Section 5 concludes the paper.

2. Bag-of-Words Representation

The bag-of-words (BoW) methodology was first proposed in the text retrieval domain problem for text document analysis, and it was further adapted for computer vision applications [24]. For image analysis, a visual analogue of a word is used in the BoW model, which is based on the vector quantization process by clustering low-level visual features of local regions or points, such as color, texture, and so forth.

To extract the BoW feature from images involves the following steps: (i) automatically detect regions/points of interest, (ii) compute local descriptors over those regions/points, (iii) quantize the descriptors into words to form the visual vocabulary, and (iv) find the occurrences in the image of each specific word in the vocabulary for constructing the BoW feature (or a histogram of word frequencies) [24]. Figure 1 describes these four steps to extract the BoW feature from images.

The BoW model can be defined as follows. Given a training dataset D containing n images represented by $D = d_1, d_2, \ldots$, and d_n, where d is the extracted visual features, a specific unsupervised learning algorithm, such as k-means, is used to group D based on a fixed number of visual words W (or categories) represented by $W = w_1, w_2, \ldots$, and w_v, where V is the cluster number. Then, we can summarize the data in a $V \times N$ cooccurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ denotes how often the word w_i occurred in an image d_i.

2.1. Interest Point Detection. The first step of the BoW methodology is to detect local interest regions or points. For feature extraction of interest points (or keypoints), they are computed at predefined locations and scales. Several well-known region detectors that have been described in the literature are discussed below [25, 26].

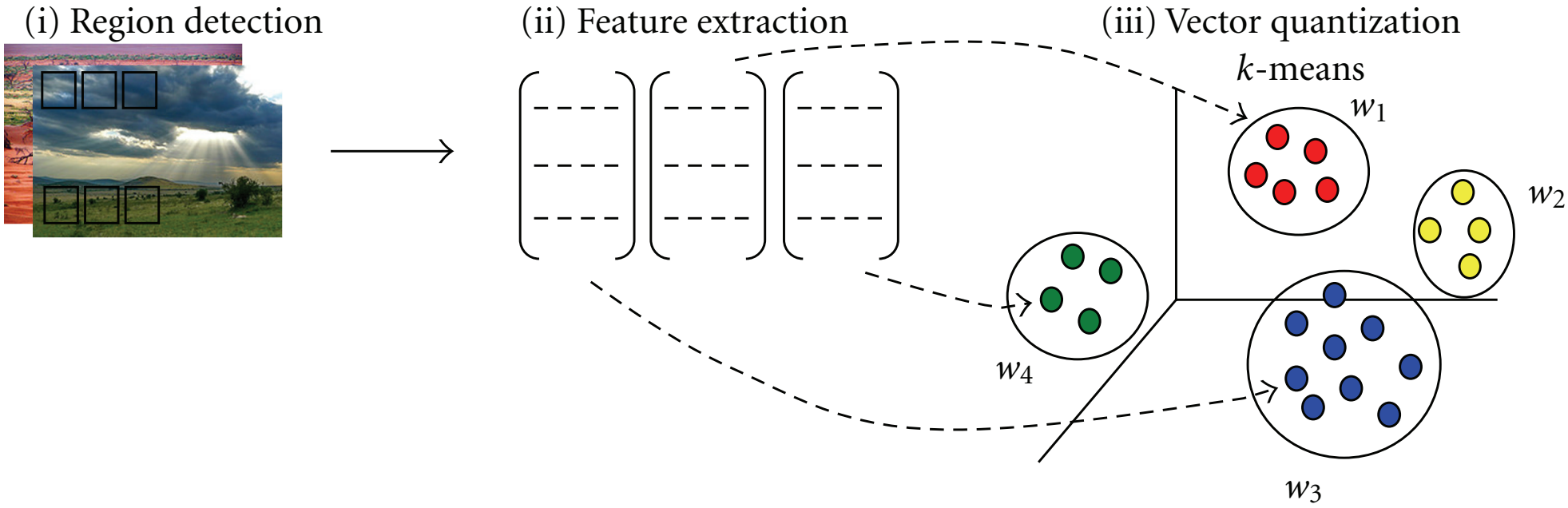
(i) Harris-Laplace regions are detected by the scale-adapted Harris function and selected in scale-space by the Laplacian-of-Gaussian operator. Harris-Laplace detects corner-like structures.

(ii) DoG regions are localized at local scale-space maxima of the difference-of-Gaussian. This detector is suitable for finding blob-like structures. In addition, the DoG point detector has previously been shown to perform well, and it is also faster and more compact (less feature points per image) than other detectors.

(iii) Hessian-Laplace regions are localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian.

(iv) Salient regions are detected in scale-space at local maxima of the entropy. The entropy of pixel intensity histograms is measured for circular regions of various sizes at each image position.

(v) Maximally stable extremal regions (MSERs) are components of connected pixels in a thresholded image.


(i) Region detection
(ii) Feature extraction
(iii) Vector quantization
k-means
w1
w2
w3
w4
(iv) Bag-of-words

as with the SIFT descriptor, the final step of extracting the BoW feature from images is based on vector quantization. In general, the k-means clustering algorithm is used for this task, and the number of visual words generated is based on the number of clusters (i.e., k). Jiang et al. [17] conducted a comprehensive study on the representation choices of BoW, including vocabulary size, weighting scheme, such as binary, term frequency (TF) and term frequency-inverse document frequency (TF-IDF), stop word removal, feature selection, and so forth for video and image annotation.

To generate visual words, many studies focus on capturing spatial information in order to improve the limitations of the conventional BoW model, such as Yang et al. [37], Zhang et al. [38], Chen et al. [39], S. Kim and D. Kim [40], Lu and Ip [41], Lu and Ip [42], Uijlings et al. [43], Cao and Fei-Fei [44], Philbin et al. [45], Wu et al. [46], Agarwal and Triggs [47], Lazebnik et al. [48], Marszałek and Schmid [49], and Monay et al. [50], in which spatial pyramid matching introduced by Lazebnik et al. [48] has been widely compared as one of the baselines.

However, Van de Sande et al. [51] have shown that the severe drawback of the bag-of-words model is its high computational cost in the quantization step. In other words, the most expensive part in a state-of-the-art setup of the bag-of-words model is the vector quantization step, that is, finding the closest cluster for each data point in the k-means algorithm.

Uijlings et al. [33] compare k-means and random forests for the word assignment task in terms of computational efficiency. By using different descriptors with different grid sizes, random forests are significantly faster than k-means. In addition, using random forests to generate BoW can provide a slightly better Mean Average Precision (MAP) than k-means does. They also recommend two BoW pipelines when the focuses are on accuracy and speed, respectively.

In their seminal work, Philbin et al. [45], the approximate k-means, hierarchical k-means, and (exact) k-means are compared in terms of the precision performance and computational cost, where approximate k-means works best. (See Section 4.3 for further discussion).

Chum et al. [52] observe that feature detection and quantization are noisy processes and this can result in variation in the particular visual words that appear in different images of the same object, leading to missed results.

2.4. Learning Models. After the BoW feature is extracted from images, it is entered into a classifier for training or testing. Besides constructing the discriminative models as classifiers for image annotation, some Bayesian text models by Latent Semantic Analysis [53], such as probabilistic Latent Semantic Analysis (pLSA) [54] and Latent Dirichlet Analysis (LDA) [55] can be adapted to model object and scene categories.

2.4.1. Discriminative Models. The construction of discriminative models for image annotation is based on the supervised machine learning principle for pattern recognition. Supervised learning can be thought as learning by examples or learning with a teacher [56]. The teacher has knowledge of the environment which is represented by a set of input-output examples. In order to classify unknown patterns, a certain number of training samples are available for each class, and they are used to train the classifier [57].

The learning task is to compute a classifier or model $\hat{f}$ that approximates the mapping between the input-output examples and correctly labels the training set with some level of accuracy. This can be called the *training* or *model generation* stage. After the model $\hat{f}$ is generated or trained, it is able to classify an unknown instance, into one of the learned class labels in the training set. More specifically, the classifier calculates the similarity of all trained classes and assigns the unlabeled instance to the class with the highest similarity measure. More specifically, the most widely developed classifier is based on support vector machines (SVM) [58].

2.4.2. Generative Models. In text analysis, pLSA and LDA are used to discover topics in a document using the BoW document representation. For image annotation, documents and discovered topics are thought of as images and object categories, respectively. Therefore, an image containing instances of several objects is modeled as a mixture of topics. This topic distribution over the images is used to classify an image as belonging to a certain scene. For example, if an image contains "water with waves", "sky with clouds", and "sand", it will be classified into the "coast" scene class [24].

Following the previous definition of BoW, in pLSA there is a latent variable model for cooccurrence data which associates an unobserved class variable $z \in Z = z_1, \ldots, z_Z$ with each observation. A joint probability model $P(w, d)$ over $V \times N$ is defined by the mixture:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d), \tag{1}$$

where $P(w \mid z)$ are the topic specific distributions and each image is modeled as a mixture of topics, $P(z \mid d)$.

On the other hand, LDA treats the multinomial weights $P(z \mid d)$ over topics as latent random variables. In particular, the pLSA model is extended by sampling those weights from a Dirichlet distribution. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, which can reduce the overfitting problem.

The goal of LDA is to maximize the following likelihood:

$$P(w \mid \phi, \alpha, \beta) = \int \sum_{z} P(w \mid z, \phi) P(z \mid \theta) P(\theta \mid \alpha) P(\phi \mid \beta) d\theta, \tag{2}$$

where θ and ϕ are multinomial parameters over the topics and words, respectively, and $P(\theta \mid \alpha)$ and $P(\phi \mid \beta)$ are Dirichlet distributions parameterized by the hyperparameters α and β.

Bosch et al. [24] compare BoW + pLSA with different semantic modeling approaches, such as the traditional global based feature representation, block-based feature representation [59] with the k-nearest neighbor classifier. They show

that BoW + pLSA performs best. Specifically, the HIS histogram + cooccurrence matrices + edge direction histogram are used as the image descriptors.

However, it is interesting that Lu and Ip [41] and Quelhas et al. [60] show that pLSA does not perform better than BoW + SVM over the Corel dataset, where the former uses blocked based HSV and Gabor texture features and the latter uses keypoint based SIFT features.

3. Extensions of BoW

This section reviews the literature regarding using BoW for some related problems. They are divided into five categories, namely, feature representation, vector quantization, visual vocabulary construction, image segmentation, and others.

3.1. Feature Representation. Since the annotation accuracy is heavily dependent on feature representation, using different region/point descriptors and/or the BoW feature representation will provide different levels of discriminative power for annotation. For example, Mikolajczyk and Schmid [34] compare 10 different local descriptors for object recognition. Jiang et al. [17] examine the classification accuracy of the BoW features using different numbers of visual words and different weighting schemes.

Due to the drawbacks that vector quantization may reduce the discriminative power of images and the BoW methodology ignores geometric relationships among visual words, Zhong et al. [61] present a novel scheme where SIFT features are bundled into local groups. These bundled features are repeatable and are much more discriminative than an individual SIFT feature. In other words, a bundled feature provides a flexible representation that allows us to partially match two groups of SIFT features.

On the other hand, since the image feature generally carries mixed information of the entire image which may contain multiple objects and background, the annotation accuracy can be degraded by such noisy (or diluted) feature representations. Chen et al. [62] propose a novel feature representation, pseudo-objects. It is based on a subset of proximate feature points with its own feature vector to represent a local area to approximate candidate objects in images.

Gehler and Nowozin [63] focus on feature combination, which is to combine multiple complementary features based on different aspects such as shape, color, or texture. They study several models that aim at learning the correct weighting of different features from training data. They provide insight into when combination methods can be expected to work and how the benefit of complementary features can be exploited most efficiently.

Qin and Yung [64] use localized maximum-margin learning to fuse different types of features during the BoW modeling. Particularly, the region of interest is described by a linear combination of the dominant feature and other features extracted from each patch at different scales, respectively. Then, dominant feature clustering is performed to create contextual visual words, and each image in the training set is evaluated against the codebook using the localized maximum-margin learning method to fuse other features, in order to select a list of contextual visual words that best represents the patches of the image.

As there is a relation between the composition of a photograph and its subject, similar subjects are typically photographed in a similar style. Van Gemert [65] exploits the assumption that images within a category share a similar style, such as colorfulness, lighting, depth of field, viewpoints and saliency. They use the photographic style for category-level image classification. In particular, where the spatial pyramid groups features spatially [48], they focus on more general feature grouping, including these photographic style attributes.

In Rasiwasia and Vasconcelos [66], they introduce an intermediate space, based on a low dimensional semantic "theme" image representation, which is learned with weak supervision from casual image annotations. Each theme induces a probability density on the space of low-level features, and images are represented as vectors of posterior theme probabilities.

3.2. Vector Quantization. In order to reduce the quantization noise, Jégou et al. [67] construct short codes using quantization. The goal is to estimate distances using vector-to-centroid distances, that is, the query vector is not quantized, codes are assigned to the database vectors only. In other words, the feature space is decomposed into a Cartesian product of low-dimensional subspaces, and then each subspace is quantized separately. In particular, a vector is represented by a short code composed of its subspace quantization indices.

As abrupt quantization into discrete bins does cause some aliasing, Agarwal and Triggs [47] focus on soft vector quantization, that is, softly voting into the cluster centers that lie close to the patch, for example, with Gaussian weights. They show that diagonal-covariance Gaussian mixtures fitted using expectation-maximization performs better than hard vector quantization.

Similarly, Fernando et al. [68] propose a supervised learning algorithm based on a Gaussian mixture model, which not only generalizes the k-means by allowing "soft assignments", but also exploits supervised information to improve the discriminative power of the clusters. In their approach, an EM-based approach is used to optimize a convex combination of two criteria, in which the first one is unsupervised and based on the likelihood of the training data, and the second is supervised and takes into account the purity of the clusters.

On the other hand, Wu et al. [69] propose a Semantics-Preserving Bag-of-Words (SPBoW) model, which considers the distance between the semantically identical features as a measurement of the semantic gap and tries to learn a codebook by minimizing this semantic gap. That is, the codebook generation task is formulated as a distance metric learning problem. In addition, one visual feature can be assigned to multiple visual words in different object categories.

In de Campos et al. [70], images are modeled as orderless sets of weighted visual features where each visual feature is associated with a weight factor that may inform re its

relevance. In this approach, visual saliency maps are used to determine the relevance weight of a feature.

Zheng et al. [71] argue that for the BoW model used in information retrieval and document categorization, the textual word possesses semantics itself and the documents are well-structured data regulated by grammar, linguistic, and lexicon rules. However, there appears to be no well-defined rules in the visual word composition of images. For instance, the objects of the same class might have arbitrarily different shapes and visual appearances, while objects of different classes might share similar local appearances. To this end, a higher-level visual representation, visual synset for object recognition is presented. First, an intermediate visual descriptor, delta visual phrase, is constructed from a frequently co-occurring visual word-set with similar spatial context. Second, the delta visual phrases are clustered into a visual synset based their probabilistic "semantics", that is, class probability distribution.

Besides reducing the vector quantization noise, another severe drawback of the BoW model is its high computational cost. To address this problem, Moosmann et al. [72] introduce extremely randomized clustering forests based on ensembles of randomly created clustering trees and show that more accurate results can be obtained as well as much faster training and testing.

Recently, Van de Sande et al. [51] proposed two algorithms to combine GPU hardware and a parallel programming model to accelerate the quantization and classification components of the visual categorization architecture.

On the other hand, Hare et al. [73] show the intensity inversion characteristics of the SIFT descriptor and local interest region detectors can be exploited to decrease the time it takes to create vocabularies of visual terms. In particular, they show that clustering inverted and noninverted (or minimum and maximum) features separate results in the same retrieval performance when compared to the clustering of all the features as a single set (with the same overall vocabulary size).

3.3. Visual Vocabulary Construction. Since related studies, such as Jegou et al. [74], Marszałek and Schmid [49], Sivic and Zisserman [14], and Winn et al. [75], have shown that the commonly generated visual words are still not as expressive as text words, in Zhang et al. [76], images are represented as visual documents composed of repeatable and distinctive visual elements, which are comparable to text words. They propose descriptive visual words (DVWs) and descriptive visual phrases (DVPs) as the visual correspondences to text words and phrases, where visual phrases refer to the frequently co-occurring visual word pairs.

Gavves et al. [77] focus on identifying pairs of independent, distant words—the visual synonyms—that are likely to host image patches of similar visual reality. Specifically, landmark images are considered, where the image geometry guides the detection of synonym pairs. Image geometry is used to find those image features that lie in a nearly identical physical location, yet are assigned to different words of the visual vocabulary.

On the other hand, López-Sastre et al. [78] present a novel method for constructing a visual vocabulary that takes into account the class labels of images. It consists of two stages: Cluster Precision Maximisation (CPM) and Adaptive Refinement. In the first stage, a Reciprocal Nearest Neighbours (RNN) clustering algorithm is guided towards class representative visual words by maximizing a new cluster precision criterion. Next, an adaptive threshold refinement scheme is proposed with the aim of increasing vocabulary compactness, while at the same time improving the recognition rate and further increasing the representativeness of the visual words for category-level object recognition. In other words, this is a correlation clustering based approach, which works as a kind of metaclustering and optimizes the cut-off threshold for each cluster separately.

Constructing visual codebook ensembles is another approach to improve image annotation accuracy. In Luo et al. [18], three methods for constructing visual codebook ensembles are presented. The first one is based on diverse individual visual codebooks by randomly choosing interesting points. The second one uses a random subtraining image dataset with random interesting points. The third one directly utilizes different patch information for constructing an ensemble with high diversity. Consequently, different types of image presentations are obtained. Then, a classification ensemble is learned by the different expression datasets from the same training set.

Bae and Juang [79] apply the idea of linguistic parsing to generate the BoW feature for image annotation. That is, images are represented by a number of variable-size patches by a multidimensional incremental parsing algorithm. Then, the occurrence pattern of these parsed visual patches is fed into the LSA framework.

Since one major challenge in object categorization is to find class models that are "invariant" enough to incorporate naturally-occurring intraclass variations and yet "discriminative" enough to distinguish between different classes, Winn et al. [75] proposed a supervised learning algorithm, which automatically finds such models. In particular, it classifies a region according to the proportions of different visual words. The specific visual words and the typical proportions in each object are learned from a segmented training set.

Kesorn and Poslad [80] propose a framework to enhance the visual word quality. First of all, visual words from representative keypoints are constructed by reducing similar keypoints. Second, domain specific noninformative visual words are detected, which are useless for representing the content of visual data but which can degrade the categorization capability. A noninformative visual word is defined as having a high document frequency and a small statistical association with all the concepts in the image collection. Third, the vector space model of visual words is restructured with respect to a structural ontology model in order to solve visual synonym and polysemy problems.

Tirlly et al. [81] present a new image representation called visual sentences that allows us to "read" visual words in a certain order, as in the case of text. Particularly, simple spatial relations between visual words are considered. In addition, pLSA is used to eliminate the noisiest visual words.

3.4. Image Segmentation. Effective image segmentation can be an important factor affecting the BoW feature generation. Uijlings et al. [43] study the role of context in the BoW approach. They observe that using the precise localization of object patches based on image segmentation is likely to yield a better performance than the dense sampling strategy, which sample patches of $8 * 8$ pixels at every 4th pixel.

Besides point detection, an image can be segmented into several or a fixed number of regions or blocks. However, very few compared the effect of image segmentation on generating the BoW feature. In Cheng and Wang [82], 20–50 regions per image are segmented, and each region is represented by a HSV histogram and cooccurrence texture features. By using contextual Bayesian networks to model spatial relationship between local regions and integrating multiattributes to infer high-level semantics of an image, this approach performs better and is comparable with a number of works using SIFT descriptors and pLSA for image annotation.

Similarly, Wu et al. [46] extract a texture histogram from the $8 * 8$ blocks/patches per image based on their proposed visual language modeling method utilizing the spatial correlation of visual words. This representation is compared with the BoW model including pLSA and LDA using the SIFT descriptor. They show that neither image segmentation nor interest point detection is used in the visual language modeling method, which makes the method not only very efficient, but also very effective over the Caltech 7 dataset.

In addition to using the BoW feature for image annotation, Larlus et al. [83] combine BoW with random fields and some generative models, such as a Dirichlet processes for more effective object segmentation.

3.5. Others

3.5.1. BoW Applications. Although the BoW model has been extensively studied for general object and scene categorization, it has also been considered in some domain specific applications, such as human action recognition [84], facial expression recognition [85], medical images [86], robot, sport image analysis [80], 3D image retrieval and classification [87, 88], image quality assessment [89], and so forth.

3.5.2. Describing Objects/Scenes for Recognition. Farhadi et al. [90] propose shifting the goal of recognition from naming to describing. That is, they focus on describing objects by their attributes, which is not only to name familiar objects, but also to report unusual aspects of a familiar object, such as "spotty dog", not just "dog", and to say something about unfamiliar objects, such as "hairy and four-legged", not just "unknown".

On the other hand, Sudderth et al. [91] develop hierarchical, probabilistic models for objects, the parts composing them, and the visual scenes surrounding them. These models share information between object categories in three distinct ways. First, parts define distributions over a common low-level feature vocabulary. Second, objects are defined using a common set of parts. Finally, object appearance information is shared between the many scenes in which that object is found.

3.5.3. Query Expansion. Chum et al. [52] adopt the BoW architecture with spatial information for query expansion, which has proven successful in achieving high precision at low recall. On the other hand, Philbin et al. [92] quantize a keypoint to the k-nearest visual words as a form of query expansion.

3.5.4. Similarity Measure. Based on the BoW feature representation, Jegou et al. [74] introduce a contextual dissimilarity measure (CDM), which is iteratively obtained by regularizing the average distance of each point to its neighborhood. In addition, CDM is learned in an unsupervised manner, which does not need to learn the distance measure from a set of training images.

3.5.5. Large Scale Image Databases. Since the aim of image annotation is to support very large scale keyword-based image search, such as web image retrieval, it is very critical to assess existing approaches over some large scale dataset(s). Chum et al. [52], Hörster and Lienhart [21], and Lienhart and Slaney [93] used datasets composed of 100000 to 250000 images belonging to 12 categories, which were downloaded from Flickr.

Moreover, Philbin et al. [45] use over 1000000 images from Flickr for experiments and Zhang et al. [94] use about 370000 images collected from Google belonging to 1506 object or scene categories.

On the other hand, Torralba and Efros [95] study some bias issues of object recognition datasets. They provide some suggestions for creating a new and high quality dataset to minimize the selection bias, capture bias, and negative set bias. Furthermore, they claim that in the state of today's datasets there are virtually no studies demonstrating cross-dataset generalization, for example, training on ImageNet, while testing on PASCAL VOC. This could be considered as an additional experimental setup for future works.

3.5.6. Integration of Feature Selection and/or (Spatial) Feature Extraction. Although modeling the spatial relationship between visual words can improve the recognition performance, the spatial features are expensive to compute. Liu et al. [96] propose a method that simultaneously performs feature selection and (spatial) feature extraction based on higher-order spatial features for speed and storage improvements.

For the dimensionality reduction purpose, Elfiky et al. [97] present a novel framework for obtaining a compact pyramid representation. In particular, the divisive information theoretic feature clustering (DITC) algorithm is used to create a compact pyramid representation.

Bosch et al. [98] investigate whether dimensionality reduction using a latent generative model is beneficial for the task of weakly supervised scene classification. In their approach, latent "topics" using pLSA are first of all discovered, and a generative model is then applied to the BoW representation for each image.

In contrast to reducing the dimensionality of the feature representation, selecting more discriminative features (e.g., SIFT descriptors) from a given set of training images has

been considered. Shang and Xiao [99] introduce a pairwise image matching scheme to select the discriminative features. Specifically, the feature weights are updated by the labeled information from the training set. As a result, the selected features corresponding to the foreground content of the images can highlight the information category of the images.

3.5.7. Integration of Segmentation, Classification, and/or Retrieval. Simultaneously learning object/scene category models and performing segmentation on the detected objects were studied in Cao and Fei-Fei [44]. They propose a spatially coherent latent topic model (Spatial-LTM), which represents an image containing objects in a hierarchical way by oversegmented image regions of homogeneous appearances and the salient image patches within the regions. It can provide a unified representation for spatially coherent BoW topic models and can simultaneously segment and classify objects.

On the other hand, Tong et al. [100] propose a statistical framework for large-scale near duplicate image retrieval which unifies the step of generating a BoW representation and the step of image retrieval. In this approach, each image is represented by a kernel density function, and the similarity between the query image and a database image is then estimated as the query likelihood.

Shotton et al. [101] utilize semantic texton forests, which are ensembles of decision trees that act directly on image pixels, where the nodes in the trees provide an implicit hierarchical clustering into semantic textons and an explicit local classification estimate. In addition, the bag of semantic textons combines a histogram of semantic textons over an image region with a region prior category distribution, and the bag of semantic textons is computed over the whole image for categorization and over local rectangular regions for segmentation.

3.5.8. Discriminative Learning Models. Romberg et al. [102] extend the standard single-layer pLSA to multiple layers, where the multiple layers handle multiple modalities and a hierarchy of abstractions. In particular, the multilayer multimodal pLSA (mm-pLSA) model is based on a two leaf-pLSAs and a single top-level pLSA node merging the two leaf-pLSAs. In addition, SIFT features and image annotations (tags) as well as the combination of SIFT and HOG features are considered as two pairs of different modalities.

3.5.9. Novel Category Discovery. In their study, Lee and Grauman [103] discover new categories by knowing some categories. That is, previously learned categories are used to discover their familiarity in unsegmented, unlabeled images. In their approach, two variants of a novel object-graph descriptor to encode 2D and 3D spatial layout of object-level cooccurrence patterns relative to an unfamiliar region, and they are used to model the interaction between an image's known and unknown objects for detecting new visual categories.

3.5.10. Interest Point Detection. Since interest point detection is an important step for extracting the BoW feature, Stottinger et al. [104] propose color interest points for sparse image representation. Particularly, light-invariant interest points are introduced to reduce the sensitivity to varying imaging conditions. Color statistics based on occurrence probability lead to color boosted points, which are obtained through saliency-based feature selection.

4. Comparisons of Related Work

This section compares related work in terms of the ways the BoW feature and experimental setup are structured. These comparisons allow us to figure out the most suitable interest point detector(s), clustering algorithm(s), and so forth used to extract the BoW feature from images. In addition, we are able to realize the most widely used dataset(s) and experimental settings for image annotation by BoW.

4.1. Methodology of BoW Feature Generation. Table 1 compares related work for the methodology of extracting the BoW feature. Note that we leave a blank if the information in our comparisons is not clearly described in these related works.

From Table 1 we can observe that the most widely used interest point detector for generating the BoW feature is DoG, and the second and third most popular detectors are Harris-Laplace and Hessian-Laplace, respectively. Besides extracting sparse BoW features, many related studies have focused on dense BoW features.

On the other hand, several studies used some region segmentation algorithms, such as NCuts [116] and Meanshift [117], to segment an image into several regions to represent keypoints.

For the local feature descriptor to describe interest points, most studies used a 128 dimensional SIFT feature, in which some considered using PCA to reduce the dimensionality of SIFT, but some "fuse" the color feature and SIFT resulting in longer dimensional features than SIFT. Except for extracting SIFT related features, some studies considered conventional color and texture features to represent local regions or points.

About vector quantization, we can see that k-means is the most widely used clustering algorithm to generate the codebook or visual vocabularies. However, in order to solve the limitations of k-means, for example, clustering accuracy and computational cost, some studies used hierarchical k-means, approximate k-means, accelerated k-means, and so forth.

For the number of visual words, related works have considered various amounts of clusters during vector quantization. This may be because the datasets used in these works are different. In Jiang et al. [17], different numbers of visual words were studied, and their results show that 1000 is a reasonable choice. Some related studies also used similar numbers of visual words to generate their BoW features.

On the other hand, the most and second most widely used weighting schemes are TF and TF-IDF. This is consistent with Jiang et al. [17], who concluded that these two weighting schemes perform better than the other weighting schemes.

Finally, SVM is no doubt the most popular classification technique as the learning model for image annotation. In particular, one of the most widely used kernel functions for

Table 1: Comparisons of interest point detection, visual words generation, and learning models.

Work	Region/point detection	Local descriptor	Clustering algorithm	No. of visual words	Weighting scheme	Learning model
			2012			
de Campos et al. [70]	DoG	SIFT				Logistic regression
Elfiky et al. [97]	Harris-Laplace	SIFT/HSV color + SIFT	k-means			SVM
Fernando et al. [68]	Harris-Laplace	PCA-SIFT/SIFT/SURF[1]	k-means	2000		SVM
Gavves et al. [77]		SIFT/SURF		200000		
Kesorn and Poslad [80]	DoG	SIFT	SLAC[2]		Binary/TF/ TF-IDF	Naïve bayes/ SVM-linear/ SVM-RBF
Lee and Grauman [103]	NCuts[3]	Texton histogram	k-means	400		SVM
Qin and Yung [64]		Color SIFT	k-means			SVM-linear/ SVM-poly/ SVM-RBF
Romberg et al. [102]		SIFT	k-means			mm-pLSA[4]
Shang and Xiao [99]		SIFT	k-means	1000		SVM
Stottinger et al. [104]	Harris-Laplace	RGB Harris with Laplacian scale selection	k-means	4000		SVM
Tong et al. [100]	Harris-Laplace	SIFT	AKM[5]			
			2011			
Hare et al. [73]	DoG/MSER	SIFT	AKM	1000–100000	IDF	
López-Sastre et al. [78]	Hessian-Laplace	SIFT	CPM and Adaptive Refinement	3818		SVM
Luo et al. [18]	DoG	SIFT	k-means	500	TF	SVM
Van Gemert [65]	Harris and Hessian-Laplace	SIFT	k-means	2000		
Yang et al. [37]		SIFT	k-means	1000		SVM
Zhang et al. [76]	DoG	SIFT	HKM[6]	32357	TF-IDF	
Zhang et al. [38]	DoG	SIFT	HKM	32400	TF-IDF	
			2010			
Bae and Juang [79]	Dense sampling			171329		
Chen et al. [62]	Hessian-Laplace	SIFT	GMM-BIC[7]	3500	TF	
Cheng and Wang [82]	Mean-shift[8]	HSV color histogram and co-occurrence matrix				SVM
Ding et al. [105]	DoG	PCA-SIFT	k-means	2000		SVM
Jégou et al. [22]	Hessian-Laplace	SIFT	k-means	200000	TF-IDF	
Jiang et al. [17]	DoG	SIFT	k-means	500–10000	Binary/TF/ TF-IDF/soft-weighting	SVM
Li and Godil [87]	DoG	SIFT	k-means	500/700/800	TF	pLSA
Qin and Yung [106]		PCA-SIFT	Accelerated k-means	32/128/2048/ 4096		SVM
Tirilly et al. [107]	Hessian-Laplace	SIFT	HKM	6556 to 117151		
Uijlings et al. [33]		PCA-SIFT	k-means/ random forest	4096		SVM
Wu et al. [69]		SIFT	k-means	2500–4500		Naïve Bayes/ SVM

Table 1: Continued.

Work	Region/point detection	Local descriptor	Clustering algorithm	No. of visual words	Weighting scheme	Learning model
			2009			
Chen et al. [39]	DoG	SIFT	k-means	1000	Spatial weighting	
Lu and Ip [41]	Dense sampling	HSV color + Gabor txture	k-means	100/200		SVM
Lu and Ip [42]	Dense sampling	HSV color + Gabor txture	k-means	100/200		LLP[9]/GLP[10]/ SVM
S. Kim and D. Kim [40]	Dense sampling	SIFT/SURF	k-means	500/1500/3000	TF	pLSA/SVM
Uijlings et al. [43]	Dense sampling	SIFT	k-means	4096		SVM
Xiang et al. [108]	NCuts	36 region features[11]				MRFA[12]
Zhang et al. [94]		SIFT	HKM	32357	TF-IDF	
			2008			
Bosch et al. [98]	Harris-Laplace	Color SIFT	k-means	1500		k-NN/SVM
Liu et al. [96]	Harris-Laplace	SIFT	k-means	1000		SVM-linear
Marszałek and Schmid [109]	Harris-Laplace	SIFT	k-means	8000		SVM
Rasiwasia and Vasconcelos [66]		DCT[13] coefficients				Hierarchical Dirichlet models/SVM
Tirilly et al. [81]		SIFT	HKM	6556/61687	TF-IDF	SVM
Van de Sande et al. [110]	Harris-Laplace	Color SIFT	k-means	4000		SVM
Zheng et al. [71]	DoG + Hessian-Laplace	SIFT + Spin[14]	k-means	1010		SVM
			2007			
Bosch et al. [24]	Dense sampling	HSV color + co-occurrence + edge	k-means	700		pLSA
Chum et al. [52]	Hessian-Laplace	SIFT	k-means		TF-IDF	
Gökalp and Aksoy [28]	Dense sampling	HSV color	k-means			Bayesian classifier
Hörster and Lienhart [21]	DoG/dense sampling	Color SIFT	k-means			LDA
Jegou et al. [74]		SIFT	k-means	30000		
Li and Fei-Fei [111]	Dense sampling	SIFT	k-means	300	TF	LDA
Lienhart and Slaney [93]		SIFT	k-means		TF	pLSA
Philbin et al. [45]	Hessian-Laplace	SIFT	AKM	1 M		
Quelhas et al. [13]	DoG	SIFT	k-means	1000		SVM/pLSA
Wu et al. [46]	Dense sampling	Texture histogram				Unigram/ bigram/trigram models
Junsong et al. [112]	DoG	PCA-SIFT	k-means	160/500		
			2006			
Agarwal and Triggs [47]	Dense sampling	SIFT	EM[15]			LDA/SVM
Bosch et al. [29]	Dense sampling	Color SIFT	k-means	1500		k-NN/pLSA
Lazebnik et al. [48]	Dense sampling	SIFT	k-means	200/400		SVM
Marszałek and Schmid [49]	Harris-Laplace	SIFT	k-means	1000	TF	SVM

Table 1: Continued.

Work	Region/point detection	Local descriptor	Clustering algorithm	No. of visual words	Weighting scheme	Learning model
Monay et al. [50]	DoG	SIFT	k-means	1000	TF	pLSA
Moosmann et al. [72]	Dense sampling/DoG	HSV color + wavelet/ SIFT	Extremely randomized trees			SVM
Perronnin et al. [113]	DoG	PCA-SIFT		1024		SVM-linear

[1] Speeded up robust features [114].
[2] Search ant and labor ant clustering algorithm [115].
[3] Normalized cuts [116].
[4] Multilayer modality pLSA.
[5] Approximate k-means.
[6] Hierarchical k-means.
[7] Gaussian mixture model with Bayesian information criterion.
[8] Mean shift region segmentation algorithm [117].
[9] Local label propagation on the k-NN graph.
[10] Global label propagation on the complete graph.
[11] Region color and standard deviation, region average orientation energy (12 filters), region size, location, convexity, first moment, and ratio of region area to boundary length squared [118].
[12] Multiple Markov random fields.
[13] Discrete cosine transform.
[14] A rotation-invariant two-dimensional histogram of intensities within an image region [71].
[15] Expectation maximization.

constructing the SVM classifier is the Gaussian radial basis function. However, some other SVM classifiers, such as linear SVM and SVM with a polynomial kernel have also been considered in the literature.

4.2. Experimental Design. Table 2 compares related work for the experimental design. That is, the chosen dataset(s) and baseline(s) are examined.

According to Table 2, most studies considered more than one single dataset for their experiments, and many of them contained object and scene categories. This is very important for image annotation that the annotated keywords should be broadened for users to perform keyword-based queries for image retrieval.

Specifically, the PASCAL, Caltech, and Corel datasets are the three most widely used benchmarks for image classification. However, the datasets used in most studies usually contain a small number of categories and images, except for the studies focusing on retrieval rather than classification. That is, similar based queries are used to retrieve relevant images instead of training a learning model to classify unknown images into one specific category.

For the chosen baselines, most studies compared BoW and/or spatial pyramid matching based BoW since their aims were to propose novel approaches to improve these two feature representations. Specifically, Lazebnik et al. [48] proposed spatial pyramid matching based BoW as the most popular baseline.

Besides improving the feature representation per se, some studies focused on improving the performance of LDA and/or pLSA discriminative learning models. Another popular baseline is that of Fei-Fei and Perona [31], who proposed a Bayesian hierarchical model to represent each region as part of a "theme."

4.3. Discussion. The above comparisons indicate several issues that were not examined in the literature. Since the local features can be represented using object-based regions by region segmentation [143, 144] or point-based regions by point detection (c.f. Section 2.1), regarding the BoW feature based on tokenizing, it is unknown which local feature is more appropriate for large scale image annotation (For large scale image annotation, this means that the number of annotated keywords is certainly large and their meanings are very broad, containing object and scene concepts.)

In addition, the local feature descriptor is the key component to the success of better image annotation; it is a fact that the number of visual words (i.e., clusters) is another factor affecting image annotation performance. Although Jiang et al. [17] conducted a comprehensive study of using various amounts of visual words, they only used one dataset, that is, TRECVID, containing 20 concepts. Therefore, one important issue is to provide the guidelines for determining the number of visual words over different kinds of image datasets having different image contents.

The learning techniques can be divided into generative and discriminative models, but there are very few studies which assess their annotation performance over different kinds of image datasets which is necessary in order to fully understand the value of these two kinds of learning models. On the other hand, a combination of generative and discriminative learning techniques [145] or hybrid models are considered for the image annotation task.

For the experimental setup, the target of most studies was not image retrieval. In other words, the performance evaluation was usually for small scale problems based on datasets containing a small number of categories, say 10. However, image retrieval users will not be satisfied with a system providing only 10 keyword-based queries to search relevant

Table 2: Comparisons of datasets used and annotation performance.

Work	Categories		Dataset	No. of categories	No. of images	Baseline
	Scene	Object				
2012						
de Campos et al. [70]		v	PASCAL′07/′08[16]	20	9292	
Elfiky et al. [97]	v	v	Sport event/15 scene/butterflies[17]/ PASCAL′07/′09	15/20	6000/21000/ 2000/160k/ 4194k	Spatial pyramid
Fernando et al. [68]		v	PASCAL′06/ Caltech 10[18]	10/10/11	5304/3044	BoW
Gavves et al. [77]	v		Oxford 5k[19]	11	5062	
Kesorn and Poslad [80]	v		Olympic organization website + Google images	8	16000	pLSA
Lee and Grauman [103]	v	v	MSRC-v0[20]/-v2/ PASCAL′08/ Corel/Gould′09	21/20/7/14	3457/591/1023/ 100/715	LDA
Qin and Yung [64]	v		SCENE-8/-15	8/15	2688/4485	BoW
Romberg et al. [102]	v	v	Flickr-10M	>300	10080251	pLSA
Shang and Xiao [99]		v	Caltech 256/ MSRC	20/20		BoW
Stottinger et al. [104]			PASCAL′07	20	9963	
Tong et al. [100]	v	v	Tattoo dataset /Oxford/Flickr		101745/5062/ 1002805	RS[21]/HKM/AKM
2011						
Hare et al. [73]	v	v	UK Bench/MIR Flickr-25000[22]			BoW
López-Sastre et al. [78]		v	Caltech 101	10	890	Mikolajczyk et al. [25]; Stark and Schiele [119]
Luo et al. [18]		v	Caltech 4/Graz-02[23]	5/2	400/200	Li and Perona [31]; Moosmann et al. [72]
Van Gemert [65]	v	v	Corel/PASCAL′09	20	2000/7054	BoW/spatial pyramid
Yang et al. [37]		v	PASCAL′08	20	8445	Divvala et al. [120]; Zhong et al. [109]
Zhang et al. [76]	v	v	Google images/ Caltech 101and256	15	376500	BoW
Zhang et al. [38]	v	v	ImageNet[24]	15 queries	1.5 million	Nister and Stewenius [121]; Zhong et al. [61]
2010						
Bae and Juang [79]	v		Corel	15	20000	LSA
Chen et al. [62]		v	Oxford buildings/ Flickr 1k	11 (55 queries)/7 (56 queries)	5062/11282	Sivic and Zisserman [14]; Philbin et al. [45]; Lazebnik et al. [48]
Cheng and Wang [82]	v		6-scene dataset	6	700	Vogel and Schiele [122]; Bosch et al. [98]; Quelhas et al. [13]; Boutell et al. [123]
Ding et al. [105]	v		TRECVID′06[25]	20	61901	Binary/TF/TF-IDF weighting
Jégou et al. [22]	v	v	Holidays[26]/Oxford 5k/U. of Kentucky object recognition[27]	500/11 (55 queries)	1491/5062/6376	BoW by HE[28]/

Table 2: Continued.

Work	Categories		Dataset	No. of categories	No. of images	Baseline
	Scene	Object				
Jiang et al. [17]		v	TRECVID′06	20	79484	
Li and Godli [87]	v	v	Corel	50	5000	Duygulu et al. [118]; Jeon et al. [124]; Lavrenko et al. [125]; Monay and Gatica-Perez, 2007 [126]
Qin and Yung [106]	v	v		8/13/15	2688/3759/ 4485	Siagian and Itti [127, 128]; Bosch et al. [29]; Li and Perona [31]; Quelhas et al. [60]; Lazebnik et al. [48]
Tirilly et al. [107]	v	v	U. of Kentucky object recognition/Oxford 5k/ Caltech 6 & 101	300/55/200 queries	10200/5062/ 8197	TF-IDF weighting
Uijlings et al. [33]		v	PASCAL′07/ TRECVID′05/ Caltech 101	20/101/15	9963/12914/4485	BoW
Wu et al. [69]		v	LabelMe[29]/ PASCAL′06	495/10		BoW; Bar-Hillel et al. [129]; Davis et al. [130]; Goldberger et al. [131]; Perronnin et al. [113]; Weinberger et al. [132]
			2009			
Chen et al. [39]	v		LabelMe	8 (448 queries)	2689	Yang et al. [133]
Lu and Ip (a) [41]	v		LabelMe + Web images	3	1239	k-NN; LDA
Lu and Ip (b) [42]	v	v	Corel/histological images	10/5	1000	pLSA/SVM
S. Kim and D. Kim [40]	v	v	Corel/histological images	10/5	1000	LLP/GLP/SVM/pLSA
Uijlings et al. [43]		v	PASCAL′07	20	9963	BoW
Xiang et al. [108]		v	Corel/TRECVID′05	50/39	5000	Feng et al. [134]
Zhang et al. [94]	v	v	Google images/ Corel/Caltech 101 and 256	1506 queries/ 50/15	376500/500/ 2250	BoW
			2008			
Bosch et al. [98]	v		6-/8-/13-/15-scene	6/8/13/15	2688/702	BoW
Liu et al. [96]	v		PASCAL′06/Caltech 4/MSRC-v2	20/5/15		Savarese et al. [135]
Marszalek and Schmid [109]	v	v	Caltech 256	256		Lazebnik et al.[48]; Zhang et al. [35]
Rasiwasia and Vasconcelos [66]	v		15-natural scene/ Corel	15/50		Bosch et al. [29]; Lazebnik et al. [48]; Li and Perona [31]; Liu and Shah [136]
Tirilly et al. [81]		v	Caltech 6 and 101	6/101	5435/8697	SVM
Van de Sande et al. [110]	v	v	PASCAL′07/ TRECVID′05	20		
Zheng et al. [71]		v	Caltech 101/ PASCAL′05	12/4		BoW

Table 2: Continued.

Work	Categories		Dataset	No. of categories	No. of images	Baseline
	Scene	Object				
2007						
Bosch et al. [24]	v		Corel	6	700	Global and block-based features + k-NN; Vogel and Schiele [122]
Chum et al. [52]	v	v	Oxford + Flickr		104844	BoW
Gökalp and Aksoy [28]	v		LabelMe	7	1050	Bag of individual regions/ bag of region pairs
Hörster and Lienhart [21]	v		Flickr	12 (60 queries)	246348	BoW/color based BoW
Jegou et al. [74]	v	v	Object recognition benchmark[30]		10200	Object recognition benchmark
Li and Fei-Fei [111]	v		8 events	8	240	LDA
Lienhart and Slaney [93]	v		Flickr	12 (60 queries)	253460	LSA
Philbin et al. [45]	v	v	Oxford 5 k/Flickr 1 and 2	11/145 and 450 tags	5062/99782/ 1040801	BoW
Quelhas et al. [13]	v		Corel + Web images	5	6680/3805/ 9457/6364	BoW; Vailaya et al. [137]
Wu et al. [46]	v	v	Caltech 7/Corel	8/6	600	LDA/pLSA
Yuan et al. [112]		v	Caltech 101	2	558	BoW
2006						
Agarwal and Triggs [47]		v	Caltech 7 + Graz/ KTH-TIPS[31]/ Cal-IPNP[32]	4/10/2	1337/810/360	LDA
Bosch et al. [29]	v		6-/8-/13-scene	6/8/13	2688/702/1071	BoW
Lazebnik et al. [48]	v	v	15-scene/Caltech 101/Graz	15/101/2		Zhang et al. [138]; Opelt et al. [139]
Marszalek and Schmid [49]		v	PASCAL′05			Wang et al. [20]
Monay et al. [50]	v		Corel	4	6600	
Moosmann et al. [72]		v	Graz-02/ PASCAL′05	3/4		BoW
Perronnin et al. [113]	v	v	Corel	10	1000	BoW; Farquhar et al. [140]; Deselaers et al. [141]

[16] http://pascallin.ecs.soton.ac.uk/challenges/VOC/.
[17] http://www.comp.leeds.ac.uk/scs6jwks/dataset/leedsbutterfly/.
[18] http://www.vision.caltech.edu/Image_Datasets/Caltech101/.
[19] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/.
[20] http://www.cs.utexas.edu/~grauman/research/datasets.html.
[21] Random seed [142].
[22] http://press.liacs.nl/mirflickr/.
[23] http://lear.inrialpes.fr/people/marszalek/data/ig02/.
[24] http://www.image-net.org/.
[25] http://www-nlpir.nist.gov/projects/tv2006/tv2006.html.
[26] http://lear.inrialpes.fr/~jegou/data.php.
[27] http://vis.uky.edu/.
[28] Hamming embedding.
[29] http://labelme.csail.mit.edu/.
[30] http://vis.uky.edu/%7Estewe/ukbench/.
[31] http://www.nada.kth.se/cvap/databases/kth-tips/.
[32] http://crl.ucsd.edu/.

images. Some benchmarks are much more suitable for larger scale image annotation, such as the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) by ImageNet (http://www.image-net.org/challenges/LSVRC/2012/index) and Photo Annotation and Retrieval 2012 by ImageCLEF (http://www.imageclef.org/2012/photo). In particular, the ImageNet dataset contains over 10000 categories and 10000000 labeled images and ImageCLEF uses a subset of the MIRFLICKR collection (http://press.liacs.nl/mirflickr/), which contains 25 thousand images and 94 concepts.

However, it is also possible that some smaller scale datasets composed of a relatively small number of images and/or categories can be combined into larger datasets. For example, the combination of Caltech 256 and Corel could be regarded as a benchmark that is more close to the real world problem.

5. Conclusion

In this paper, a number of recent related works using BoW for image annotation are reviewed. We can observe that this topic has been extensively studied recently. For example, there are many issues for improving the discriminative power of BoW feature representations by such techniques as image segmentation, vector quantization, and visual vocabulary construction. In addition, there are other directions for integrating the BoW feature for different applications, such as face detection, medical image analysis, 3D image retrieval, and so forth.

From comparisons of related work, we can find the most widely used methodology to extract the BoW feature which can be regarded as a baseline for future research. That is, DoG is used as the kepoint detector and each keypoint is represented by the SIFT feature. The vector quantization step is based on the k-means clustering algorithm with 1000 visual words. However, the number of visual words (i.e., the k values) is dependent on the dataset used. Finally, the weighting scheme can be either TF or TF-IDF.

On the other hand, for the dataset issue in the experimental design, which can affect the contribution and final conclusion, the PASCAL, Caltech, and/or Corel datasets can be used as the initial study.

According to the comparative results, there are some future research directions. First, the local feature descriptor for vector quantization usually by point-based SIFT feature can be compared with other descriptors, such as a region-based feature or a combination of different features. Second, a guideline for determining the number of visual words over what kind of datasets should be provided. The third issue is to assess the performance of generative and discriminative learning models over different kinds of datasets, such as different dataset sizes and different image contents, for example, a single object per image and multiple objects per image. Finally, it is worth examining the scalability of BoW feature representation for large scale image annotation.

References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] M. L. Kherfi, D. Ziou, and A. Bernardi, "Image retrieval from the World Wide Web: issues, techniques, and systems," *ACM Computing Surveys*, vol. 36, no. 1, pp. 35–67, 2004.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, article 5, 2008.

[4] Y. Choi and E. M. Rasmussen, "Users' relevance criteria in image retrieval in American history," *Information Processing and Management*, vol. 38, no. 5, pp. 695–726, 2002.

[5] M. Markkula, M. Tico, B. Sepponen, K. Nirkkonen, and E. Sormunen, "A test collection for the evaluation of content-based image retrieval algorithms—a user and task-based approach," *Information Retrieval*, vol. 4, no. 3-4, pp. 275–293, 2001.

[6] A. Goodrum and A. Spink, "Image searching on the Excite Web search engine," *Information Processing and Management*, vol. 37, no. 2, pp. 295–311, 2001.

[7] C. F. Tsai and C. Hung, "Automatically annotating images with keywords: a review of image annotation systems," *Recent Patents on Computer Science*, vol. 1, no. 1, pp. 55–68, 2008.

[8] A. Hanbury, "A survey of methods for image annotation," *Journal of Visual Languages and Computing*, vol. 19, no. 5, pp. 617–627, 2008.

[9] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, pp. 346–362, 2011.

[10] A. Pinz, "Object categorization," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 4, pp. 255–353, 2006.

[11] C. F. Tsai, K. Mcgarry, and J. Tait, "CLAIRE: a modular support vector image indexing and classification system," *ACM Transactions on Information Systems*, vol. 24, no. 3, pp. 353–379, 2006.

[12] W.-C. Lin, M. Oakes, J. Tait, and C.-F. Tsai, "Improving image annotation via useful representative feature selection," *Cognitive Processing*, vol. 10, no. 3, pp. 233–242, 2009.

[13] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[14] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, pp. 1470–1477, October 2003.

[15] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 370–377, October 2005.

[16] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1816–1823, October 2005.

[17] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: a comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[18] H. L. Luo, H. Wei, and L. L. Lai, "Creating efficient visual codebook ensembles for object categorization," *IEEE Transactions on Systems, Man, and Cybernetics Part A*, vol. 41, no. 2, pp. 238–253, 2010.

[19] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic

concepts," in *Proceedings of the 12th ACM International Conference on Multimedia (MM '04)*, pp. 540–547, October 2004.

[20] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1597–1604, June 2006.

[21] E. Hörster and R. Lienhart, "Fusing local image descriptors for large-scale image retrieval," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[22] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] A. Bosch, X. Muñoz, and R. Martí, "Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, 2007.

[25] K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1792–1799, October 2005.

[26] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007.

[27] K. Mikolajczyk, T. Tuytelaars, C. Schmid et al., "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[28] D. Gökalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[29] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *European Conference on Computer Vision*, pp. 517–530, 2006.

[30] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 604–610, October 2005.

[31] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the 6th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 524–531, June 2005.

[32] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. 506–513, July 2004.

[33] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 665–681, 2010.

[34] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[35] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[36] Z. Li, Z. Shi, X. Liu, Z. Li, and Z. Shi, "Fusing semantic aspects for image annotation and retrieval," *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 798–805, 2010.

[37] L. Yang, N. Zheng, and J. Yang, "A unified context assessing model for object categorization," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 310–322, 2011.

[38] S. Zhang, Q. Tian, G. Hua et al., "Modeling spatial and semantic cues for large-scale near-duplicated image retrieval," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 403–414, 2011.

[39] X. Chen, X. Hu, and X. Shen, "Spatial weighting for bag-or-visual-words and its application in content-based image retrieval," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 867–874, 2009.

[40] S. Kim and D. Kim, "Scene classification using pLSA with visterm spatial location," in *Proceedings of the 1st ACM International Workshop on Interactive Multimedia for Consumer Electronics (IMCE '09)*, pp. 57–66, October 2009.

[41] Z. Lu and H. H. S. Ip, "Image categorization with spatial mismatch kernels," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 397–404, June 2009.

[42] Z. Lu and H. H. S. Ip, "Image categorization by learning with context and consistency," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 2719–2726, June 2009.

[43] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "What is the spatial extent of an object?" in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 770–777, June 2009.

[44] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[45] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[46] L. Wu, M. Li, Z. Li, W. Y. Ma, and N. Yu, "Visual language modeling for image classification," in *Proceedings of the 9th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval (MIR '07)*, pp. 115–124, September 2007.

[47] A. Agarwal and B. Triggs, "Hyperfeatures—multilevel local coding for visual recognition," in *Conference on Computer Vision*, pp. 30–43, 2006.

[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.

[49] M. Marszałek and C. Schmid, "Spatial weighting for bag-of-features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2118–2125, June 2006.

[50] F. Monay, P. Quelhas, J. M. Odobez, and D. Gatica-Perez, "Integrating co-occurrence and spatial contexts on patch-based scene segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 14–21, June 2006.

[51] K. E. A. Van De Sande, T. Gevers, and C. G. M. Snoek, "Empowering visual categorization with the GPU," *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 60–70, 2011.

[52] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: automatic query expansion with a generative

feature model for object retrieval," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[53] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal for the American Society for InFormation Science*, vol. 41, no. 6, pp. 391–407, 1990.

[54] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[55] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[56] T. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[57] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 1999.

[58] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[59] M. Summer and R. W. Picard, "Indoor-outdoor image classification," *IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp. 42–50, 1998.

[60] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 883–890, October 2005.

[61] W. Zhong, K. Qifa, M. Isard, and S. Jian, "Bundling features for large scale partial-duplicate web image search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 25–32, June 2009.

[62] K. T. Chen, K. H. Lin, Y. H. Kuo, Y. L. Wu, and W. H. Hsu, "Boosting image object retrieval and indexing by automatically discovered pseudo-objects," *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 815–825, 2010.

[63] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '09)*, pp. 221–228, 2009.

[64] J. Qin and N. H. Yung, "Feature fusion within local region using localized maximum-margin learning for scene categorization," *Pattern Recognition*, vol. 45, pp. 1671–1683, 2012.

[65] J. C. Van Gemert, "Exploiting photographic style for category-level image classification by generalizing the spatial pyramid," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11)*, pp. 1–8, April 2011.

[66] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[67] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[68] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Supervised learning of Gaussian mixture models for visual vocabulary generation," *Pattern Recognition*, vol. 45, pp. 897–907, 2011.

[69] L. Wu, S. C. H. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and applications," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1908–1920, 2010.

[70] T. de Campos, G. Csurka, and F. Perronnin, "Images as sets of locally weighted features," *Computer Vision and Image Understanding*, vol. 116, pp. 68–85, 2012.

[71] Y. T. Zheng, M. Zhao, S. Y. Neo, T. S. Chua, and Q. Tian, "Visual synset: towards a higher-level visual representation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[72] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *International Conference on Neural Information Processing Systems*, pp. 985–992, 2006.

[73] J. S. Hare, S. Samangooei, and P. H. Lewis, "Efficient clustering and quantisation of SIFT features: exploiting characteristics of the SIFT descriptor and interest region detectors under image inversion," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11)*, pp. 1–8, April 2011.

[74] H. Jegou, H. Harzallah, and C. Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[75] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1800–1807, October 2005.

[76] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Guo, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2664–2677, 2011.

[77] E. Gavves, C. G. M. Snoek, and A. W. Smeulders, "Visual synonyms for landmark image retrieval," *Computer Vision and Image Understanding*, vol. 116, pp. 238–249, 2012.

[78] R. J. López-Sastre, T. Tuytelaars, F. J. Acevedo-Rodríguez, and S. Maldonado-Bascón, "Towards a more discriminative and semantic visual vocabulary," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 415–425, 2011.

[79] S. H. Bae and B. H. Juang, "IPSILON: incremental parsing for semantic indexing of latent concepts," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1933–1947, 2010.

[80] K. Kesorn and S. Poslad, "An enhanced bag-of-visual words vector space model to represent visual content in athletics images," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 211–222, 2012.

[81] P. Tirilly, V. Claveau, and P. Gros, "Language modeling for bag-of-visual words image categorization," in *Proceedings of the International Conference on Image and Video Retrieval (CIVR '08)*, pp. 249–258, July 2008.

[82] H. Cheng and R. Wang, "Semantic modeling of natural scenes based on contextual Bayesian networks," *Pattern Recognition*, vol. 43, no. 12, pp. 4042–4054, 2010.

[83] D. Larlus, J. Verbeek, and F. Jurie, "Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 238–253, 2010.

[84] Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.

[85] B. Fasel, F. Monay, and D. Gatica-Perez, "Latent semantic analysis of facial action codes for automatic facial expression

recognition," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, pp. 181–188, October 2004.

[86] J. Wang, Y. Li, Y. Zhang et al., "Bag-of-features based medical image retrieval via multiple assignment and visual words weighting," *IEEE Transactions on Medial Imaging*, vol. 30, no. 11, pp. 1996–2011, 2011.

[87] X. Li and A. Godil, "Investigating the bag-of-words method for 3D shape retrieval," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, Article ID 108130, 2010.

[88] R. Toldo, U. Castellani, and A. Fusiello, "A bag of words approach for 3D object categorization," in *International Conference on Computer Vision/Computer Graphics Collaboration Techniques*, pp. 116–127, 2009.

[89] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 2012.

[90] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1778–1785, June 2009.

[91] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 291–330, 2008.

[92] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: improving particular object retrieval in large scale image databases," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[93] R. Lienhart and M. Slaney, "PLSA on large scale image databases," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. IV1217–IV1220, April 2007.

[94] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proceedings of the 17th ACM International Conference on Multimedia (MM '09)*, pp. 75–84, October 2009.

[95] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 1521–1528, 2011.

[96] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[97] N. M. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez,"Discriminative compact pyramids for object and scene recognition," *Pattern Recognition*, vol. 45, pp. 1627–1636, 2012.

[98] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[99] L. Shang and B. Xiao, "Discriminative features for image classification and retrieval," *Pattern Recognition Letters*, vol. 33, pp. 744–751, 2012.

[100] W. Tong, F. Li, R. Jin, and A. Jain, "Large-scale near-duplicate image retrieval by kernel density estimation," *International Journal of Multimedia Information Retrieval*, vol. 1, pp. 45–58, 2012.

[101] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[102] S. Romberg, R. Lienhart, and E. Horster, "Multimodal image retrieval: fusing modalities with multilayer multimodal pLSA," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 1, pp. 31–44, 2012.

[103] Y. J. Lee and K. Grauman, "Object-graphs for context-aware visual category discovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 346–358, 2012.

[104] J. Stottinger, A. Hanbury, N. Sebe, and T. Gevers, "Sparse color interest points for image retrieval and object categorization," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2681–2692, 2012.

[105] G. Ding, J. Wang, and K. Qin, "A visual word weighting scheme based on emerging itemsets for video annotation," *Information Processing Letters*, vol. 110, no. 16, pp. 692–696, 2010.

[106] J. Qin and N. H. C. Yung, "Scene categorization via contextual visual words," *Pattern Recognition*, vol. 43, no. 5, pp. 1874–1888, 2010.

[107] P. Tirilly, V. Claveau, and P. Gros, "Distances and weighting schemes for bag of visual words image retrieval," in *Proceedings of the ACM SIGMM International Conference on Multimedia Information Retrieval (MIR '10)*, pp. 323–332, March 2010.

[108] Y. Xiang, X. Zhou, T. S. Chua, and C. W. Ngo, "A revisit of generative model for automatic image annotation using markov random fields," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1153–1160, June 2009.

[109] M. Marszalek and C. Schmid, "Constructing category hierarchies for visual recognition," in *European Conference on Computer Vision*, pp. 479–491, 2008.

[110] K. E. A. Van De Sande, T. Gevers, and C. G. M. Snoek, "A comparison of color features for visual concept classification," in *Proceedings of the International Conference on Image and Video Retrieval (CIVR '08)*, pp. 141–150, July 2008.

[111] L. J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[112] Y. Junsong, W. Ying, and Y. Ming, "Discovery of collocation patterns: from visual words to visual phrases," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[113] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *European Conference on Computer Vision*, pp. 464–475, 2006.

[114] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[115] H. Lee, G. Shim, Y. B. Kim, J. Park, and J. Kim, "A search ant and labor ant algorithm for clustering data," in *International Conference on Ant Colony Optimization and Swarm Intelligence*, pp. 500–501, 2006.

[116] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[117] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[118] 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 *European Conference on Computer Vision*, pp. 97–112, 2002.

[119] M. Stark and B. Schiele, "How good are local features for classes of geometric objects," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[120] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1271–1278, June 2009.

[121] D. Nister and H. Stewenius, "Scalable recognition with vocabulary tree," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1470–1477, 2006.

[122] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[123] M. R. Boutell, J. Luo, and C. M. Brown, "Scene parsing using region-based generative models," *IEEE Transactions on Multimedia*, vol. 9, no. 1, pp. 136–146, 2007.

[124] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 119–126, 2003.

[125] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *International Conference on Neural Information Processing Systems*, pp. 553–560, 2003.

[126] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, 2007.

[127] C. Siagian and L. Itti, "Gist: a mobile robotics application of context-based vision in outdoor environment," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 1063–1069, 2005.

[128] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007.

[129] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 11–18, August 2003.

[130] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 209–216, June 2007.

[131] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *International Conference on Neural Information Processing Systems*, pp. 513–520, 2004.

[132] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," in *International Conference on Neural Information Processing Systems*, pp. 1473–1480, 2006.

[133] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the 9th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval (MIR '07)*, pp. 197–206, September 2007.

[134] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. 1002–1009, July 2004.

[135] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2033–2040, June 2006.

[136] J. Liu and M. Shah, "Scene modeling using co-clustering," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[137] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. J. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[138] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: an in-depth study," Tech. Rep. RR-5737, INRIA Rhône-Alpes, 2005.

[139] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *European Conference on Computer Vision*, pp. 71–84, 2004.

[140] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor, "Improving "bag-of-keypoints" image categorication," Tech. Rep., University of Southampton, 2005.

[141] T. Deselaers, D. Keysets, and H. Ney, "Classification error rate for quantitative evaluation of content-based image retrieval systems," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 505–508, August 2004.

[142] F. Li, W. Tong, R. Jin, A. K. Jain, and J. E. Lee, "An efficient key point quantization algorithm for large scale image retrieval," in *Proceedings of the 1st ACM Workshop on Large-Scale Multimedia Retrieval and Mining (LS-MMRM '09)*, pp. 89–96, October 2009.

[143] A. K. Bhogal, N. Singla, and M. Kaur, "Comparison of algorithms for segmentation of complex scene images," *International Journal of Advanced Engineering Sciences and Technologies*, vol. 8, no. 2, pp. 306–310, 2011.

[144] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: a survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[145] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: using generative information in discriminative classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1249–1262, 2012.