Contextual Bag-of-Words for Visual Categorization

Teng Li, Tao Mei, In-So Kweon, *Member, IEEE,* and Xian-Sheng Hua, *Member, IEEE*

***Abstract*—Bag-of-words (BOW), which represents an image by the histogram of local patches on the basis of a visual vocabulary, has attracted intensive attention in visual categorization due to its good performance and flexibility. Conventional BOW neglects the contextual relations between local patches due to its Naïve Bayesian assumption. However, it is well known that contextual relations play an important role for human beings to recognize visual categories from their local appearance. This paper proposes a novel contextual bag-of-words (CBOW) representation to model two kinds of typical contextual relations between local patches, i.e., a semantic conceptual relation and a spatial neighboring relation. To model the semantic conceptual relation, visual words are grouped on multiple semantic levels according to the similarity of class distribution induced by them, accordingly local patches are encoded and images are represented. To explore the spatial neighboring relation, an automatic term extraction technique is adopted to measure the confidence that neighboring visual words are relevant. Word groups with high relevance are used and their statistics are incorporated into the BOW representation. Classification is taken using the support vector machine with an efficient kernel to incorporate the relational information. The proposed approach is extensively evaluated on two kinds of visual categorization tasks, i.e., video event and scene categorization. Experimental results demonstrate the importance of contextual relations of local patches and the CBOW shows superior performance to conventional BOW.**

***Index Terms*—Bag-of-words, conceptual relation, local patches context, neighboring relation.**

I. Introduction

THE POPULARITY of the internet has caused an exponential increase in the amount of online video data and in the number of users. Visual categorization, which can be used for indexing, searching, filtering, and mining large amounts of video data, becomes increasingly important for users. For example, we can group the video frames according to the high-level concepts they contain or group the scenes they happened in such as *Indoor*, *Beach*, *People_Marching*, and so on, for efficient browsing.

Conventional methods of visual categorization usually represent an image based on the low level global features such as "gist," Gabor filters, color moment, texture from the whole image or from a fixed spatial layout [1], [22], [37], which is convenient for categorization and is computationally efficient. [5] tried using high-level global features to determine the semantic class of a scene utilizing a semantic object detector and generative scene-configuration models. The main drawback of global feature-based methods is their sensitivity to scale, pose and lighting condition changes, clutter, and occlusions. Recently, categorization based on local features in the image has attracted intensive attention in visual categorization, for its robustness to intra-class variations. Local features detected from an image are of various numbers and in a different order, therefore we cannot apply the classification algorithms directly. Some methods are proposed to define a matching function [9], [11] or find the correspondences [4] to measure the similarity between local feature sets directly. Though effective, they are unpractical to be applied to large scale datasets, such as TRECVID corpus [33], due to the high-computational complexity.

Originating from the text categorization area, bag-of-words (BOW) has become a popular method for visual categorization for its effectiveness and flexibility. With extracted local features from images, a visual vocabulary is built by clustering the local features to visual words, which are analogous to the words in text documents. Then each local feature is encoded by mapping to a visual word, and an image can be represented as a BOW, or specifically, a vector containing the count of each visual word in that image [8]. In this process, the visual vocabulary provides an intermediate helping to convert the chaotic local feature set to a regular representation vector, based on which it is convenient to apply the machine learning techniques, such as support vector machine (SVM), to yield good performance. Joining the robustness of local feature matching and the practicality of vector representation, the BOW model has been applied to various tasks, such as image categorization [40], video object retrieval [31], near duplicate detection [36], etc., and shown excellent performance. On several benchmark datasets, for example the PASCAL visual object classes, the BOW-based methods achieved a state of the art performance [28].

Though shown to be very effective [40], the BOW assumes that local features in an image are independent to each other given the class, i.e., the Naïve Bayesian assumption, which means the contextual relations between local patches are neglected. Contextual information is important in the recognition process of human beings. A white image patch is likely to be the cloud if it is in a sky area, while it could be a sheep if surrounded by grass. In the text area, the relation between words can be utilized to help understanding. One can expect to find certain letters occurring regularly in particular arrangement with other letters. With the visual words analogy, encoding the local features as visual words, it is natural that

Manuscript received September 8, 2008; revised April 13, 2009, and July 8, 2009. Date of publication January 29, 2010; date of current version April 1, 2011. This work was performed while T. Li was visiting Microsoft Research Asia as an Intern. This paper was recommended by Associate Editor, L. Guan.

T. Li and I.-S. Kweon are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: tengli@rcv.kaist.ac.kr; iskweon@rcv.kaist.ac.kr).

T. Mei and X.-S. Hua are with Microsoft Research Asia, Beijing 100190, China (e-mail: tmei@microsoft.com; xshua@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2010.2041828

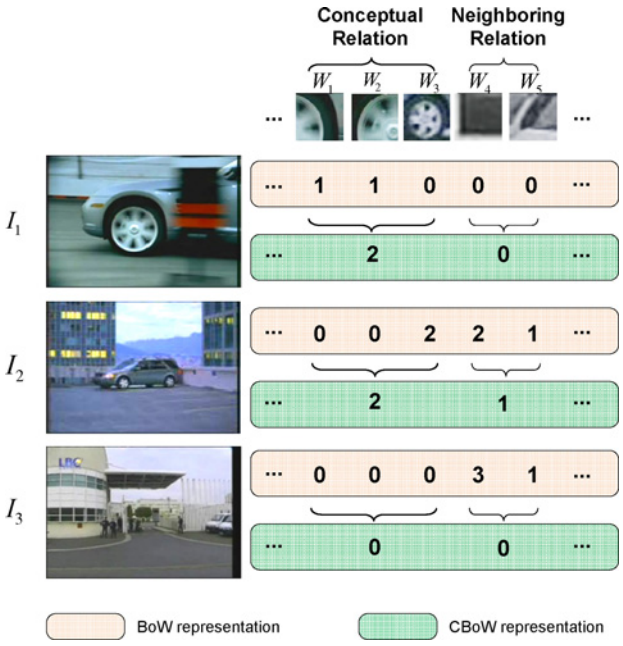

Fig. 1. BOW and CBOW for three video keyframes in terms of five visual words. The representations are shown to the right of each keyframe. Keyframes I_1 and I_2 belong to the car concept while I_3 does not. Regarding BOW, I_2 has more histogram intersection with I_3, while with CBOW, where the contextual relations are considered, I_2 has more intersection with I_1.

the context of related visual words can be considered for better categorization.

Two relations between local patches in images or video keyframes can be important for categorization. First, there is the semantic conceptual relation between patches. An image or video frame can be described by the composition of objects such as cars, buildings, and persons. The objects can be further described in terms of parts, e.g., a wheel of a car, a window of a building, or the face of a person. On the bottom level, the local patches have the relation of appearing on the "same part," "same object," or "same category" [32]. In BOW, visual words are usually learned by clustering over features, in terms of the visual appearance. Different words may correspond to the same concept, i.e., they are *conceptually related*. As a result the corresponding features in different images may be encoded to different bins and classification performance is affected. Second, the spatial neighboring context of patches is totally neglected in the BOW. In many cases, two patches appearing together, i.e., having the *neighboring relation*, can give more information for classification than appearing separately.

Fig. 1 illustrates the effect of these two relations for categorization using three video keyframes and five visual words. Keyframes I_1 and I_2 belong to the car concept and keyframe I_3 does not. Visual words W_1, W_2, and W_3 all contain the concept of "tire" but have different appearances due to the imaging variation and limitation of the local patch extractor. Visual words W_4, W_5 occur in both I_2 and I_3 while they are a neighbor forming an informative part only in I_2. By the original BOW, the representation for each keyframe is shown in the first line of its right area, according to which I_2 has more histogram intersection with I_3 than with I_1, which is not expected in categorization. As marked on the top, we consider the conceptual relation of W_1, W_2, and W_3, and the neighboring relation of W_4 and W_5. For *conceptual relation*, the occurrence of a word also indicates the concept of its relational words, and for *neighboring relation*, the occurrence of relational words in the neighborhood should be considered. Therefore, we group the patches of the "tire" concept together and count the occurrence of neighboring W_4 and W_5 to obtain a new contextual bag-of-words (CBOW) representation for each keyframe, as shown in the second line of its right. By the CBOW, obviously I_2 is more matched with I_1 than with I_3, therefore categorization can be facilitated.

Although the contextual relations between local patches are useful, they have not been well explored in the BOW-based visual categorization. Most approaches group local features into separate bins of visual words and treat these words independently when comparing or categorizing. Different weighting schemes, such as binary or term frequency (TF) [8], [40], term frequency–inverse document frequency [31], and binary [27], have been proposed for considering the significance of individual words. However, the visual words context has not been considered in this process. In [17], Lazebnik *et al.* consider the spatial layout relation of local features by partitioning an image into increasingly fine grids and computing the BOW inside each grid cell. It shows better performance than the original BOW and validates the importance of the local patches context for visual categorization. The spatial layout relation of local patches is still rough, though, and the two contextual relations addressed in this paper have not been considered yet.

In this paper, we propose a novel visual categorization algorithm to model the two contextual relations between local patches based on the BOW representation. Firstly, the semantic conceptual relation is measured according to the class distribution induced by the visual words. The distributional similarity is measured by the Kullback–Leibler (KL) [15] divergence. With different similarity criteria, relational visual words are grouped on multiple semantic levels and images are represented accordingly. The multiple level conceptual relation is integrated into the classification by a kernel design based on the pyramid matching theory. Moreover, to evaluate the neighboring relation of visual words, the automatic term extraction from the text area is adopted, which calculates a confidence value that neighboring words can form an informative part. Informative word groups with high confidence are then extracted and their statistical information is combined with the BOW representation.

We studied the effectiveness of the proposed contextual relations modeling method on two visual categorization tasks: scene categorization and video event categorization. Experiments are conducted on the 15 scene categories dataset and TRECVID2005 events detection corpus. Comparisons with previous methods are taken. The spatial layout relation is further combined to extensively explore the effective of local patches context in categorization. In the following, we also use conceptual relation and neighboring relation to denote semantic conceptual relation and spatial neighboring relation, respectively.

The rest of this paper is organized as follows. Section II briefly reviews the related works. Section III presents the de-

tails of the proposed visual categorization approach, including conceptual relation modeling, neighboring relation modeling, and classification scheme. Section IV gives experimental results on two benchmark datasets. Finally, Section V concludes this paper.

II. Related Work

Since being introduced, BOW has attracted intensive attention. Some works have studied the parameters or feature settings of BOW comparatively to yield high-categorization performance [38]. Many algorithms were also proposed to improve the method itself. Different clustering techniques, such as agglomerative [13], mean-shift [18] or hierarchical k-means [26], have been adopted for visual vocabulary learning. To introduce discrimination to the visual words or integrate the visual words learning step into the classification scheme, Winn *et al.* [34] proposed to build a compact and discriminative codebook by pair wise merging of visual words based on the information bottleneck principle, and Moosmann *et al.* [24] applied the randomized forest method to codebook learning. Recently, [16] proposed to learn the codebook by minimizing information loss. These algorithms aim at improving the visual words to encode local features efficiently, therefore lead to better categorization performance or high speed.

In contrast, to minimize the gap between visual words encoding and the semantic concepts, [10], [30] try to extract the middle level topics based on BOW and model categories in terms of the semantic topics. They apply the probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA), which originate from the text processing area. Each topic has a probabilistic distribution over the words and image categories are modeled in terms of the distribution of topics. The learning process can be unsupervised and pLSA is also applied to human action categorization [25] and video object discovery [19]. Though appealing in theory, later work shows that BOW still keeps the prior position in terms of categorization performance [17]. None of the above algorithms have considered the contextual relations between local patches in the image.

In [17], the spatial layout relation of local features is considered assuming that similar parts of scene categories often appear in similar areas of 2-D image space. Images are partitioned into increasingly fine grids and histograms are computed for patches found inside each grid cell, based on which the pyramid matching is adapted for classification, named spatial pyramid matching (SPM). It obtains better categorization performance than the original BOW, but the spatial layout relation is still rough and the contextual relations we are going to address in this paper have not been considered. Liu *et al.* [20] proposes to group visual words to intermediate concepts by co-clustering and reducing the dimensions of image representation for efficient computing, where visual words are clustered semantically. However, their aim is to be more efficient and the conceptual relation is not well combined for better performance.

The neighboring relation between local patches has been explored in Bayesian framework-based object categorization and image retrieval tasks. Wu *et al.* [35] proposed to use the visual language model (VLM) to statistic distribution of the neighboring visual words (N-grams) to describe the image category. Modeling the probability of every possible N-gram in a Bayesian framework, the VLM cannot be naturally combined with the BOW representation, thus it is weak in categorization performance. Zheng *et al.* [42] and Zhang *et al.* [41] extract visual phases, i.e., neighboring words groups, simply according to their occurrence frequency for image retrieval. This paper is different from them in that we explore the contextual relations between local patches for more effective visual categorization in the BOW framework.

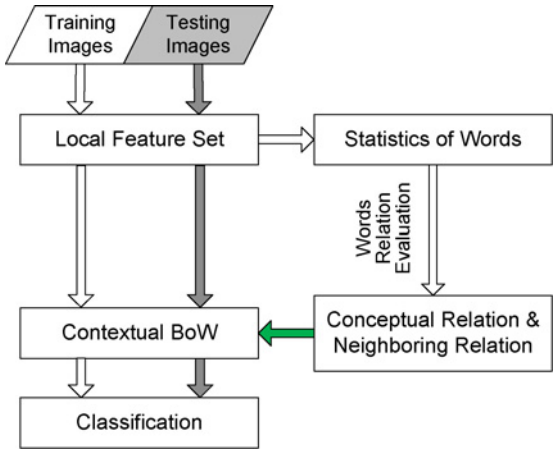

Fig. 2. Overview of the proposed visual categorization approach. White and gray pipelines mark the flow of training and test keyframes, respectively, and the green pipeline represents the contextual relations that are used in the representation computing process.

III. Contextual Bag-of-Words Categorization

There are three main steps in the proposed approach for visual categorization. First, similar to BOW, local features are extracted from images or video keyframes and translated to the feature descriptors; then visual words are learned by clustering. Second, the occurrence numbers of visual words are counted and the two contextual relations between visual words are measured from the statistics. Finally, the images are represented and classification is taken considering the relation information. Fig. 2 shows an overview of the proposed approach. White and gray pipelines mark the flow of training and test images or video keyframes respectively, and the green pipeline represents the contextual relations that are used in the CBOW representation computing process.

A. Feature Extraction and Visual Words Learning

In the feature extraction, local patches used are extracted densely from the images and translated to scale invariant feature transform (SIFT) descriptors [21]. Some previous works use interest point detectors for local patches extraction, but the features detected are usually too sparse to describe the visual characteristics. Recent research shows that extracting local patches densely can yield better performance [27]. Thus, we extract the patches centering on a regular grid with spacing

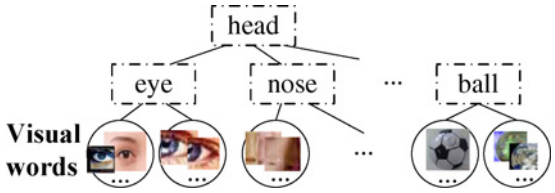

Fig. 3. Illustration of the patches conceptual relation. Patches of the same concept may correspond to different visual words, which is referred to as "conceptual relation" here. The appearance of a word implies the information of its relational words.

M pixels and calculate the descriptor of each local patch. As a result each image is translated to a set of local features. The SIFT descriptor is widely used and is shown to be effective in the performance evaluation of [23], and the PCA-SIFT [14], which extends the original SIFT descriptors, has been shown to be more compact and distinctive. Conventionally, SIFT is computed for eight orientation planes and each gradient region is sampled over a 4×4 grid of locations. Thus, the dimension of the resulting descriptor is 128. In this paper, since the main orientation of the sampled local patches is unknown, for each sampled patch we use two orthogonal orientations as the main orientation and concatenate the descriptors calculated accordingly, resulting in a vector of 256 dimensions. Then the principle component analysis (PCA) is applied to transform the features to 80 dimensions to reduce the computation and storage cost. The PCA transform is learned using a randomly selected subset of training features and applied to all the local features.

The visual words are learned from a collection of local patches sampled from the training images using the k-means clustering algorithm, which efficiently groups visually similar patches into one cluster. The visual words are set as the cluster centroids. With a visual vocabulary, each local feature can be encoded as a word by a vector quantization algorithm, i.e., the nearest word to it. For each keyframe or image, the occurrence number of each word can be counted to form a histogram representation.

B. Conceptual Relation Modeling

The conventional BOW encodes local patches purely according to the visual appearance and considering visual words independently. However, as illustrated in Fig. 3, some words are closely related in concept while some are not. The occurrence of a visual word conveys similar information to its relational words, which should be considered in the categorization. This section details the process that measures and incorporates the conceptual relation between local patches in the categorization.

Evaluating the conceptual relation between local patches has rarely been addressed in visual categorization. However, in the text area, the relation between words can be obtained from the WordNet [29], which is built manually. In informative clustering, semantic distance between words is measured implicitly or explicitly to group the semantically similar words [3]. Most of the methods are essentially based on the information gain criterion. Among them, an effective way is to measure the word relation by the KL divergence between distributions of the classes induced by the words, and it also measures the distance explicitly and is convenient to be incorporated into categorization frameworks [2], [3]. The distribution is computed according to the statistical information of words' occurrence [2]. In BOW, with the visual words analogy, the conceptual relation between visual words can be measured similarly. The core intuition behind this measurement is that visual words related to the same objects or object parts are more likely to distribute similarly over the categories. For example, patches of "eye" or "nose" tend to occur frequently in face images and less in other categories.

To illustrate the conceptual relation of visual words derived from their distribution, in Fig. 4, four visual words from 1000 words learned by k-means clustering (from a 15 scene categories dataset) are used. The above figure shows some sample patches of these visual words extracted from the training images in row sequence. We can see visual words #2, #352, and #503 are closely related to the "coastline" part of the coast category and therefore have strong conceptual relation, while visual word #4 has a very small conceptual relation with them. The below figure plots the class distributions induced by these words in the training set. The horizontal axis represents the class variable, the vertical axis indicates the probability of each class given the word, and the shape of the line shows the distribution. As we can see, the line shape of distributions of the three relational words is quite similar, while that of word #4 is obviously different. Thus, using the induced class distribution, the conceptual relation between visual words can be calculated. Considering the classification task, the graph of class distributions can also be interpreted as a picture of how much the word votes for each of the classes whenever it occurs, and it can be seen from Fig. 4 that the three relational visual words vote mostly for the coast category, with the other voting mostly for the forest category.

Consider the distribution of a particular word W_t over a class C_j, i.e., the probability of C_j given W_t

$$P(C_j|W_t) = \frac{P(C_j, W_t)}{P(W_t)}. \tag{1}$$

This probability is approximately calculated by counting the occurrence number of visual words, i.e., the number W_t occurring in class C_j versus its occurrence number in all the classes. To measure the difference between two conditional distributions, the KL divergence, also called information divergence, is used. The KL divergence between the distributions of class variable C induced by W_t and W_s is defined as

$$KL\Big(P(C|W_t)||P(C|W_s)\Big) = \sum_{j=1}^{|C|} P(C_j|W_t) \log\Big(\frac{P(C_j|W_t)}{P(C_j|W_s)}\Big). \tag{2}$$

In the context of information theory, the KL divergence can be intuitively understood as a measure of inefficiency that occurs when messages are sent according to one distribution, $P(C|W_t)$, but encoded with a code that is optimal for a different distribution, $P(C|W_s)$.

Since the KL divergence is not symmetric, and it is infinite when an event with nonzero probability in the first distribution

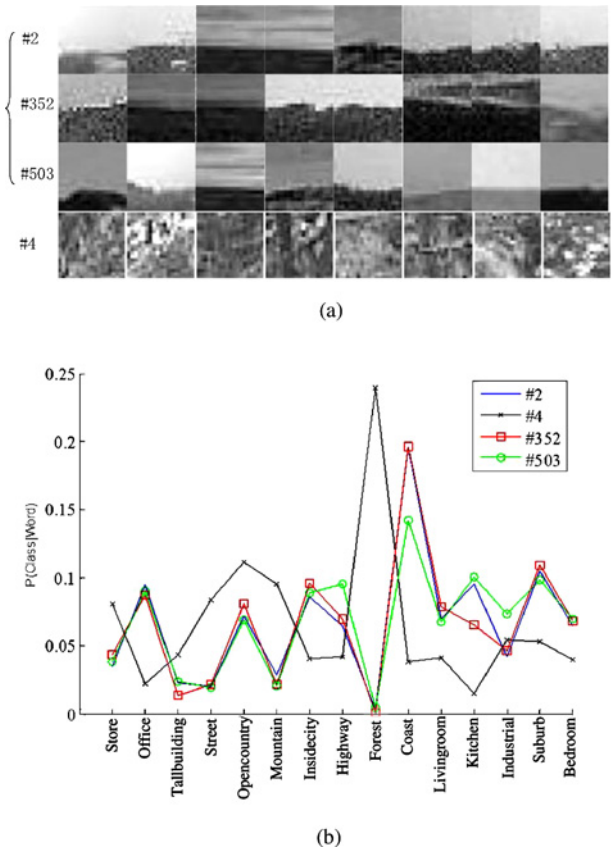

Fig. 4. Conceptual relation of patches from the class distribution induced by visual words. (a) Sample patches corresponding to words #2, #352, #503, and #4 from 1000 words leaned by k-means in row sequence. (b) Plots the class probability distribution of these words. Words #2, #352, and #503 have similar distribution and are grouped as relational in the proposed method while word #4 is not.

has zero distribution in the second distribution, here we use a related measure that does not have these problems. It is a weighted average of the KL divergence of each distribution to their mean distribution

$$D(W_t, W_s) = \frac{P(W_t)}{P(W_t \vee W_s)} * KL\Big(P(C|W_t)||P(C|W_t \vee W_s)\Big) + \frac{P(W_s)}{P(W_t \vee W_s)} * KL\Big(P(C|W_s)||P(C|W_t \vee W_s)\Big). \quad (3)$$

When words W_t and W_s are considered as relational and grouped together, the new distribution $P(C|W_t \vee W_s)$ is the weighted average of the individual distributions. This metric can be understood as the expected amount of inefficiency incurred if, instead of encoding two distributions optimally with their own code, we use the code that would be optimal for their mean. Here it describes the difference of effect to the image representation that relational words that formerly generated their own individual statistics now generate combined statistics.

With the informative measurement of (3), from an initial vocabulary of K words, we can group the visual words to a pre-defined number of relational groups by an agglomerative procedure as in the following, where the group number is denoted as K_C.

1) Calculate the conditional distribution over the class variable of each original word $P(C_j|W_t)$. According to (2) and (3), calculate the semantic distance $D(W_t, W_s)$ between each words pair (W_t, W_s).
2) Initially, each word itself is considered as a relational group. Iteratively merge the two groups whose distance value is the smallest until the group number equals to K_C. The distance between two groups (WC_t, WC_s) is defined as the shortest distance that a word in cluster WC_t to a word in cluster WC_s

$$D(WC_t, WC_s) = \min_{i \in I, j \in J} D(W_i, W_j) \quad (4)$$

where I, J represent the collection of words in group t and group s, respectively.

Since words of the same conceptual group reflect similar semantic meaning, their occurrence is considered as the occurrence of the group. The occurrence numbers of all the groups can form a histogram representation similar to that of the BOW, which is used in the categorization. Furthermore, semantic relation between patches can be interpreted in multiple levels; containing same scene, object, object parts, or intermediate concepts. It is impossible to precisely measure visual words relation corresponding to these levels; here we propose to incorporate the conceptual relations between visual words on multiple levels in an approximate way, which will be detailed in Section III-D.

C. Neighboring Relation Modeling

In images or video keyframes, some patches can be combined to form a meaningful object or object part, which is similar to the terms constituted by closely relevant words in text. These patches are considered as having a "neighboring relation," as illustrated in Fig. 1. To incorporate the neighboring relation into the BOW, we measure the information that the neighboring visual word groups give for classification and use the occurrence number of informative groups in image representation.

In natural language words' contextual information is usually modeled by the N-gram language model in text categorization, according to the grammar which restricts the words connection and order. The N-gram model estimates the conditional probability of word sequences of length $N(N \geq 2)$, and set as the prior knowledge for understanding the text. Many works model the neighboring relation between words using the conditional probability assuming the Markov property or words sequence, such as VLM [35]. As we have explained, it is hard to integrate with the BOW. To directly adopt the N-gram in the BOW, and count the number of N-grams in the image to construct a "Bag of N-grams," we encounter the practical problem that the vector dimension for representing an image is too high to compute. For example, the number of Bi-grams constituted by 1000 visual words is 500 000 without considering the order. At the same time, many word pairs are useless for the classification and even seldom appear. Therefore, we need to extract the informative neighboring visual words groups and model them based on the BOW.

TABLE I
OBSERVED FREQUENCIES OF WORD PAIR

	$W_t = B$	$W_t \neq B$	
$W_s = A$	O_{11}	O_{12}	Ro_1
$W_s \neq A$	O_{21}	O_{22}	Ro_2
	Co_1	Co_2	

TABLE II
EXPECTED FREQUENCIES OF WORD PAIR

	$W_t = B$	$W_t \neq B$
$W_s = A$	$E_{11} = \frac{Ro_1 \times Co_1}{N}$	$E_{12} = \frac{Ro_1 \times Co_2}{N}$
$W_s \neq A$	$E_{21} = \frac{Ro_2 \times Co_1}{N}$	$E_{22} = \frac{Ro_2 \times Co_2}{N}$

Various feature selection or key term extraction techniques proposed in the text area can be applied. Among them, the automatic terms extraction technique with the chi-square criterion, which is commonly used and is shown to be very effective [39], is adopted. Automatic terms extraction is used to find relevant word groups through the statistic information from the corpus. It calculates a confidence value that a pair of words constitutes a term. Considering any two consecutive words W_s and W_t in the corpus forming a word pair (W_s, W_t), for each word pair we can get the contingency table of observed frequencies O_{ij} as Table I. Where O_{11} represents the frequency of word pair (W_s, W_t) in the corpus when W_s is A and W_t is B. O_{12} represents the frequency that W_s is A and W_t is not B. O_{21} represents the frequency that W_s is not A and W_t is B. O_{22} represents the frequency that W_s is not A and W_t is not B. Based on Table I, the expected values of these frequencies, if consecutive words A and B form a meaningful term, can be calculated according to Table II. The E_{ij} values are the expected occurrence number of the corresponding cases to the O_{ij} values in Table I. N is the frequency of all word pairs in the corpus. Using the above two tables, for any of the two words A and B, the confidence that they form a term can be calculated by the following (5). Word pairs with a high-confidence value are considered as the terms

$$\text{Confidence}(A, B) = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}. \tag{5}$$

In the BOW representation, an image is analogous to a document, thus by the above method the confidence value that a group of neighboring visual words form an informative part can also be calculated. For convenience, here we use N-gram to name the group of N neighboring words and define the Bi-gram as the visual word pair occurring in neighbor, and the $(N + 1)$-gram as the neighboring pair of an N-gram and a word. Fig. 5 illustrates the definition of "visual N-gram." After translating the features to visual words, we consider their 8-neighborhood relation in the image space and extract the neighboring word groups as shown in Fig. 5. Using the above confidence evaluating method, informative neighboring N-grams are extracted by the following procedure.

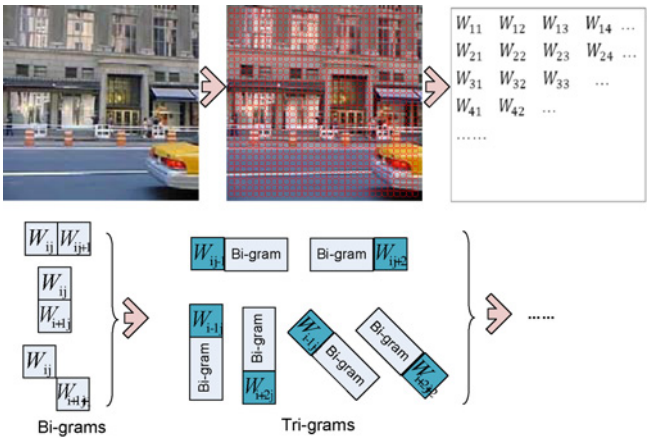

Fig. 5. Illustration of the visual N-grams extraction. The top middle image shows the extracted local features from the top left image, and the features are converted to visual words as at the top right image. The below figures illustrate the "visual N-grams" we define.

1) Count the occurrence of all the word pairs in the training set, calculate the confidence value of each word pair and extract those whose confidence is high to construct the Bi-gram terms.
2) Find the informative $(N + 1)$-grams by calculating the confidence of an N-gram with a word and extract these with high-confidence values. This step can be iterated.

In practice, visual word groups with a small occurrence number are neglected. Finally, we count the occurrence number of the informative neighboring N-grams and concatenate to the original BOW representation.

D. Classification

Considering the two contextual relations together, an image can be finally represented by a multilevel BOW representation by first incorporating the neighboring relation then applying the multiple level conceptual relation modeling method. With the conceptual distance between visual words measured by the method of the previous section, the initial K visual words can be grouped to K_C relational groups by an agglomerative procedure. To incorporate the words relation information on multiple levels, we group the relational words with different similarity criterions resulting in different levels of the group numbers K_Cs. K_Cs are set as the following:

$$K\_C_l = K/2^l, \quad l = 0, \ldots, L-1 \tag{6}$$

where L represents the number of conceptual levels considered. On each level a histogram representation can be computed; as a result, an image is represented as a multiresolution histogram of the pyramid structure, where higher level features are encoded semantically more coarsely. Based on this representation categorization is taken using the kernel SVM. Here we adopted the idea of the pyramid matching [11] for combining multilevel information in matching.

The pyramid matching works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution, with coarser levels assigned smaller

weights. A sequence of grids at different resolutions $0, \ldots, L$ is constructed over the feature space. At any resolution, two points are considered as a match if they fall into the same cell of the grid. Matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. Specifically, consider that the feature dimension is d and the grid at level l has 2^l cells along each dimension; thus there are in total $T = 2^{dl}$ cells. Let H_x^l and H_y^l denote the histograms of x and y at this resolution, $H_x^l(i)$ and $H_y^l(i)$ are the numbers of points from x and y that fall into the i_{th} cell of the grid. Then the number of matches at level l is given by the histogram intersection function

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{T} \min(H_X^l(i), H_Y^l(i)). \tag{7}$$

Note that the number of matches found at level l also includes all the matches found at the finer level $l+1$. Therefore, the number of new matches found at level l is given by $I^l - I^{l+1}$ for $l = 0, \ldots, L-1$ (abbreviate $I(H_x^l, H_y^l)$ as I^l). The weight associated with level l is set to $1/2^{L-l}$, which is inversely proportional to cell width at that level. Intuitively, matches found in larger cells should be weighted lower because the corresponding features are increasingly dissimilar. In summary, a pyramid match kernel (PMK) is defined as

$$\begin{aligned} K_{\mathrm{PMK}}^L(x, y) &= I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^l + 1) \\ &= \frac{1}{2^L} I^0 + \sum_{l=0}^{L} \frac{1}{2^{L-l}+1}. \end{aligned} \tag{8}$$

Both the histogram intersection and the pyramid match kernel are Mercer kernels [11].

SVM has demonstrated its effectiveness in many categorization tasks and shows excellent performance for feature combination. The kernel is important for SVM. In previous works using BOW, several kernel types have been adopted, such as histogram intersection and linear. The proposed matching kernel is defined based on the Laplacian radial basis function (LRBF) proposed in [7], which has shown superior performance in histogram-based image categorization. Given two images which are represented by K-dimensional vectors x and y, respectively, the LRBF kernel is defined as

$$K_{\mathrm{LRBF}}(x, y) = \exp\left[-\frac{Dis(x, y)}{A}\right] \tag{9}$$

with the distance function

$$Dis(x, y) = \sum_{i=1}^{K} |x(i) - y(i)|. \tag{10}$$

In this paper, representing an image on L levels, x and y each contain L vectors of different dimensions as defined in (6). The multilevel representation of images corresponds to different fine levels in conceptual relations. A lower level means the words groups have a finer relation and the matched local features are semantically more similar. Motivated the pyramid matching kernel, the kernel function for image matching is defined as

$$Dis(x, y) = \sum_{l=0}^{L-1} a_l d_l(x, y) \tag{11}$$

where a_l is the weight at level l defined as $a_l = 1/2^l$, d_l represents the distance between x and y at level l, and is defined as

$$d_l(x, y) = \sum_{i=1}^{K-C_l} |x^l(i) - y^l(i)|. \tag{12}$$

The parameter A in the LRBF kernel is set as the mean value of the defined distances between all training images in the implementation. This parameter setting works well in experiments.

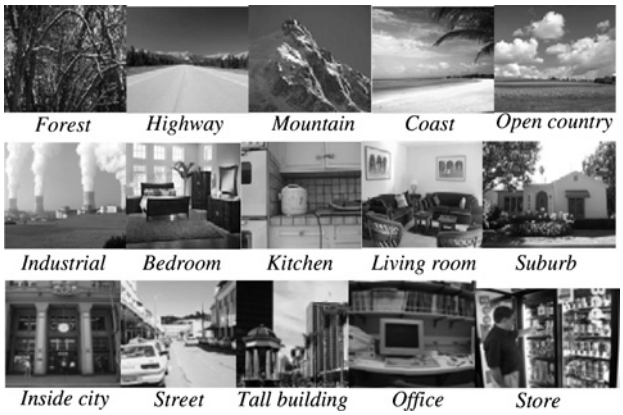

Fig. 6. Example images of the 15 scene categories.

IV. Experimental Result

In this section, we evaluate the proposed algorithm for two kinds of visual categorization tasks: scene and video event categorization, which can be used for understanding what and where the video is. Two widely used benchmarks, the TRECVID2005 video events database [33] and 15 scene categories database [17], are used in the experiments. For feature extraction, gray level images are used for both datasets. The sampling interval M of local patches is set as 8 and 10 for the two datasets, respectively. The scale of extracted patches is randomly sampled between 10 to 30 pixels. The SVM is implemented using LIBSVM [6]. Multiclass classification is done with the SVM trained using the one-versus-all rule. The parameters of SVM, such as the cost value C, are set empirically and fixed in all the tests for a database. Though cross validation can be used for parameter selection, its computation is high for large datasets and our parameters setting also does well as we observed in experiments.

A. Evaluation on Scene Categorization

The scene database is composed of 15 scene categories: *Store*, *Office*, *Tall building*, *Street*, *Open country*, *Mountain*, *Inside city*, *Highway*, *Forest*, *Coast*, *Living room*, *Kitchen*, *Industrial*, *Suburb*, and *Bedroom*. Each category has 212 to

TABLE III
COMPARISON OF DIFFERENT KERNEL TYPES WITH BOW

Kernel	Intersection	Linear	Polynomial	LRBF
AR	79.8%	78.2%	77.9%	81.3%

TABLE IV
PERFORMANCE OF BOC AND BOW WITH DIFFERENT VOCABULARY SIZES

K/K_C	100	200	400	1000	1500
BOW	74.7%	76.0%	80.3%	81.3%	80.9%
BoC	77.4%	79.8%	81.3%	X	X

TABLE V
COMPARISON OF DIFFERENT LEVEL NUMBERS OF CBOW

L	1	2	3	4
AR	82.4%	83.1%	83.3%	83.4%

410 images, and average image size is 240×352 pixels. The major sources of the pictures in the dataset include the COREL collection, personal photographs, and Google image search. This is one of the most complete scene category datasets used in the literature thus far. Fig. 6 shows the example images. Following previous works on this dataset [17], [20], we randomly choose 100 images per class for training and use the rest for testing. The classification accuracy rate (AR) is adopted as evaluation criterion, i.e., the number of correctly classified images versus the number of all test images.

We first test the baseline BOW method with different parameter settings. Table III compares the performance of different kernel types with the BOW. The performance of LRBF is shown to be better than the other three popular kernel types, which validates the effectiveness of the proposed CBOW matching kernel based on LRBF. The vocabulary size for the baseline BOW is also tested and the results are shown in the row "BOW" in Table IV. The best performance is achieved when $K = 1000$. The conceptually relational words groups obtained by the proposed method can also be considered as the intermediate concepts, based on which we can construct a "Bag of Concepts" on a single semantic level. The row "BoC" in Table IV gives the performance of the Bag of Concepts with a different words groups number K_C. They are learned from an original vocabulary of 1000. As shown in the table, the BoC is better than the original BOW with the same vocabulary size. This result verifies the effectiveness of the proposed conceptual relation measuring method and shows that encoding the local features semantically is more effective than using the words learned by pure clustering.

The proposed neighboring relation modeling method is then evaluated. An original vocabulary of size $K = 400$ is used in this test since large vocabularies cause much computational load, and the confidence calculation of (5) relies on the precise statistics of visual words in the training corpus, while large vocabularies have too many word pairs to estimate and require much training data. The informative Bi-grams with high-confidence values are extracted and used to construct a "Bag of Bi-grams." The best performance is achieved when using 800 Bi-grams, giving an accuracy rate of 74.1%. It's lower than the original BOW due to the information loss when discarding most word pairs. However, the combination of the 800 Bi-grams and the original 400 words yields an accuracy rate of 82.4%. It is better than the BOW with large vocabularies and proves that the neighboring relation can be helpful. With the proposed method, informative N-grams ($N > 2$) can be further extracted. However, due to the lack of training data for statistics, there are only a few Tri-grams with a high-confidence value.

Using the combination of 800 Bi-grams and 400 words as the initial vocabulary, the conceptual relation modeling method is applied as the proposed CBOW. Table V lists the performance of CBOW with different level numbers, i.e., L in (6). The performance improves as L increases for the consideration of more conceptual relation. However, it does not change much when L is high, since as in (6), the dimension of higher level representation becomes lower and the incorporated information is small. Finally, we used four levels in the following tests on the scene dataset.

TABLE VI
PERFORMANCE COMPARISON ON THE 15 SCENE CATEGORIES

Alg.	VLM	BOW	CBOW	CBOW_SL	SPM	SPM_IC
AR	53.6%	81.3%	83.4%	85.1%	83.4%	83.3%

Table VI gives the performance of the proposed CBOW and the combination of spatial layout context with the CBOW (CBOW_SL), which partitions images into 2×2 regions and concatenates the representation of all the regions. There, recent local feature-based methods, the VLM [35], BOW, SPM, and the SPM with intermediate concepts (SPM_IC) [20], are compared. The best performance of SPM is achieved with vocabulary size $K = 400$ and level number $L = 3$, as it surpasses previous reported results on this dataset by "gist" [1] and pLSA [10]. The proposed CBOW and the previous SPM obtain similar improvements over the baseline BOW, which demonstrates the importance of spatial layout and the proposed contextual relations of local patches for categorization. Note that the dimension for image representation in the SPM is thus $400 * (1 + 4 + 16) = 8400$, while that of the CBOW is $1200 + 600 + 300 + 150 = 2250$, which is much lower. It means the CBOW costs much less computation than the SPM with a similar performance level. To classify the 2985 test images of this dataset, the BOW with 1000 visual words takes about 113 s, the SPM takes 944 s, while the proposed CBOW method requires about 248 s (2.4 GHz CUP and 3G RAM; not including feature extraction time). High performance is achieved by integrating all these contextual relations between local patches together. As shown in Table VI, CBOW_SL yields the best result. This result proves the importance of local patches context for categorization. The VLM shows low performance for its Bayesian classification framework. Trying to model the conditional dependence between visual words, it

Fig. 7. Example keyframes of TRECVID2005 10 high-level video event concepts.

does not utilize the classification power of machine learning techniques such as SVM. Comparatively, the BOW-based methods can be more flexible and therefore more effective.

B. Evaluation on Events Categorization

The TRECVID2005 video event concepts dataset contains a total of 61 901 keyframes from 39 annotated categories. These keyframes are extracted from a variety of real TV news programs. The size of the keyframes is 240×352; it is a very challenging dataset. In the TRECVID, high-level feature extraction task [33], the following ten concepts were chosen for evaluation: *Building*, *Car*, *Explosion_Fire*, *Flag_US*, *Maps*, *Mountain*, *People_Marching*, *Prisoner*, *Sports*, and *Water*. In this test, we also evaluate the ten concepts detection and compare with previous work [37]. Fig. 7 shows the example keyframes. We follow the experimental settings of [22], [37] to partition the data for training and test. In [22], [37], the annotated dataset is partitioned to training, validation and test sets. However, in this experiment, the validation set is not needed for choosing the parameters since the parameter values are fixed for all categories. Therefore, the training and validation partitions are used for training together and the test set is used for test. Since in the training set, the number of negative samples is much larger than the number of positive ones, for each category, the negative training keyframes are down sampled by ten in model learning. The keyframes in the TRECVID2005 set may contain several overlapping concepts and each keyframe may be classified into multiple categories. In this experiment, we posed the multilabel video annotation task into a binary classification problem using binary SVM as the classifier, which classifies the keyframes from one category as positive, with the rest being treated as negative. For each concept a binary classifier is learned and the ten concept classifiers are applied to each test keyframe one by one; finally, the positive ones are output as the resulting multiple labels.

Following previous work on this dataset [22], [37], the average precision (AP) is adopted as the performance measure. For each concept, assuming N retrieved keyframes are ranked, and R of them are relevant ($R < N$), we can define the AP as follows:

$$AP = \frac{1}{R} \sum_{j=1}^{N} \frac{R_j}{j} * B_j \tag{13}$$

where $B_j = 1$ if the j-th shot is relevant, otherwise 0. R_j is the number of relevant keyframes in the top j retrieved keyframes. Precision and recall were also popular performance measures when evaluating detection algorithms [12]. However, both recall and precision must be taken into account simultaneously, which is not convenient in the case of multiple concepts detection. Usually, the AP is the most commonly used performance measure and it takes into account both recall and precision [43]. To observe the performance extensively, we also plot the Precision–Recall curves of some concepts in the results.

Two typical previous works on this dataset, Columbia [37] and MSRA [22], are compared in the test. They use global features with the SVM classifier and yield excellent performance. In [22], several kinds of global features are combined. Here only the best one, 5×5 color moment, is implemented for comparison since the BOW can also be considered as one feature. The original BOW [8], [40] and the SPM [17] are also implemented for comparison. It is difficult to apply other local feature-based algorithms not based on BOW to the large dataset due to the high computation. In [38], the BOW was applied to TRECVID data with extensive comparisons of different parametric settings such as visual vocabulary size and weighting scheme. However, the sparse local feature extraction they used degrades the performance, and the large visual vocabularies used cause a high-computational load. Due to the huge computational load for this large dataset, we fix the initial vocabulary size as $K = 1000$ for each concept and use the TF weighting method. This setting proves to be effective in experiments.

Based on this baseline the proposed conceptual relation modeling method is applied with three levels, i.e., $L = 3$ in (6). The neighboring relation modeling was also tried. 500 Bi-grams with high-relevant confidence were extracted and incorporated to the BOW. However, for this dataset the proposed neighboring relation modeling of patches does not provide improvement and the AP values go down to about 0.02 lower than the original BOW. The reason is that the variation of keyframes in this set is large and there are only a small amount of positive samples for most concepts; the expected occurrence numbers and confidence values of word pairs cannot be estimated precisely. In the following result, the proposed CBOW only contains the conceptual relation. As shown in the scene categorization experiment, with the dataset containing less variation, the confidence estimation is more precise and the proposed neighboring relation modeling does help.

In Fig. 8, we compare the resulting AP values of Columbia [37], MSRA [22], the BOW, the SPM of two levels, and the

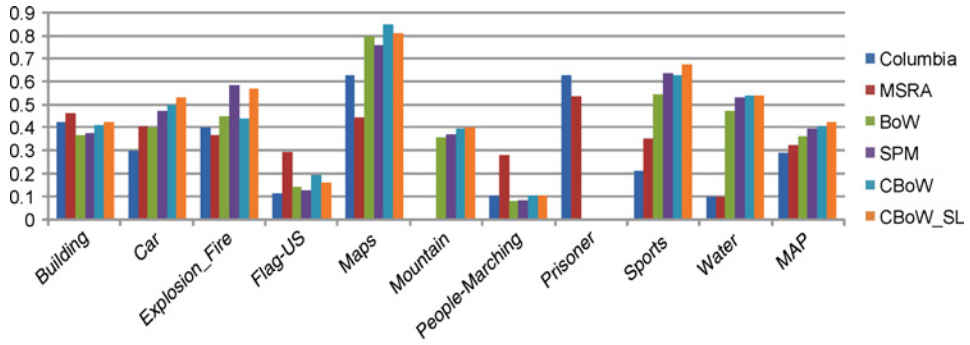

Fig. 8. Comparison of APs of six methods on the TRECVID2005 10 video event concepts.

TABLE VII
PERFORMANCE COMPARISON ON THE TRECVID2005 TEN VIDEO CONCEPTS

Alg.	Columbia	MSRA	BOW	SPM	CBOW	CBOW_SL
MAP	0.2881	0.3214	0.3580	0.3928	0.4026	0.4193

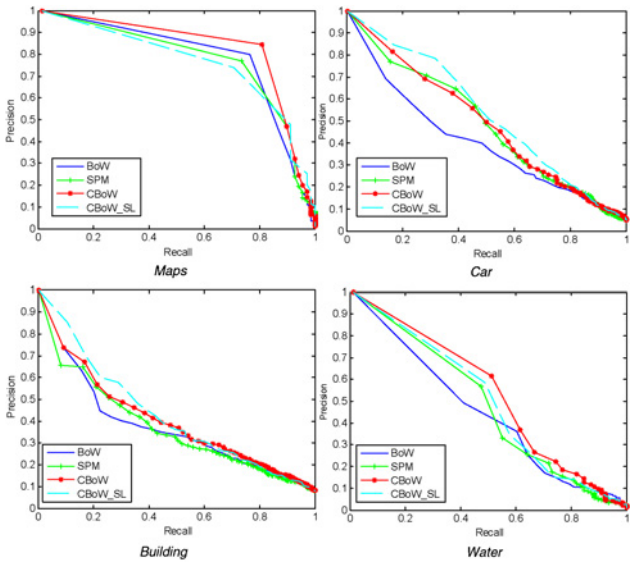

Fig. 9. Precision–Recall curves of four video event concepts.

proposed CBOW. Spatial layout context is also combined to the proposed CBOW to study the context of local patches further, and the result is listed for comparison (CBOW_SL). Table VII lists the mean AP (MAP) values of these methods. Compared to the global feature-based methods of Columbia and MSRA, local feature-based BOW related methods yield better performance for the good representation ability. The contextual relations of CBOW and the spatial layout relation of SPM both yield significant improvements over the BOW while the CBOW is superior. The performance is further improved by introducing the spatial layout information to the CBOW by partitioning keyframes into 2×2 regions and concatenating the representation of all the regions together. Fig. 9 plots the Precision–Recall curves of four example concepts for better observation and comparison. We can see the proposed CBOW or CBOW_SL shows good performance on these concepts from the curves, which coincides with the measurement of the AP values.

From Fig. 8, we can also see that for some specific concepts the global feature-based methods are much better than the proposed method. Thus, they are complementary and could be combined for better performance. It is noticeable that for the concept *Mountain*, the global feature-based methods yield very low APs and for *Prisoner* the local feature-based methods perform poorly. One reason is that the global description of *Mountain* is not discriminative from other concepts while the local characteristics of *Prisoner* can be confused with concepts such as *People_Marching*, etc. Another reason causing the big variation of AP values is the small number of positive test samples of the two concepts, i.e., three for *Prisoner* and 40 for *Mountain* out of more than 6000 test keyframes.

V. CONCLUSION

In this paper, we addressed the problem of the BOW algorithm that local patches context is neglected and proposed a new algorithm, named CBOW, modeling the conceptual context and neighboring context of local patches based on the BOW. First, the measurements for these relations between visual words were introduced. Then visual words were grouped on multiple levels according to the conceptual relation. Images were represented and matched accordingly. Visual words groups, which have informative neighboring relation, were extracted and their statistics were incorporated in the image representation. The proposed method was tested for scene categorization and video event categorization tasks. Experimental results showed that the contextual relations between local patches can be very useful for categorization and the proposed algorithm achieves significant improvement over the original BOW. Furthermore, in experiments the spatial layout context also was combined to extensively study the importance of local patches context for categorization with a high performance being achieved.

ACKNOWLEDGMENT

T. Li would like to thank L. Wu for providing visual language model results, and also grateful to M. Callcut for proof reading the paper.

REFERENCES

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[2] L. D. Baker and A. K. McCallum, "Distributional clustering of words for text classification," in *Proc. Assoc. Comput. Machinery Special Interest Group Informat. Retrieval (SIGIR)*, 1998, pp. 96–103.

[3] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "On feature distributional clustering for text categorization," in *Proc. Assoc. Comput. Machinery Special Interest Group Informat. Retrieval (SIGIR)*, 2001, pp. 146–153.

[4] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2005, pp. 26–33.

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

[6] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[7] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1055–1064, Sep. 1999.

[8] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Statist. Learning Comput. Vision Workshop (SLCV)*, 2004, pp. 1–22.

[9] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: The kernel recipe," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2003, pp. 257–264.

[10] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2005, pp. 524–531.

[11] K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2005, pp. 1458–1465.

[12] M. John, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. Defense Advanced Research Projects Agency (DARPA) Broadcast News Workshop*, 1999, pp. 249–252.

[13] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Beijing, China, 2005, pp. 604–610.

[14] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2004, pp. 506–513.

[15] S. Kullback, "The Kullback–Leibler distance," *Am. Statistician*, vol. 41, no. 4, pp. 340–341, 1987.

[16] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1294–1309, Jul. 2009.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2006, pp. 2169–2678.

[18] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vision Special Issue Learn. Recognit. Recognit. Learn.*, vol. 77, nos. 1–3, pp. 259–289, 2008.

[19] D. Liu and T. Chen, "Discov: A framework for discovering objects in video," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 200–208, Feb. 2008.

[20] J. Liu and M. Shah, "Scene modeling using co-clustering," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2007, pp. 1–7.

[21] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, Y. Liu, Z. Gu, G.-J. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *Proc. Text Retrieval Conf. Video Retrieval Evaluation Online (TRECVID)*, 2007 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2007

[23] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[24] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Proc. Adv. Neural Informat. Process. Syst. (NIPS)*, 2006, pp. 985–992.

[25] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[26] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2006, pp. 2161–2168.

[27] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Graz, Austria, 2006, pp. 490–503.

[28] *Pascal visual object classes* [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/.

[29] *Wordnet* [Online]. Available: http://wordnet.princeton.edu/perl/webwn?s=word-you-want

[30] J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman, "Discovering objects and their locations in images," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Beijing, China, 2005, pp. 370–377.

[31] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2003, pp. 1470–1477.

[32] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2005, pp. 1331–1338.

[33] *Trec Video Retrieval Evaluations (TRECVID)* [Online]. Available: http://www.nlpir.nist.gov/projects/trecvid/

[34] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. IEEE Int. Conf. Comput. Vision*, 2005, pp. 1800–1807.

[35] L. Wu, Y. Hu, M. Li, N. Yu, and X.-S. Hua "Scale-invariant visual language modeling for object categorization," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 286–294, Feb. 2009.

[36] X. Wu, W.-L. Zhao, and C.-W. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," in *Proc. Assoc. Comput. Machinery Int. Conf. Image Video Retrieval (ACM MIR)*, 2007, pp. 162–169.

[37] A. Yanagawa, S. F. Chang, L. Kennedy, and W. Hsu, "Columbia University's baseline detectors for 374 LSCOM semantic visual concepts," Columbia Univ., New York, ADVENT Tech. Rep. 222-2006-8, 2007.

[38] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Assoc. Comput. Machinery Int. Conf. Multimedia Informat. Retrieval (ACM MIR)*, Augsburg, Germany, 2007, pp. 197–206.

[39] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. Machine Learn. (ICML)*, 1997, pp. 412–420.

[40] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[41] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. 17th Assoc. Comput. Machinery Int. Conf. Multimedia (ACM MM)*, Beijing, China, 2009, pp. 75–84.

[42] Q. Zheng, W. Wang, and W. Gao, "Effective and efficient object-based image retrieval using visual phrases," in *Proc. Assoc. Comput. Machinery Int. Conf. Multimedia (ACM MM)*, Santa Barbara, CA, 2006, pp. 77–80.

[43] M. Zhu, "Recall, precision, and average precision," Dept. Statistics Actuarial Sci., Univ. Waterloo, CA, Tech. Rep. 9, 2004.

Teng Li received the B.Eng. degree in automation from the University of Science and Technology of China, Hefei, in 2001, the M.Eng. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2004, and the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, in 2010.

His current research interests include visual categorization, multimedia retrieval, machine learning, and computer vision.

Tao Mei received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, in 2001 and 2006, respectively.

He joined Microsoft Research Asia, Beijing, China, as a Researcher Staff Member in 2006. His current research interests include multimedia content analysis, computer vision, and multimedia applications such as search, advertising, presentation, social networks, and mobile applications. He is the editor of one book, and the author of over 100 journal and conference papers as well as book chapters in these areas. He holds more than 25 filed patents or pending applications.

Dr. Mei serves as an Associate Editor for the *Journal of Multimedia and Neurocomputing*, a Guest Editor for IEEE MULTIMEDIA, *ACM/Springer Multimedia Systems*, and the *Journal of Visual Communication and Image Representation*. He was the Principle Designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He received the Best Paper and Best Demonstration Awards in ACM Multimedia in 2007, the Best Poster Paper Award in IEEE MMSP in 2008, and the Best Paper Award in ACM Multimedia in 2009. He was awarded Microsoft Gold Star in 2010.

In-So Kweon (M'95) received the B.S. and M.S. degrees in mechanical design and production engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the M.S. and Ph.D. degrees in robotics from the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 1986 and 1990, respectively.

He was with the Toshiba Research and Development Center, Kanagawa, Japan, and joined the Department of Automation and Design Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 1992. He is currently a Professor with the Department of Electrical Engineering, KAIST. Specific research topics include invariant-based visions for recognition and assembly, 3-D sensors and range data analysis, color modeling and analysis, robust edge detection, and moving object segmentation, and tracking. His research interests include computer vision, robotics, pattern recognition, and automation.

Dr. Kweon is a Member of the Institute for Computer Applications in Science and Engineering and the Association for Computing Machinery.

Xian-Sheng Hua (M'05) received the B.S. and Ph.D. degrees from Peking University, Beijing, China, in 1996 and 2001, respectively, both in applied mathematics.

Since 2001, he has been with Microsoft Research Asia, Beijing, China, where he is currently a Lead Researcher with the Media Computing Group. He is now an Adjunct Professor of University of Science and Technology of China. He has authored, or co-authored, more than 160 publications in these areas and has more than 40 filed patents or pending applications. When he was with Peking University, his major research interests included the areas of image processing and multimedia watermarking. His current research interests include the areas of video content analysis, multimedia search, management, authoring, sharing, mining, advertising, and mobile multimedia computing.

Dr. Hua serves as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, Associate Editor of the *Association for Computing Machinery (ACM) Transactions on Intelligent Systems and Technology*, an Editorial Board Member of the *Advances in Multimedia and Multimedia Tools and Applications*, and an Editor of the *Scholarpedia (Multimedia Category)*. He won the Best Paper Award and the Best Demonstration Award in the ACM Multimedia in 2007, the Best Poster Paper Award in the 2008 IEEE International Workshop on Multimedia Signal Processing. He also won the 2008 MIT Technology Review TR35 Young Innovator Award, and was named as one of the "Business Elites of People under 40 to Watch" by Global Entrepreneur. He is a Senior Member of the Association for Computing Machinery.