Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study

Yu-Gang Jiang, Jun Yang, Chong-Wah Ngo*, *Member, IEEE*, Alexander G. Hauptmann, *Member, IEEE*

***Abstract*—Based on the local keypoints extracted as salient image patches, an image can be described as a "bag-of-visual-words (BoW)" and this representation has appeared promising for object and scene classification. The performance of BoW features in semantic concept detection for large-scale multimedia databases is subject to various representation choices. In this paper, we conduct a comprehensive study on the representation choices of BoW, including vocabulary size, weighting scheme, stop word removal, feature selection, spatial information, and visual bi-gram. We offer practical insights in how to optimize the performance of BoW by choosing appropriate representation choices. For the weighting scheme, we elaborate a soft-weighting method to assess the significance of a visual word to an image. We experimentally show that the soft-weighting outperforms other popular weighting schemes such as TF-IDF with a large margin. Our extensive experiments on TRECVID data sets also indicate that BoW feature alone, with appropriate representation choices, already produces highly competitive concept detection performance. Based on our empirical findings, we further apply our method to detect a large set of 374 semantic concepts. The detectors, as well as the features and detection scores on several recent benchmark data sets, are released to the multimedia community.**

***Index Terms* — Bag-of-visual-words, representation choice, semantic concept detection.**

I. Introduction

Semantic concept detection is a research topic of great interest as it provides semantic filters to help analysis and search of multimedia data. It is essentially a classification task that determines whether an image or a video shot is relevant to a given semantic concept. The semantic concepts cover a wide range of topics such as those related to objects (e.g., *car, airplane*), indoor/outdoor scenes (e.g., *meeting, desert*), events (e.g., *people_marching*), etc. Automatically detecting these concepts is challenging especially in the presence of within-class variation, occlusion, background clutter, pose and lighting changes in images and video shots. Global features are known to be limited in face of these difficulties, which stimulated the development of local invariant features (keypoints) in recent years. Keypoints are salient patches that contain rich local information about an image. The most popular keypoint-based representation is bag-of-visual-words (BoW) [1]. In BoW, a visual vocabulary is generated through grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. By mapping the keypoints of an image back into visual words of the vocabulary, we can represent the image as a histogram of visual words and use it as the feature for classification.

The BoW image representation is analogous to the bag-of-words representation of text documents in terms of both form and semantics. This makes techniques for text categorization readily applicable to the problem of semantic concept detection. As it is true with text categorization, where feature representation has a large impact on its performance, the performance of semantic concept detection is also sensitive to various representation choices. In this paper, we conduct a comprehensive study on the representation choices of BoW feature and their impact to the performance of semantic concept detection. Some of the representation choices are related to text categorization techniques, including word weighting scheme, stop word removal, feature selection, and visual bi-gram, while the others are unique to concept detection in images and videos, including vocabulary size (number of keypoint clusters) and spatial information of the keypoints. Particularly, for the visual word weighting scheme, we provide in-depth analysis of a soft-weighting method, which was initially proposed in our earlier work [2]. We generate BoW features based on different representation choices and evaluate their performance in large scale concept detection experiments. Besides, we also study the choice of kernel functions used in the classification of BoW features.

This study fills the gap in the existing works on image classification based on local features, where most of the effort focused on various keypoint detectors, keypoint descriptors and clustering algorithms [1], [3], [4], [5], [6]. Few have paid attention to various representation choices regarding this visual-word feature (e.g., feature selection and weighting) and studied their impacts on the classification performance. Although some researchers adopted techniques like TF-IDF weighting and stop word removal [1], the effectiveness of these techniques have been taken for granted without empirical evidence. In addition, most existing evaluations of methods using local features have also been of small scale. This paper provides the first comprehensive study on the representation

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 118906), and a grant from City University of Hong Kong (Project No. 7002241).

Y.-G. Jiang and C.-W. Ngo are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: cwngo@cs.cityu.edu.hk). Y.-G. Jiang is also with the Department of Electrical Engineering, Columbia University, New York, NY 10027, USA (e-mail: yjiang@ee.columbia.edu).

J. Yang is with Google Inc., Mountain View, CA 94043, USA (e-mail: yangjun@google.com).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA (e-mail: alex@cs.cmu.edu).

* Corresponding author.

choices of visual-word features in semantic concept detection. By evaluating various representation choices, we intend to answer the question of what BoW representation choices (w.r.t dimension, weighting, selection, etc) are likely to produce the best performance in terms of both accuracy and efficiency.

We evaluate semantic concept detection performance based on various visual-word representations on TRECVID data sets. Our experiments indicate that, with carefully chosen representation choices, BoW feature offers very competitive performance. Additionally, to stimulate innovation of new techniques and reduce the re-implementation effort of our approach, we apply our method to detect a large set of 374 LSCOM [7] semantic concepts, namely VIREO-374. We release our features, classifier models, and detection scores on several popular data sets. Compared to existing large-scale concept detector sets [8], [9], VIREO-374 is better in terms of scale and/or performance.

The remaining sections are organized as follows. Section II reviews the existing works. Section III describes the generation process of BoW image representation. Section IV outlines the representation choices of BoW feature, including vocabulary size, weighting scheme, stop word removal, feature selection, spatial information and visual bi-gram. Section V introduces the kernel choices of BoW classification. Section VI evaluates the representation and kernel choices on TRECVID 2006 data set. Section VII further discusses the generalizability of our empirical findings to other popular data sets and extends our method to detect a large set of 374 concepts. Finally, Section VIII concludes this paper.

II. Related Works

Semantic concept detection aims to annotate images or video shots with respect to a semantic concept. In existing works, this task is often conducted in a diverse setting where the emphasis usually includes feature selection, multi-modality fusion, and machine learning on huge multimedia data sets [10]. Here we focus our review on feature-level analysis which is related to our latter experimental comparison. In [11], rich sets of features (visual, motion, text, face) and classifiers were demonstrated to have excellent performance on concept detection. Visual features, in particular, were extracted simultaneously from global, grid, region and keypoints levels, activating more than 100 SVM classifiers for learning a single concept. While technically impressive, it becomes expensive to scale up such a system, for instance, when thousands of semantic concepts are considered for detection. Meanwhile, the approaches in [6], [12], [13], [14], [15], [16] used less features but still shown comparable performance to that of [11]. The features include color and texture (in global and grid levels), motion, text, etc. BoW is also used in [2], [17], [12], [14], [15]. Specifically, [12], [14] adopted single type of keypoint and the SIFT descriptor [18], while [6], [15], [16] used a combination of different keypoint sampling methods (including sparse detectors such as Harris Laplace and Boosted ColorHarris Laplace, as well as dense sampling) and keypoint descriptors (SIFT, HueSIFT, and etc). The ColorHarris Laplace and HueSIFT are constructed by integrating color information into Harris Laplace and SIFT respectively [19]. Improvements of the color boosted features over the traditional ones were observed in [6], [15], [16].

In addition, [20] also used local feature for semantic concept detection, but in a different way. They adopted geometric blur features [21] as keypoint descriptor. The geometric blur features were computed based on 200 randomly sampled points with high edge energy from a keyframe. A total of 1291 training example keyframes are picked as references. Given a test keyframe, online point-to-point matching was performed between the keyframe and the exemplars. Each keyframe was then represented as a 1291 dimensional vector with each component indicating the distance of the keyframe to a reference. The feature vectors were used directly for SVM learning. In this representation, for each keypoint in the test keyframe, the number of keypoint comparisons is as high as 1291×200. This is computationally more expensive than the BoW representation where the number of comparison for each test keypoint is equal to the number of visual words used (usually a few thousands; cf. Section IV-A).

In computer vision, BoW has already exhibited surprisingly good performance for object retrieval and categorization across several data sets (e.g., [2], [3], [4], [5], [22], [23] among others). In our recent work [2], a study on keypoint detectors, feature weighting and vocabulary size was given. In [5], Zhang et al. conducted a comprehensive study on the local feature based object and texture classification. They provided comparisons on the choice of a few keypoint detectors and proposed to use χ^2 RBF kernel for SVM learning. In [4], Nowak et al. studied the sampling strategies of BoW to compare dense (grid-based local image patches) and sparse (keypoints) representation. They claimed that sample size is critical for building vocabularies and thus the randomly sampled image patches could offer a more powerful representation than the sparse keypoints. In [22], Grauman et al. proposed to use pyramid matching kernel (PMK) for image comparison based on local keypoint features. The orderless keypoint feature sets were mapped to multi-resolution histograms and weighted histogram intersection was used as kernel response. In [3], Lazebnik et al. exploited the spatial location of keypoints and proposed a spatial pyramid kernel, in which an image was firstly divided into multi-level equal-sized grids and each grid was described by a separate BoW. The BoWs from image grids at each level were concatenated and finally, similar to PMK, the weighted histogram intersection was used as kernel response. Recently, in both [23] and [24], the effects of soft and hard weighting schemes in generating BoW features for object retrieval are contrasted.

In this paper, we assess and improve the performance of BoW for semantic concept detection in large-scale multimedia corpus, extending our previous works [2], [17] with results on two more recent data sets, ample result analysis, and an extension to detect 374 semantic concepts. Different from [3], [4], [5], [6], [14], [15], [16], [22], [23], we first separately and then jointly consider various representation choices such as feature weighting, vocabulary size, feature selection and visual bi-gram, which could govern the BoW performance but have not yet been seriously studied in other works.

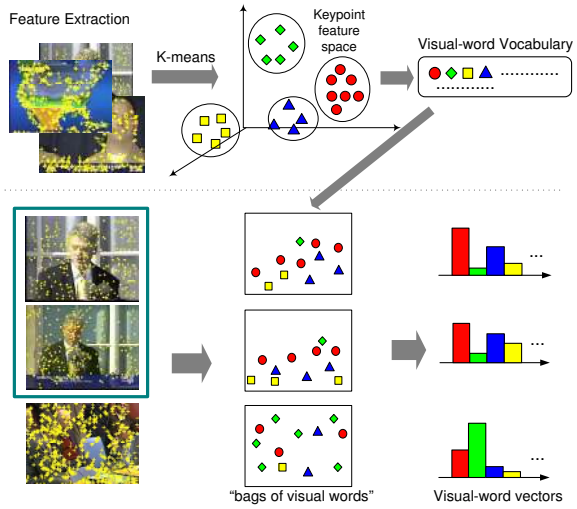

Fig. 1. Image representation using bag-of-visual-words.

III. Bag-of-Visual-Words (BoW) Feature

Similar to terms in a text document, an image has local interest points or keypoints defined as salient patches that contain rich local information about the image. Shown as small crosses in the images on the left of Figure 1, keypoints are usually around the corners and edges of image objects, such as the edges of the map and people's faces, etc. Keypoints can be automatically detected by various detectors [25] and described by different descriptors [26].

Images can be represented by sets of keypoint descriptors, but the sets vary in cardinality and lack meaningful ordering. This creates difficulties for learning methods (e.g., classifiers) which usually demand feature vectors of fixed dimension as input. To address this problem, we adopt vector quantization (VQ) technique to cluster the keypoint descriptors in their feature space into a large number of clusters using the k-means clustering algorithm, and then encodes each keypoint by the index of the cluster to which it belongs. We conceive each cluster as a *visual word* that represents a specific local pattern shared by the keypoints in that cluster. The clustering process generates a *visual word vocabulary* describing different local patterns. The number of clusters is the size of the vocabulary, which usually varies from hundreds to over tens of thousands. Mapping the keypoints to the visual words, we can represent an image as a bag-of-visual-words (BoW). This representation is analogous to the bag-of-words document representation in terms of form and semantics. Both representations are sparse and high-dimensional, and just as words convey meanings of a document, visual words reveal local patterns characteristics of the whole image.

The BoW representation can be converted into a visual word vector, which is similar to the term vector of a document. This visual word vector may contain the presence/absence of each visual word in the image, the count of each visual word (i.e., the number of keypoints in the corresponding cluster), or weights of each visual word by other factors (see section IV-B). This visual word vector is used in classifying the semantic concepts. The process of generating visual word representation is illustrated in Figure 1.

IV. Representation Choices

This section introduces various factors that can affect the performance of BoW feature for semantic concept detection. Some are widely used in text categorization, such as term weighting, stop word removal, feature selection, and bi-grams (word co-occurrence), while others are unique to images, such as changing the vocabulary size and encoding the spatial information. We discuss these techniques below.

A. Vocabulary Size

Since the visual words are generated by clustering local keypoint features, the size of a visual vocabulary is controlled by the number of keypoint clusters in the clustering process. This is different from the vocabulary of a text corpus whose size is relatively fixed. A small vocabulary may lack the discriminative power since two keypoints may be assigned into the same cluster even if they are not similar to each other. A large vocabulary, on the other hand, is less generalizable, less forgiving to noises, and incurs extra processing overhead.

The trade-off between discrimination and generalization motivates the study of visual vocabulary size. Our survey shows that previous works used a wide range of vocabulary sizes, leading to difficulties in interpreting their findings. For instance, Lazebnik et al. [3] adopted 200-400 visual words, Zhang et al. [5] adopted 1,000, Sivic et al. [1] adopted 6,000 -10,000, Philbin et al. [23] adopted as high as 1 million, etc. In our study, we experiment with vocabularies of various numbers of visual words.

B. Weighting Schemes

Term weighting is known to have a critical impact on text information retrieval (IR). Whether such impact extends to visual keywords is an interesting question. A fundamental difference is that: text words are natural entities in a language context, while visual words are the outcomes of feature clustering. The former carries semantic sense of natural language, while the latter infers statistical information of repetitive local image patterns. The existing work on BoW mostly adopted conventional weighting schemes in IR, which are based on term frequency (TF) and/or inverse document frequency (IDF). In [1], Sivic et al. adopted TF-IDF, while most of the other works chose TF directly [3], [5]. In [4], binary weighting, which indicates the presence and absence of a visual word with values 1 and 0 respectively, was used.

All these weighting schemes perform the nearest neighbor search in the vocabulary in the sense that each keypoint is mapped to the most similar visual word (i.e., the nearest cluster centroid). For visual words, however, assigning a keypoint only to its nearest neighbor is not an optimal choice, given the fact that two similar points may be clustered into different clusters when increasing the size of visual vocabulary. Moreover, simply counting the votes (e.g., TF) is not optimal as well. For instance, two keypoints assigned to the same visual word are not necessarily equally similar to that visual word, i.e., their distances to the cluster centroid are different. Ignoring their similarity with the visual word during weight

assignment causes the contribution of two keypoints equal, and thus it becomes more difficult to assess the importance of a visual word in an image.

In order to tackle the aforementioned problems, in our earlier work [2], we proposed a *soft-weighting* scheme to weight the significance of visual words. For each keypoint in an image, instead of mapping it only to its nearest visual word, in soft-weighting we select the top-N nearest visual words. Suppose we have a visual vocabulary of K visual words, we use a K-dimensional vector $\mathbf{w} = [w_1, ..., w_t, ..., w_K]$ with each component w_t representing the weight of a visual word t in an image such that

$$w_t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j,t), \quad (1)$$

where M_i represents the number of keypoints whose *i*th nearest neighbor is the visual word *t*. The measure $sim(j,t)$ represents the Cosine similarity between keypoint *j* and the visual word *t*. Notice that in Eqn 1 the contribution of a keypoint is its similarity to word *k* weighted by $\frac{1}{2^{i-1}}$, representing that the visual word is its *i*th nearest neighbor.

C. Stop Word Removal

Stop word removal is a standard technique in text categorization. The question is, are there also "visual stop words" that represent local patterns totally redundant for image retrieval and classification? Sivic and Zisserman [1] claimed that the most frequent visual words in images are also "stop words" and need to be removed from the feature representation. There is however no empirical evidence that shows doing that improves image classification performance. Since it is very difficult to judge whether a visual word is a stop word, we focus on the relationship between frequent visual words as a group and the classification performance.

D. Feature Selection

Feature selection is an important technique in text categorization for reducing the vocabulary size and consequently the feature dimension. It uses a specific criterion for measuring the "informativeness" of each word and eliminates the non-informative words. Yang et al. [27] found out that, when a good criterion is used, over 90% of the unique words in the vocabulary can be removed without loss of text categorization accuracy. In semantic concept detection of images and videos, feature selection is potentially important as the size of the visual-word vocabulary is usually very high, but it has not been seen in any existing work. We experiment with five feature selection criteria used in text categorization [27]:

- ***document frequency (DF)***: *DF* is the number of images (documents) in which a visual word (word) appears. In text categorization, words with small *DF* are removed since rare words are usually non-informative for category prediction. Not knowing whether frequent visual words or rare ones are more informative, we adopt two opposite selection criteria based on *DF*: *DF_max* removes rare words by choosing visual words with *DF* above a predefined threshold, while *DF_min* removes frequent words by choosing visual words with *DF* below a threshold.
- χ^2 ***statistics (CHI)***: The χ^2 statistics measures the level of (in)dependence between two random variables [27]. Here we compute $\chi^2(t, c_i)$ between the presence/absence of a specific visual word t and the binary label of an image class c_i. A large value of $\chi^2(t, c_i)$ indicates a strong correlation between t and c_i, and vice versa. Since $\chi^2(t, c_i)$ depends on a specific class, we compute the average statistics across all the image classes as $\chi^2_{avg}(t) = \frac{1}{C}\sum_{i=1}^{C} \chi^2(t, c_i)$, where C is the number of classes in the corpus. We then eliminate visual words with $\chi^2_{avg}(t)$ below a threshold.
- ***Information gain (IG)***: *IG* is another measure of the dependence between two random variables. The *IG* between a visual word t and a class label c_i is computed as:

$$IG(t, c_i) = \sum_{t \in \{0,1\}} \sum_{c_i \in \{0,1\}} P(t, c_i) \log \frac{P(t, c_i)}{P(t)P(c_i)}. \quad (2)$$

  We compute $IG_{avg}(t) = \frac{1}{C}\sum_{i=1}^{C} IG(t, c_i)$, and remove visual words with $IG_{avg}(t)$ below a threshold.
- ***Mutual information (MI)***: *MI* is related to *IG*. It uses one term in the sum of Eqn 2 to measure the association between a visual word t and a class label c_i:

$$MI(t, c_i) = \log \frac{P(t = 1, c_i = 1)}{P(t = 1)P(c_i = 1)}. \quad (3)$$

  Similar to *CHI* and *IG*, visual words with small $MI_{avg}(t)$ are eliminated from the vocabulary.

E. Spatial Information

Where within a text document a certain word appears is usually not very relevant to the category of this document. The spatial locations of keypoints in an image, however, carry important information for classifying the image. For example, an image showing a beach scene typically consists of sky-like keypoints on the top and sands-like keypoints at the bottom. The plain BoW representation described in Section III ignores such spatial information and may result in inferior classification performance. To integrate the spatial information, we follow [3] to partition an image into equal-sized rectangular regions, compute the visual-word feature from each region, and concatenate the features of these regions into a single feature vector. There can be many ways of partitioning, e.g., 3×3 means cutting an image into 9 regions.

This region-based representation has its downside in terms of cost and generalizability. First, if we divide each image into $m \times n$ regions, and compute a K-dimensional feature on each region, the concatenated feature vector is of $K \times m \times n$ dimension, which can be prohibitively expensive to deal with. Besides, encoding spatial information can make the representation less generalizable. Suppose an image class is defined by the presence of a certain object, say, *airplane*, which may appear anywhere in an image. Using region-based representation can cause a feature mismatch if the objects in the training images are in different regions from those in

the testing images. Another risk is that many objects may cross region boundaries. These considerations prefer relatively coarse partitions of image regions to fine-grained partitions.

F. Visual Bi-gram

Besides the location of individual visual words, the spatial proximity of different visual words is also important for classification because it captures the geometrical structure of an image. For example, visual words depicting *face* may frequently co-occur with visual words characterizing *necktie*. The spatial co-occurrence of visual words is analogous to the bi-grams or n-grams in text categorization [28], [29]. Because the keypoints are sparsely distributed in an image and are not necessarily adjacent to each other in our representation, we name it as sparse visual bi-gram.

We use a two-dimensional co-occurrence histogram to represent an image based on the visual bi-grams. Suppose there are K visual words, a $K \times K$ matrix (2-dimensional histogram) G_r is constructed with each entry $G_r(s,t)$ indicating the frequency of visual bi-gram $\{s,t\}$ appearing with $d(s,t) \leq r$, where $d(\cdot)$ is the Euclidean distance of the two words s and t in the image and r is a threshold. Multiple histograms with various r can be used to capture the visual bi-grams of different word distances.

The visual bi-gram offers a perspective of modeling the spatial co-occurrence of visual words. Similar works include recent studies of Lazebnik et al. [30] and Nowozin et al. [31]. The former used semi-local parts (groups of neighboring keypoints) for texture and object recognition, while the latter attempted to mine significant spatial co-occurrent visual word patterns for object categorization. By using the most informative visual word patterns, better categorization performance is observed in [31].

V. Kernel Choices of BoW Classification

Once images are represented by BoW features, we can classify images in the same way we classify text documents. The general approach is to build supervised classifiers from labeled images based on BoW features and apply them to predict the labels of other images.

In our experiments, we adopt Support Vector Machines (SVM) for semantic concept detection. SVM has been one of the most popular classifiers for BoW-based image classification [2], [3], [4], [5], [6], [11]. For two-class SVM, the decision function for a test sample x has the following form:

$$g(x) = \sum_i \alpha_i y_i \mathcal{K}(x_i, x) - b, \tag{4}$$

where $\mathcal{K}(x_i, x)$ is the response of a kernel function for the training sample x_i and the test sample x, which measures the similarity between the two data samples; y_i is the class label of x_i; α_i is the learned weight of the training sample x_i, and b is a learned threshold parameter.

The choice of an appropriate kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ is critical to the classification performance. $\mathcal{K}$ should be positive definite and symmetric (a.k.a. Mercer's condition), to guarantee the convergence of SVM training. Although there are a number of general-purpose kernel functions, it is unclear which one is the most effective for BoW features in the context of semantic concept detection. In [22], histogram intersection is implicitly used in the proposed pyramid match kernel. In [5], Zhang et al. adopted the χ^2 RBF kernel which has shown good performance, while the authors of many other existing works, to our knowledge, chose the traditional linear kernel or Gaussian RBF kernel. In this paper, we will evaluate the following kernels for BoW-based visual classification:

- ***Linear kernel***:
$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}, \tag{5}$$
where $\mathbf{x}$ and $\mathbf{y}$ are two input vectors.
- ***Histogram intersection kernel***: The Histogram Intersection kernel was proposed and proven to be Mercer kernel in [32]:
$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \sum_i \min\{x_i, y_i\}, \tag{6}$$
- ***Generalized forms of RBF kernels***:
$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\rho d(\mathbf{x}, \mathbf{y})}, \tag{7}$$
where $d(\mathbf{x}, \mathbf{y})$ can be chosen to be any distance in the feature space. Since BoW is a histogram of visual words with discrete densities, the χ^2 distance may be more appropriate:
$$d_{\chi^2}(\mathbf{x}, \mathbf{y}) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}, \tag{8}$$
which gives a χ^2 RBF kernel. The χ^2 RBF kernel satisfies Mercer's condition [33].
In addition to χ^2, there are another series of generalized RBF kernels with the distance function defined as:
$$d_b(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|^b. \tag{9}$$
With this distance function, Eqn 7 becomes the Laplacian RBF kernel when $b = 1$ and the sub-linear RBF kernel when $b = 0.5$. These kernels are popularly used in image retrieval with color histogram as feature, and have shown to generate better performance than Gaussian RBF kernel ($b = 2$) [34]. The functions $e^{-\rho d_b(\mathbf{x}, \mathbf{y})}$ satisfy Mercer's condition if and only if $0 \leq b \leq 2$ [35].

VI. Empirical Study

In this section we conduct extensive experiments to evaluate the choices of BoW representations and classification kernels.

A. Experimental Setup

1) Data set: We use TRECVID 2006 data set to empirically study the choices described in the previous sections. The data set was used for TREC Video Retrieval Evaluation 2006 [36], where the training and testing sets consist of 61,901 and 79,484 video shots respectively. One video frame is extracted from each shot as its keyframe. In the experiments, we use the 20 semantic concepts which were officially evaluated in TRECVID 2006. The labels of these concepts in the training set are provided by LSCOM [7]. Figure 2 shows keyframe

Fig. 2. Keyframe examples of 20 semantic categories in TRECVID 2006 data set.

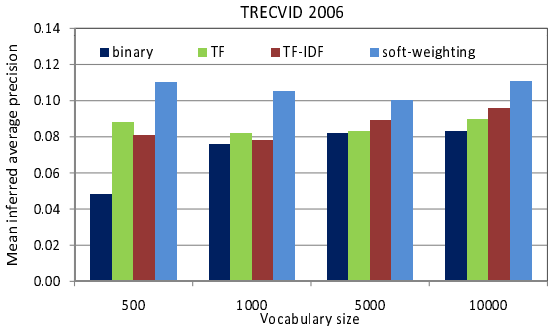

Fig. 3. The MinfAP of concept detection on TRECVID 2006 using different weighting schemes and vocabulary sizes. It can be clearly seen that soft-weighting produces consistently better performance than the other weighting schemes for all the vocabulary sizes.

examples of the 20 semantic concepts. These concepts cover a wide variety of topics, including objects, indoor/outdoor scenes, people, events, etc. The goal of concept detection is to rank the 79,484 video keyframes according to the presence of each of the 20 semantic concepts. Note that this data set is a multi-label data set, which means each keyframe may belong to multiple classes or none of the classes (concepts), e.g., the example of *weather* in Figure 2 also belongs to concept *map*.

2) BoW generation: The keypoints are detected by DoG (Difference of Gaussian) detector [18] and described by SIFT descriptor [18]. This results in an average of 235 keypoints per keyframe. In the experiments, we use k-means clustering algorithm to generate visual vocabularies. To reduce the computational cost, we sample the training set and cluster 550,000 SIFT features. While in the BoW representation there is an issue of data dependent vocabulary versus universal vocabulary, we will not elaborate this challenging question due to space limitation. The parameters N in the soft-weighting scheme and the parameter r in the visual bi-gram generation are empirically chosen as 4 and 40 respectively.

The classification is conducted independently for each concept. Using the SVM, we build 20 binary classifiers for the 20 semantic concepts, where each classifier is for determining the presence of one specific concept.

3) Evaluation criteria: We use inferred average precision (infAP) for performance evaluation. The infAP is an approximation of the conventional average precision (AP). The main advantage of the infAP is that it can save significant judging effort during the annotation of ground-truth for large testing data set [36]. Following the TRECVID evaluation, the infAP is computed over the top 2,000 ranked shots according to the outputs of the SVM classifiers. To aggregate the performance of multiple semantic concepts, mean infAP (MinfAP) is used.

B. Weighting Schemes and Vocabulary Sizes

In this section, we examine the keyword weighting schemes, vocabulary sizes, and their impact on classification performance. We use the χ^2 RBF kernel for SVM learning. The observations from the other kernel choices are similar. The results are summarized in Figure 3.

First, let us evaluate the influence of different weighting schemes. Our soft-weighting outperforms the other popular weighting schemes across different vocabulary sizes. This indicates that the visual words are indeed correlated to each other and such correlation needs to be considered in feature representation. For that reason, our soft-weighting method which is tailored for the weighting of visual words performs much better. Next, we move on the see the relationship between *binary* and TF. We see that TF outperforms *binary* by a large margin only when the vocabulary size is small. This is due to the fact that, with a larger vocabulary size, the count of most visual keywords is either 0 or 1 and thus TF features are similar with *binary* features.

The IDF, which weighs visual words according to their distribution among the images, is only slightly helpful in some of our experiments. We observe that the impact of IDF is sensitive to the vocabulary size. This is not surprising because a frequent visual word (cluster) may be split into several rare words (clusters) when increasing the vocabulary size. Thus the IDF weight of a certain keypoint is not stable at all.

Finally, let us examine the impact of different vocabulary sizes. When using *binary* weighting, we observe that an appropriate size of vocabulary is 10,000 or larger. However, when more sophisticated weighting schemes are employed, the impact of vocabulary size turns to be less significant. Less sensitive to vocabulary size is an important merit for a weighting scheme, since using small vocabulary size reduces the computational time in both the vector quantization and the classification processes. The MinfAP performance of the soft-weighting scheme over different vocabulary sizes (500-10,000) varies just in a small range of 0.01, while the performance of *binary* weighting changes for almost 0.04. The small performance fluctuation of the soft-weighting scheme is probably due to the use of k-means algorithm which is sensitive to the initial selection of cluster centers.

C. Stop Word Removal

Do the most frequent visual words behave like "stop words"? We approach this question by examining the classification performance using pruned vocabularies with the most frequent visual words removed. We use the 10,000-d vocabulary, which produces the best performance in the

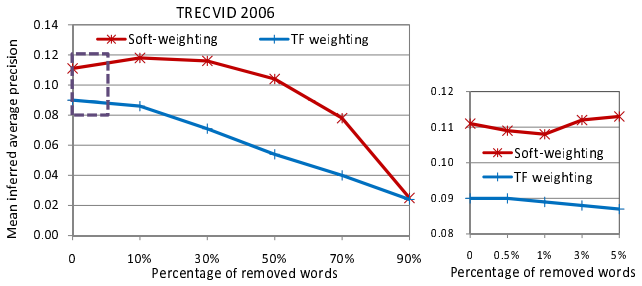

Fig. 4. Concept detection performance on TRECVID 2006 with stop word removal. For soft-weighting, removing a moderate amount of frequent words (30%) does not hurt the performance. The smaller figure on the right enlarges the part within the dotted bounding box.

last experiment. As shown in Figure 4, for soft-weighting, removing up to 30% of the words basically does not hurt the performance. But after that, the performance drops at a faster rate. While for TF weighting, the performance always decreases, and the degree of degradation is basically linear to the proportion of removed words. From this experiment we have two observations. First, the most frequent words are indeed not that informative and we can remove some of them without hurting the performance. However, it is still premature to say that they are all stop words, as reserving them will not hurt the performance as well, which is different from text retrieval where stop words hurt performance. Second, the soft-weighting is more robust than TF when pruning more words. This is probably due to the fact that soft-weighting assigns a keypoint to multiple words, which can increase the discriminative power of the remaining words.

D. Feature Selection

In this section we examine the five feature selection criteria discussed in Section IV-D, which are *DF-max*, *DF-min*, *CHI*, *IG*, and *MI*. We reduce the vocabulary size by removing the most uninformative words determined by each criterion, and evaluate the concept detection performance. Results based on the 10,000-d vocabulary are shown in Figure 5.

We see that when effective criteria like *IG* and *CHI* are used, there is only a minimum loss of performance when the vocabulary is cut by 50%. It is interesting to see that even when the vocabulary is cut by as high as 90% (retain 1,000 words), the performance drops 45% (soft-weighting). However, as shown in Figure 3, using a small vocabulary of 1,000 visual words without selection still achieves very good performance. Thus we conclude that a reduction of up to 50% can be carried out using feature selection, but for larger reductions, the performance may not be better than directly constructing a smaller vocabulary. As a comparison, in text categorization a vocabulary can be reduced by 90% or even 98% without loss of classification accuracy [27], which implies that the percentage of uninformative (noisy) terms in text may be larger than that in images.

Among different feature selection methods, *CHI* and *IG* are obviously top performers, followed by *DF_max*, while the performances of *DF_min* and *MI* are lower than the

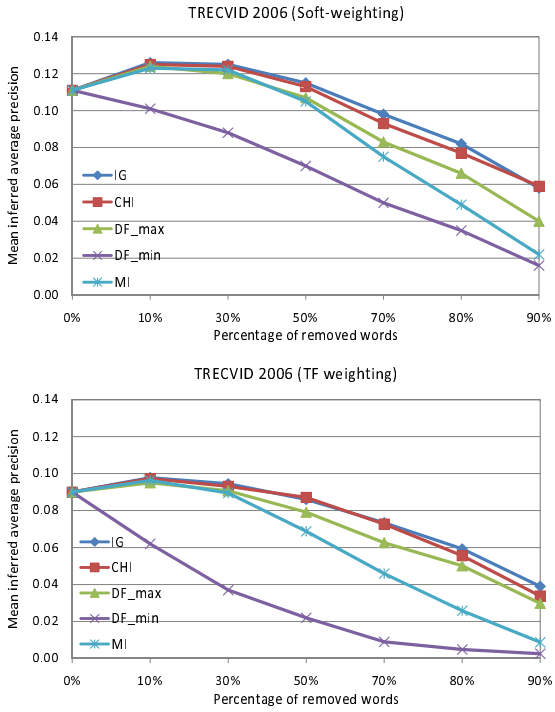

Fig. 5. Concept detection performance on TRECVID 2006 using visual vocabularies pruned using various feature selection criteria. Under both TF weighting and soft-weighting, as high as 50% of the visual words can be removed with very minor performance degradation when criteria such as *IG* and *CHI* are used.

others. This order is basically consistent with that in the text categorization [27].

E. Spatial Information

The importance of spatial information can be seen by comparing the classification performance between the plain visual-word features (BoW) and the region-based ones. We examine four ways of partitioning images, including 1×1 (whole image), 2×2 (4 regions), 3×3 (9 regions), and 4×4 (16 regions). Figure 6 shows the performance using different spatial partitions, vocabulary sizes, and weighting schemes.

We see that the 2×2 partition substantially improves the classification performance. As the partition changes from 2×2 to 4×4, the MinfAP drops for most of the vocabulary sizes. This can be explained based on our discussions in section IV-E that using more regions will make the representation less generalizable and may cause the feature mismatch problem. When investigating the per-concept performance, we find that spatial information is more useful for classifying scenes than for classifying objects, since the former usually occupy a whole keyframe, while the latter can appear anywhere in a keyframe. For large scale semantic detection in diversified data set, using 2×2 partition might be enough. Our conclusion is a bit different from the results of scene and object categorization in [3] where 8×8 regions are still useful. This is probably due to the fact that the objects in the data set they used (Caltech-101) are centered, for which spatial information always helps.

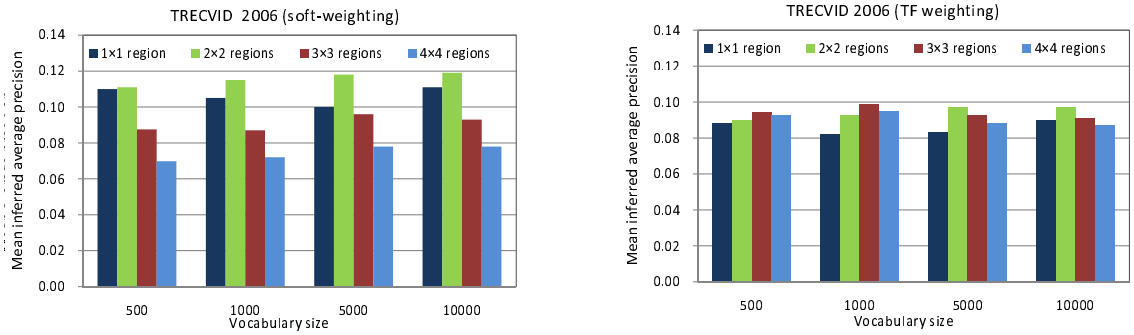

Fig. 6. Concept detection performance on TRECVID 2006 using region-based features computed from different spatial partitions. Due to the feature mismatch problem caused by spatial partition, relatively coarse region partition such as 2×2 is preferred.

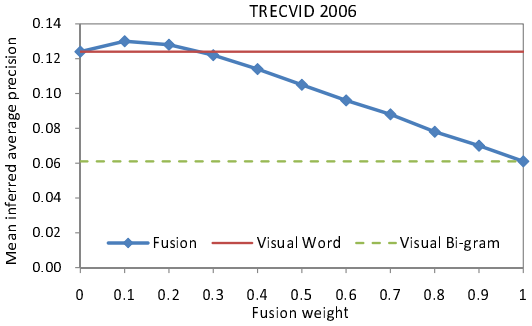

Fig. 7. Performance on TRECVID 2006 using fusion of visual bi-grams and visual words. With good choice of fusion parameter, visual bi-grams are able to improve the detection performance by 5%.

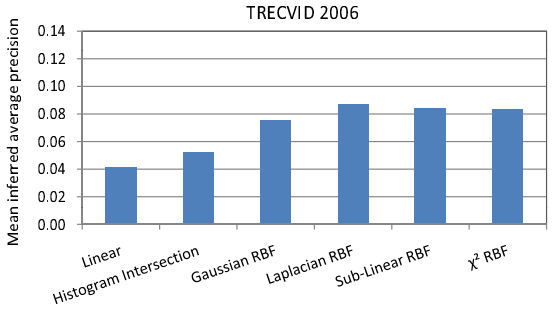

Fig. 8. Concept detection performance on TRECVID 2006 using SVM with different *kernels*. RBF kernels with linear exponential decay (χ^2 and Laplacian) are the most suitable choices for BoW classification.

F. Visual Bi-gram

In this section we examine the effectiveness of the visual bi-gram. As introduced in section IV-F, there are in total $K \times K$ possible bi-grams in a vocabulary of K visual words. It is risky to concatenate the visual bi-grams with the original BoW feature into a single feature vector, as the large number of bi-grams may overwhelm the visual words. Instead, we build a separate SVM model based on visual bi-grams. The combination of visual bi-grams and visual words is done by 'late fusion', i.e., the final decision is made by fusing of the outputs of separate classifiers. While the raw output of SVM in Eqn 4 can be used as a detector response, we prefer the Platt's method [37], [38] to convert the raw output into a posterior probability. This is more reasonable especially for the fusion of multiple feature modalities, since the raw outputs of SVM for different modalities may be in different scales, which will make the feature with larger scale dominating the others. In this experiment, we use linear weighted fusion defined as $\lambda \times p_{bi-gram}(x) + (1-\lambda) \times p_{word}(x)$, where $p(x)$ is the probability output of SVM for test sample x.

We fuse the result of visual bi-gram with the best visual word based result (10,000 words with 30% of them removed by CHI). Figure 7 shows the fusion performance with various λ. We see that the MinfAP of visual bi-gram alone is 0.06, and the fusion with visual word can improve the performance by 5% when $\lambda = 0.1$. The improvement of using visual bi-grams is consistent with that in text categorization where bi-grams can improve the performance by about 10% or less [29]. Thus we may conclude that the visual bi-grams describing the geometric structure of an image are useful for semantic detection. It can be used as a complement to the visual word features, but careful selection of the fusion parameter (e.g., using cross validation) is necessary.

G. Kernel Choice

In this experiment, we investigate the impact of different kernels in SVM on BoW-based concept detection performance. We use TF weighting on the vocabulary with five thousands visual words. Figure 8 summarizes the performances of various kernels. The results of other weighting schemes and vocabulary sizes are similar. For the generalized RBF kernels, we vary the parameter ρ in a reasonable range and choose the best one via cross validation. Overall, the generalized RBF kernels perform better than Linear kernel and histogram intersection kernel with non-trivial margin. This indicates that the semantic classes are correlated to each other in BoW feature space and thus are not linearly separable.

Among all the generalized RBF kernels, the χ^2 RBF kernel, Laplacian RBF kernel, and sub-linear RBF kernel consistently outperform the traditional Gaussian RBF kernel. This can be attributed to the responses of the kernels to background variance. Ideally, a kernel should only emphasize regions containing the target concept, while tolerating the background variance without amplifying the effect. Take Figure 9 as an example. It is easier for us to perceive the common region (flag) when comparing their relevancy to the concept *flag-US*.

Fig. 9. Instances of *flag-US* with different backgrounds in TRECVID data set.

An ideal kernel should thus reduce the impact of backgrounds. With reference to Figure 9, suppose there is a bin (visual word) representing people. This bin should have a nonzero weight w for the keyframe I_1 on the right hand side, but its weight is zero for the other keyframe. The responses of different kernels at this particular bin are:

$$\mathcal{K}_{sub-linear}(I_1, I_2) = e^{-\rho|w-0|^{0.5}} = e^{-\rho w^{0.5}}$$
$$\mathcal{K}_{Laplacian}(I_1, I_2) = e^{-\rho|w-0|} = e^{-\rho w}$$
$$\mathcal{K}_{\chi^2}(I_1, I_2) = e^{-\rho\frac{(w-0)^2}{w+0}} = e^{-\rho w}$$
$$\mathcal{K}_{Gaussian}(I_1, I_2) = e^{-\rho(w-0)^2} = e^{-\rho w^2}.$$

The sub-linear RBF kernel has a sub-linear exponential decay, while the Laplacian RBF and χ^2 RBF kernels have a linear exponential decay, and the Gaussian RBF kernel has a quadratic exponential decay. An ideal distance function should give small response (or equivalently a larger kernel response) to the background variance. Thus the kernels with linear/sub-linear exponential decay appear as better choices than the Gaussian RBF kernel. This conclusion is consistent with the observation of [34] using color histogram for image classification.

Among different kernels, the computational time of linear kernel and histogram intersection kernel is shorter than that of the generalized RBF kernels. The sub-linear RBF kernel is the slowest since it contains a time-consuming square root for nonzero components of every support vector. For the BoW representation, as shown in our experiments, we suggest to use kernels with linear exponential decay, i.e., the Laplace RBF kernel or the χ^2 RBF kernel. In the rest of our experiments, χ^2 RBF kernel is employed.

VII. Discussion

In this section we further evaluate and discuss the effectiveness of our BoW representation using data sets other than the TRECVID 2006. We first evaluate the generalizability of our empirical findings to two recent data sets. We then study the degree of performance improvement when fusing the BoW feature with global features such as color and texture. Finally, we extend our method to detect a large set of 374 concepts and discuss the detection performance.

A. Generalizability to Other Data Sets

We use TRECVID 2008 and PASCAL VOC 2007 data sets to study the generalizability of our empirical findings. Different from the TRECVID 2006 data set which is composed of broadcast news videos, the TRECVID 2008 data set mainly consists of documentary videos from the Netherlands Institute for Sound and Vision, where the training and test sets contains 43,616 and 35,766 shots respectively. There are 20 semantic concepts evaluated in TRECVID 2008.

The PASCAL VOC 2007 data set was used for the PASCAL Visual Object Classes Challenge 2007 [39]. In total, there are 9,963 images, which were divided evenly into training and test sets. 20 semantic concepts are evaluated on this data set, covering four major topics: person, animals, vehicles, and indoor scenes. Note that the detection performance on PASCAL VOC 2007 is measured by the conventional AP and mean AP (MAP) is used to aggregate the performance of multiple concepts. Compared with the TRECVID data sets, the PASCAL VOC data set is also smaller and less diversified. We choose it as it has been frequently used as a benchmark for evaluating keypoint-based features.

For both data sets, two detectors, DoG and Hessian Affine [25], are used to extract local keypoints. The keypoints are then described using SIFT. Here we choose two detectors because there is plenty of evidence in recent work which shows that the combination of various keypoint detectors leads to better performance [5], [6], [40], [41]. For each keypoint detector, we sample and cluster around 550,000 keypoints to generate a visual vocabulary of 500 visual words for each data set. The soft-weighting and the χ^2 RBF kernel SVM are then adopted for constructing and classifying BoW features respectively. The classification outputs of features generated from different keypoint detectors are combined using average fusion.

The per-concept detection performances on both data sets are shown in Figure 10. Based on our observations in Section VI-E, relatively coarse spatial partition is preferred. Thus in this experiment, we include one more spatial partition (1×3) and test three choices, 1×1 (whole keyframe), 1×3, and 2×2. The choice 1×3 is also adopted in the best performing system of PASCAL VOC 2007 [39]. From Figure 10 we see that the overall performances of each spatial partition are very close. This is consistent with the results on TRECVID 2006 data set (Section VI-E). On TRECVID 2008 data set, some scene concepts such as *street* and *mountain* benefit from using spatial information, while the performances of object concepts such as *bridge* and *airplane* are degraded by spatial partition. Similar observations also hold for the PASCAL VOC 2007 data set. We further combine the detection outputs from different spatial partition choices using average fusion. As can be seen in Figure 10, the combination of different spatial partitions greatly improves the performance (16% on TRECVID 2008 and 8% on PASCAL VOC 2007). The results indicate that although different types of concepts favor different spatial partitions, the fusion of multiple partition choices is helpful for most concepts, and thus should be used for better performance.

Figure 11 and 12 further compare our results with the state-of-the-art approaches on both data sets. Our submitted runs in TRECVID 2008 (dark blue bars) [41] are based on the BoW feature representation discussed in this paper, which achieve very competitive performance, ranking top-10 out of all the 200 official submissions. Among the top-20 runs, 15 are based on BoW feature [6], [41]. More interestingly, all the top-14 runs used soft-weighting techniques (our soft-weighting method and [24]). This indeed proves the effectiveness of soft-

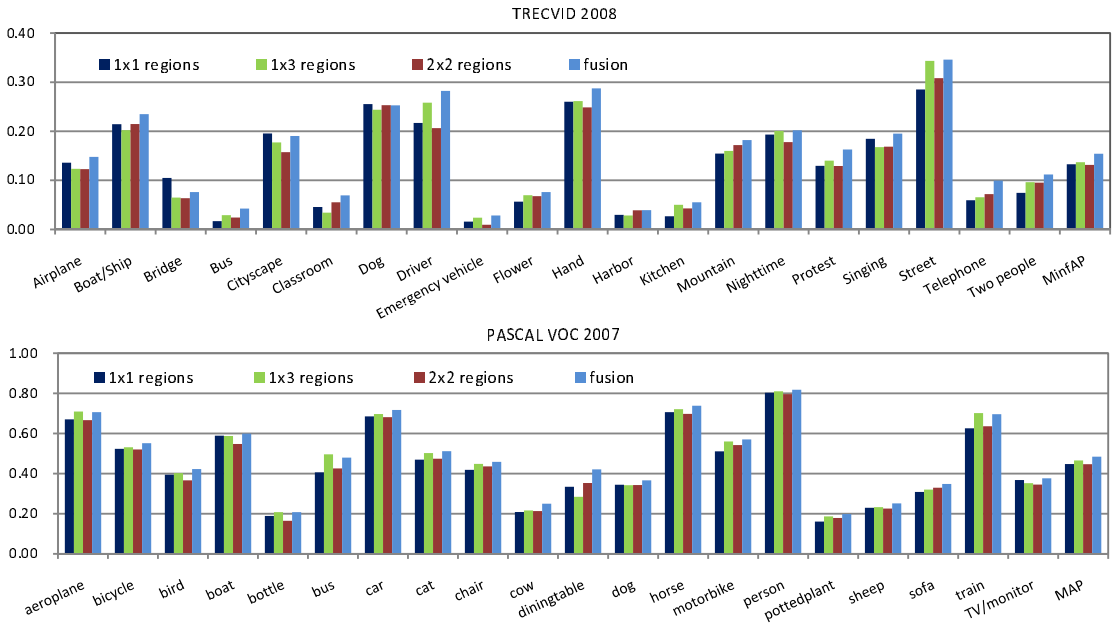

Fig. 10. Concept detection performance on TRECVID 2008 and PASCAL VOC 2007 data sets using BoW features computed from different spatial partitions. At each single partition choice, the MinfAP/MAP performances are similar. However, the fusion of different spatial partition choices shows noticeable performance improvement (16% and 8% MinfAP/MAP improvements for TRECVID 2008 and PASCAL VOC 2007 respectively).

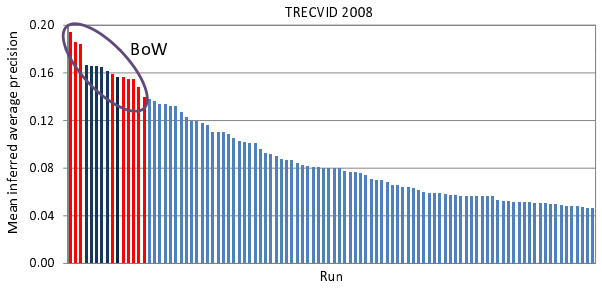

Fig. 11. Performance of the top-100 (out of 200) official runs in TRECVID 2008. Within the top-20 runs, 15 (circled) are based on BoW features. All of our 6 submissions (dark blue bars) ranked top-10 [41].

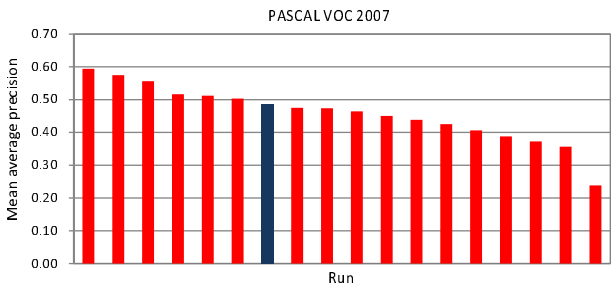

Fig. 12. Performance comparison of our result (blue bar) with the 17 official submissions of PASCAL VOC 2007. All of the methods are based on local features. Our result is shown in blue.

weighting in semantic concept detection. On the other hand, all the submissions in PASCAL VOC 2007 relied on local keypoint features, with emphasis ranging from keypoint detectors, descriptors, to advanced machine learning techniques. Compared with the 17 official submissions, our result ranks the 7th. It is also interesting to note that many of the runs in PASCAL VOC 2007 utilized not only sparse sampled keypoints (Harris Laplace and Laplace of Gaussian), but also densely sampled image patches. Since the large number of densely sampled local image patches demands heavy computation load for SIFT calculation and vector quantization, our BoW representation has the advantage of speed efficiency.

B. Fusion with Color/Texture Features

Global features such as color and texture are extensively used in image and video classification. While keypoint features describe the local structures in an image and do not contain color information, global features are statistics about the overall distribution of color, texture, or edge information. Global features have been used for concept detection in many previous studies [2], [11], [13], [15]. It is interesting not only to compare the performance of the two features, but also to see whether their combination further improves the performance.

We experiment with two types of global features: color moment (CM) and wavelet texture (WT). In CM, we calculate the first 3 moments of 3 channels in Lab color space over 5×5 grid partitions, and aggregate the features into a 225-d feature vector. For WT, we use 3×3 grids and each grid is represented by the variances in 9 Haar wavelet sub-bands to form a 81-d feature vector. We compare their performance with that of the local features (BoW) on the TRECVID 2008 data set.

Average fusion is used to combine different features. Table I shows the results on the TRECVID 2008 data set. We can see that BoW (with soft-weighting and χ^2 RBF kernel) significantly outperforms CM, WT and their combination. This indeed proves the effectiveness of local features for semantic concept detection, even though they contain no color information. By fusing BoW with global features, the performance is slightly improved by 3-4%. The degree of improvement,

TABLE I
MINFAP PERFORMANCE OF FUSING BOW WITH COLOR MOMENT (CM) AND/OR WAVELET TEXTURE (WT) ON TRECVID 2008 DATA SET. THE PERCENTAGE IN THE PARENTHESIS SHOWS THE DEGREE OF IMPROVEMENT OVER THE BOW ONLY PERFORMANCE (THE 5TH COLUMN).

	CM	WT	CM+WT	BoW	BoW+CM	BoW+WT	BoW+CM+WT
MinfAP	0.050	0.031	0.060	0.154	0.159 (3%)	0.154 (0%)	0.160 (4%)

however, is not as apparent as that on the TRECVID 2006 data set, which is as high as 50% [2]. This is due to the fact that TRECVID 2006 data set is composed of broadcast news videos which contain plenty commercial advertisements. The repetitive commercials result in many near-duplicate video shots on which global features work very well. For PASCAL VOC 2007 data set which does not contain near-duplicate images, similar observation is also noted where fusion with global features does not lead to apparent improvement.

C. VIREO-374: LSCOM Semantic Concept Detectors

We further apply our method to detect a large set of 374 semantic concept detectors, namely VIREO-374. With the goal of stimulating innovation in concept detection technique and providing better large scale concept detectors for video search, the detectors as well as features and detection scores on recent years' TRECVID data sets (2005–2009) have been released[1].

The VIREO-374 detectors are trained on TRECVID 2005 development set using three features (BoW, CM, and WT). On a leave-out validation set (a subset of the TRECVID 2005 development set), the mean performances of the 374 concepts are 0.150 for BoW and 0.174 for the fusion of BoW and the global features (CM and WT). The fusion with global features improves the performance by 16%. This is probably due to the fact that there are also many near-duplicates in the TRECVID 2005 data set. The effectiveness of the detectors is also evidenced in other works [42], [43], [44] which adopted VIREO-374 detectors for semantic video indexing. These works have reported promising search performance by utilizing the 374 detectors to perform query by text keywords [42], [43] and query by multimedia examples [44].

VIII. CONCLUSION

We have investigated various representation choices in BoW feature for semantic concept detection. By jointly considering the vocabulary size, weighting scheme, stop-word removal, feature selection, spatial information and visual bi-gram, the BoW shows surprisingly strong performance regardless the colorless and essentially orderless representation.

We have shown that all the six investigated representation choices, together with the kernel choice in SVM classifier, are influential to the performance of BoW. The vocabulary size, however, exhibits less or even insignificant impact when our soft-weighting scheme is in use. This indeed motivates and verifies the need of a weighting scheme specifically for visual words to alleviate the impact of clustering on vocabulary generation. In addition, we show that using appropriate feature selection methods (*IG* and *CHI*) can remove half of vocabulary without hurting the performance, this will significantly reduce the computational time especially when detecting thousands of concepts in large-scale multimedia databases.

Currently our works are grounded on keyframes and thus temporal information within a video shot is not considered. When extending to multiple frames or the whole frame sequence per shot, the detection performance may be further improved, but with additional computational cost of feature extraction and classification. Nevertheless, the temporal information has been shown to be effective particularly for the detection of event-type concepts in [45], [46]. Whether there is a more efficient way of utilizing the temporal information still deserves future research.

[1]Download site: http://vireo.cs.cityu.edu.hk/research/vireo374/

REFERENCES

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

[2] Y. G. Jiang, C. W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM CIVR*, 2007.

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

[4] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV*, 2006.

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

[6] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. van Gemert, and et al., "The MediaMill TRECVID 2008 semantic video search engine," in *TRECVID workshop*, 2008.

[7] "LSCOM lexicon definitions and annotations," in *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3*, 2006.

[8] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 LSCOM semantic visual concepts," Columbia University, Tech. Rep., March 2007.

[9] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM Multimedia*, 2006.

[10] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *ACM MIR*, 2006.

[11] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, and et al., "Intelligent multimedia group of Tsinghua university at TRECVID 2006," in *TRECVID workshop*, 2006.

[12] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, and et al., "Columbia university trecvid-2006 video search and high-level feature extraction," in *TRECVID workshop*, 2006.

[13] M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, and et al., "IBM research TRECVID-2006 video retrieval system," in *TRECVID workshop*, 2006.

[14] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and et al., "Multi-lingual broadcast news retrieval," in *TRECVID workshop*, 2006.

[15] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, and et al., "The Mediamill TRECVID 2006 semantic video search engine," in *TRECVID workshop*, 2006.

[16] K. E. van de Sande, T. Gevers, and C. G. M. Snoek, "A comparison of color features for visual concept classification," in *ACM CIVR*, 2008.

[17] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *ACM MIR*, 2007.

[18] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[19] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *ECCV*, 2006.

[20] S. Petrov, A. Faria, P. Michaillat, A. Berg, D. Klein, and et al., "Detecting categories in news video using acoustic, speech, and image features," in *TRECVID workshop*, 2006.

[21] A. C. Berg and J. Malik, "Geometric blur for template matching," in *CVPR*, 2001.

[22] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *ICCV*, 2005.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[24] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on PAMI*, vol. 99, no. 1, 2009.

[25] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, pp. 63–86, 2004.

[26] ——, "A performance evaluation of local descriptors," *IEEE Trans. on PAMI*, vol. 27, no. 10, 2005.

[27] Y. Yang and X. Liu, "A comparative study on feature selection in text categorization," in *ICML*, 1997.

[28] W. B. Cavnar and J. M. Trenkle, "N-gram based text categorization," in *3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[29] C. M. Tan, Y. F. Wang, and C. D. Lee, "The use of bigrams to enhance text categorization," *Journal of Information Processing and Management*, vol. 30, no. 4, pp. 529–546, 2002.

[30] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *IEEE ICCV*, 2005.

[31] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. Bakir, "Weighted substructure mining for image analysis," in *CVPR*, 2007.

[32] F. Odone, A. Barla, and A. Verri, "Building kernels from binary strings for image matching," *IEEE Trans. on Image Processing*, vol. 14, no. 2, 2005.

[33] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE Trans. on PAMI*, vol. 26, no. 2, 2004.

[34] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, 1999.

[35] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[36] TREC Video Retrieval Evaluation (TRECVID), *http://www-nlpir.nist.gov/projects/trecvid/*.

[37] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[38] J. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, 2000, pp. 61–74.

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[40] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, and et al., "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1/2, pp. 43–72, 2005.

[41] S. F. Chang, J. He, Y. G. Jiang, E. E. Khoury, C. W. Ngo, and et al., "Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search," in *TRECVID workshop*, 2008.

[42] X.-Y. Wei and C.-W. Ngo, "Fusing semantics, observability, reliability and diversity of concept detectors for video search," in *ACM Multimedia*, 2008.

[43] R. Aly, D. Hiemstra, A. de Vries, and F. de Jong, "A probabilistic ranking framework using unobservable binary events for video search," in *ACM CIVR*, 2008.

[44] S. Tang, J.-T. Li, M. Li, C. Xie, Y.-Z. Liu, and et al., "TRECVID 2008 participation by MCG-ICT-CAS," in *TRECVID workshop*, 2008.

[45] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and et al., "Hierarchical spatio-temporal context modeling for action recognition," in *CVPR*, 2009.

[46] F. Wang, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and visual relatedness," in *ACM Multimedia*, 2008.