Evaluating Bag-of-Visual-Words Representations in Scene Classification

Jun Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
juny@cs.cmu.edu

Yu-Gang Jiang
Dept of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
yjiang@cs.cityu.edu.hk

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alex@cs.cmu.edu

Chong-Wah Ngo
Dept of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

Based on keypoints extracted as salient image patches, an image can be described as a "bag of visual words" and this representation has been frequently used in the classification of imagery data. The representation choices regarding the dimension, selection, and weighting of visual words are crucial to the classification performance but have not been thoroughly studied in existing works. Given the analogy between this image representation and the bag-of-words representation of text documents, we apply techniques widely used in text categorization, including term weighting, stop word removal, feature selection, to generate image representations that differ in the dimension, selection, and weighting of visual words. The impact of these representations choices to scene classification is studied through extensive experiments on the TRECVID and PASCAL collections. This study provides an empirical basis for designing visual-word representations that are likely to produce superior classification performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

1. INTRODUCTION

Classifying images or video scenes into semantic categories is a problem of great interest in both research and practice. For example, an online collection of photos needs to be grouped into categories like "landscape", "portrait", and "animal" to support efficient browsing. To search over a large archive of news video, we want to classify video frames by the presence of certain scenes (e.g., meeting) and objects (e.g., buildings) and by semantic topics (e.g., politics). Scene classification is typically based on real-valued feature vectors describing the color, texture, and other visual properties of images. This representation is significantly different from the sparse and discrete term-vector document representation in text categorization, and therefore, there has been little connection between the two streams of research.

Recently, there is a trend of using *image keypoints* or *local interest points* in the retrieval and classification of imagery data [6, 7, 3, 14, 20, 19]. Keypoints are salient image patches that contain rich local information about an image, which can be automatically detected using various detectors [9, 19] and represented by many descriptors [10]. Keypoints are then grouped into a large number of clusters with those with similar descriptors assigned into the same cluster. By treating each cluster as a *"visual word"* that represents the specific local pattern shared by the keypoints in that cluster, we have a *visual-word vocabulary* describing all kinds of such local image patterns. With its keypoints mapped into visual words, an image can be represented as a *"bag of visual words"*, or specifically, as a vector containing the (weighted) count of each visual word in that image, which is used as feature vector in the classification task.

This visual-word image representation is analogous to the bag-of-words representation of text documents in terms of both form and semantics, which makes techniques for text categorization readily applicable to the problem of scene classification. In this paper, we use text categorization techniques, including term weighting and normalization, stop word removal, and feature selection, to generate image representations with different dimension, selection, and weighting of visual words and study their effectiveness in scene classification tasks. The goal is to provide a missing link in the existing works, where most of the effort has been on various keypoint detectors, keypoint descriptors, and clustering and classification algorithms [6, 7, 3, 14, 20, 19]. In comparison, the issues studied in this paper focus on the representation choices of the visual-word features, which are critical to the classification performance but yet to be thoroughly studied. By empirically examining various representation choices, we

WOODSTOCK '97 El Paso, Texas USA

intend to answer the question of what visual-word representation choices (w.r.t dimension, weighting, selection, etc) are likely to give the best classification performance in terms of accuracy and efficiency.

We evaluate the image classification performance based on various visual-word representations generated by text categorization techniques on two benchmark corpora, TRECVID and PASCAL, in order to study the impact of different representation choices. The experiments lead to the following important observations: (1) the size of an effective visual-word vocabulary varies from thousands to tens of thousands; (2) binary visual-word features are as effective as *tf* or *tf-idf* weighted features; (3) using selection criteria such as chi-square and mutual information, half of the visual words in the vocabulary can be eliminated with minimum loss of classification performance; (4) frequent visual words are usually very informative and must not be removed; (5) the spatial information of keypoints is helpful under small vocabularies. These observations are critical to designing the most effective visual-word representation for image classification and other related tasks.

In Section 2, we briefly review the existing works on image classification and text categorization. We describe the generation of bag-of-visual-words image representation in Section 3, and discuss the text categorization techniques for generating various representations in Section 4. We introduce the testing corpora and explore the distribution of visual words in Section 5. The experiment results and conclusions are presented in Section 6 and Section 7, respectively.

2. RELATED WORK

Representing images by effective features is crucial to the performance of image retrieval and classification. The most popular image representation has been the low-level visual features, which describes an image by the overall distribution of color, texture, or other properties. Features like color histograms and Gabor filters belong to this category. To include spatial information, an image is partitioned into either rectangular regions or segments of objects and backgrounds, and features computed from these regions/segments are concatenated into a single image feature vector. These conventional image representations are in the form of real-valued feature vectors, which is different from the sparse term vectors representing text documents.

Recently, the computer vision community has found keypoints to be an effective image representation for tasks varying from object recognition to image classification. Keypoints are salient image patches that contain rich local information of an image. They can be automatically detected using various keypoint detectors, which are surveyed in [9] and [19]. Keypoints are depicted by descriptors like SIFT (scale-invariant feature transform) [8] and its variant PCA-SIFT [5]. The keypoint descriptors are surveyed in [10]. Keypoint features can be used in their raw format for direct image matching [20], or vector-quantized into a representation analogous to the bag-of-words representation of text documents. There have been works using this vector-quantized keypoint feature, or bag-of-visual-word representation, for image classification [6, 7, 3, 14, 20, 19]. Our work examines the effectiveness of various representation choices,

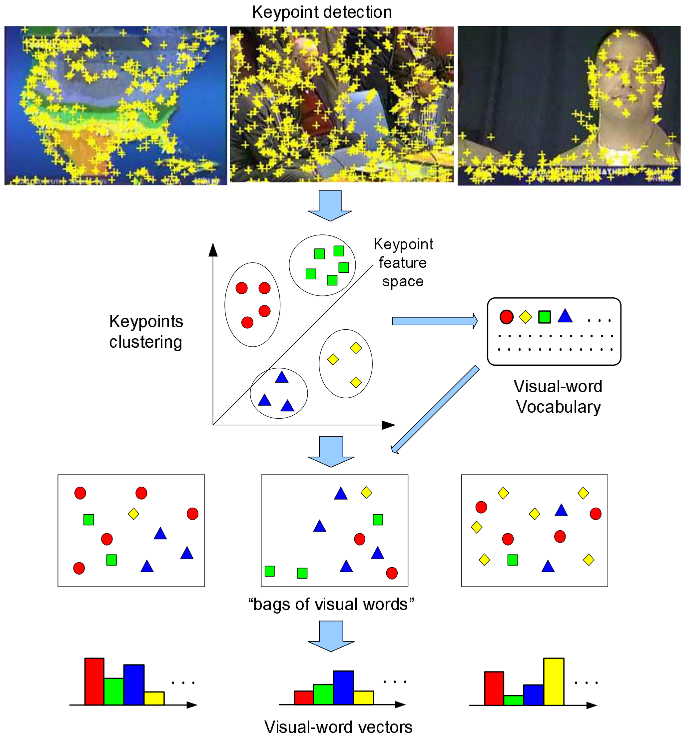

Figure 1: Visual-word image representation based on vector-quantized keypoint features

which is yet to be thoroughly studied in existing works.

Text categorization (TC) is a well studied area in IR. In TC, documents are represented as "bags of words" after stop-word removal and stemming. Each document is described either by a binary vector indicating the presence or absence of terms (e.g., [2]), or by a vector consisting of the *tf* or *tf-idf* weights of the terms (e.g., [4], [17]). Yang et al. [18] has studied the feature selection methods in TC, and found that up to 98% of the unique terms in the vocabulary can be eliminated without sacrificing classification accuracy. Different learning algorithms have been applied to TC, including SVM, k-Nearest Neighbor, Naive Bayes, Linear Least Square Fit, etc, which are surveyed in [17] and [2].

3. BAG-OF-VISUAL-WORDS

Similar to terms in a text document, an images has local interest points or keypoints defined as salient image patches (small regions) that contain rich local information of the image. Denoted by small crosses in the three images in Figure 1, keypoints are usually around the corners and edges in image objects, such as the edges of the map and around people's faces. We use the Difference of Gaussian (DoG) detector [8] to automatically detect keypoints from images. The detected keypoints are depicted using the PCA-SIFT descriptor, which is a 36-dimensional real-valued feature vector [5].

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) that require feature vectors of fixed dimension as input. To address this problem, we use the vector quantization (VQ) technique which clusters the keypoint descriptors in their feature space into a large number of clusters using the K-means clustering algorithm and 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. Thus, the clustering process generates a *visual-word vocabulary* describing different local patterns in images. The number of clusters determines the size of the vocabulary, which can vary from hundreds to over tens of thousands. Mapping the keypoints to visual words, we can represent each image as a *"bag of visual words"*. 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 characteristic of the whole image.

The bag-of-visual-words representation can be converted into a visual-word vector similar to the term vector of a document. The visual-word vector may contain the presence or absence information of each visual word in the image, the count of each visual word (i.e., the number of keypoints in the corresponding cluster), or the count weighted by other factors (see Section 4.3). Visual-word vectors are used in our image classification approach. The process of generating visual-word representation is illustrated in Figure 1.

4. REPRESENTATION CHOICES

Once images are represented as bags of visual words, we can classify them in the same way we classify text documents. The general approach is to build supervised classifiers from labeled images and apply them to predict the labels of other images. Specifically, there are many techniques that can affect the visual-word feature representation. Some are widely used in text categorization, such as term weighting, stop word removal, and feature selection, while others are unique to images, such as changing the vocabulary size and encoding the spatial information. We discuss these techniques below.

4.1 Vocabulary size

Unlike the vocabulary of a text corpus whose size is relatively fixed, the size of a visual-word vocabulary is controlled by the number of keypoint clusters in the clustering process. Choosing the right vocabulary size involves the trade-off between discriminativity and generalizability. With a small vocabulary, the visual-word feature is not very discriminative because dissimilar keypoints can map to the same visual word. As the vocabulary size increases, the feature becomes more discriminative, but meanwhile less generalizable and forgiving to noises, since similar keypoints can map to different visual words. Using a large vocabulary also increases the cost of clustering keypoints, computing visual-word features, and running supervised classifiers.

There is no consensus as to the appropriate size of a visual-word vocabulary. The vocabulary size used in existing works varies from several hundreds [6, 19], to thousands and tens of thousands [14, 20]. Their results are not directly comparable due to the difference on corpus and classification methods. To find out the proper range of vocabulary size, we experiment with vocabularies with sizes varying from 200 to 320,000. We are also interested in comparing the size of a visual-word vocabulary to that of a text vocabulary, which is usually around thousands to tens of thousands.

Table 1: Weighting schemes for visual-word feature

Name	Factors	Value for t_i
bxx	*binary*	1 if t_i is present, 0 if not
txx	*tf*	tf_i
txc	*tf, normalization*	$\frac{tf_i}{\sum_i tf_i}$
tfx	*tf, idf*	$tf_i \cdot \log(N/n_i)$
tfc	*tf, idf, normalization*	$\frac{tf_i \cdot \log(N/n_i)}{\sum_i tf_i \cdot \log(N/n_i)}$

4.2 Stop word removal

Stop word removal is a standard technique in text categorization. Are there also "visual stop words" that represent local image patterns totally useless for retrieval and classification? Sivic and Zisserman [14] claimed that the most frequent visual words in images are "stop words" and need to be removed from the feature representation. There is however no empirical evidence showing that removing them improves the image classification performance. Since it is very difficult to judge whether each visual word is a stop word, we focus on the relationship between the most frequent visual words and the classification performance.

4.3 Weighting schemes

Since term weighting is a key technique in IR [13, 1], we explore its use in adjusting visual-word vectors. Two major factors in term weighting are *tf* (term frequency) and *idf* (inverse document frequency). A third factor is the normalization factor, which converts the feature into unit-length vector to eliminate the difference between short and long documents. Many text categorization methods use weighting schemes based on these factors, such as *"tfc"* in [4], *"lfc"* in [18], while some simply use binary term vectors [2].

We apply popular term weighting schemes in IR to the visual-word feature vectors. These schemes are summarized in Table 1, where they are named after the convention in [13]. These schemes are chosen to allow us to study the impact of *tf*, *idf*, and the normalization factor. Note that tf_i is the number of times a visual word t_i appears in an image, N is the total number of images in the corpus, and n_i is the number of images having visual word t_i.

This is the first study on the weighting schemes of visual-word features. We have seen the use of vectors containing the counts of visual words (which are essentially *tf* features) for image classification [6, 19], and the use of *tf-idf* weighted features for image search [14, 20], but no comparisons have been made with other weighting schemes. As we will see, the best weighting scheme in IR does not guarantee good performance in image classification. In particular, the normalization factor, while eliminating the difference on the size of images, may have a negative effect. Even among images of the same size, the number of keypoints (visual words) varies according to the complexity of the image content. For example, an image showing a complex street scene may have over 1000 keypoints, while an image showing a smooth sky background may have less than 100 keypoints. An image with many keypoints usually has very different content from one with fewer keypoints, even though the relative distribution of the keypoints after mapping to visual words is similar. Normalization eliminates such difference and makes the two images less distinguishable.

4.4 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. [18] found out that, when a good criterion is used, up to 98% of the unique words in the vocabulary can be removed without loss of text categorization accuracy. In image classification, feature selection is potentially important as the size of the visual-word vocabulary is usually very high, but it has not been used in any existing work. We experiment with five feature selection criteria used in text categorization [18]:

- ***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 for image classification, we adopt two opposite selection criteria based on *DF*: *DF_max* chooses visual words with *DF* above a predefined threshold, while *DF_min* chooses visual words with *DF* below a threshold.

- x^2 ***statistics (CHI)***: The x^2 statistics measures the level of (in)dependence between two random variables [18]. Here we compute $x^2(t, c_i)$ between a specific visual word t and the binary label of an image class c_i. A large value of $x^2(t, c_i)$ indicates a strong correlation between t and c_i, and vice versa. Since $x^2(t, c_i)$ depends on a specific class, we compute the average statistics across a total of M image classes in the corpus as $x^2_{avg}(t) = \frac{1}{M}\sum_{i=1}^{M} x^2(t, c_i)$. We then eliminate visual words with $x^2_{avg}(t)$ below a threshold.

- ***Mutual information (MI)***: *MI* is another measure of the dependence between two random variables. The *MI* between a visual word t and a class label c is:

$$MI(t, c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} P(t, c) \log \frac{P(t, c)}{P(t)P(c)} \qquad (1)$$

  We compute $MI_{avg}(t) = \frac{1}{M}\sum_{i=1}^{M} MI(t, c_i)$, and remove visual words with $MI_{avg}(t)$ below a threshold.

- ***Pointwise Mutual information (PMI)***: *PMI* is directly related to *MI*. It uses one term in the sum of Eq.(1) to measure the association between a visual word t and a class label c:

$$PMI(t, c) = \log \frac{P(t = 1, c = 1)}{P(t = 1)P(c = 1)} \qquad (2)$$

  Visual words with small $PMI_{avg}(t)$ are eliminated from the vocabulary.

4.5 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 bag-of-visual-words representation described in Section 3 ignores such spatial information and may result in inferior classification performance. To integrate the spatial information, we 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 in 3 rows and 3 columns.

This region-based representation has its downside in terms of cost and generalizability. Dividing an image into $m \times n$ regions increases the feature dimension by $m \times n$ times, making the feature computationally expensive. 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 airplane in the training images are in different regions from that 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.

5. DATA EXPLORATION

We use two corpora to study the bag-of-visual-word representation and its use in image classification: the TRECVID 2005 corpus and the PASCAL 2005 corpus.

The TRECVID corpus contains 34-hour footage of broadcast news video of 6 programs, which was used for TREC Video Retrieval Evaluation 2005 [15]. The video has been segmented into a total of 29,252 shots, and a video frame is extracted from each shot as its keyframe. The data have been annotated with labels of 39 semantic concepts in the LSCOM-Lite project [11]. We rank the 39 concepts by frequency (i.e., the number of shots where the concept is present) and select the 20 most frequent concepts since the rare concepts have insufficient training data. These 20 concepts cover many different types, including outdoor scenes (e.g., waterscape, mountain), indoor scenes (e.g., meeting, studio), objects (e.g., car, computer), people activities (e.g., marching). The goal is to classify the 29,252 video frames according to the presence of any of the 20 semantic concepts. Note that this is a multi-label corpus in that there can be zero or more than one concept present in a video frame. This is a huge corpus with highly diversified content, as it contains any possible scenes from broadcast news, which makes the classification task very challenging.

The PASCAL corpus was used for the PASCAL Visual Object Classes Challenge 2005. It has 1578 labeled images from multiple sources, which belong to 4 categories as motorbikes, bicycles, people, and cars. Compared with TRECVID, PASCAL is smaller and less diversified, and its images are less noisy and cluttered than the video frames in TRECVID. We

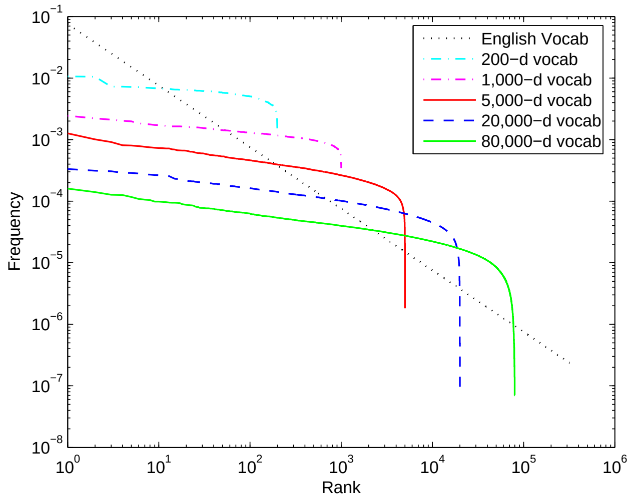

Figure 2: The frequency of visual words from vocabulary of different sizes plot against their frequency ranks in log-log scale.

choose it since it has been frequently used as a benchmark for evaluating keypoint-based features. Using a second and very different corpus also makes the conclusions in this paper more convincing.

The keypoints in both corpora are detected by the DoG detector [8] and described by the PCA-SIFT descriptor [5]. This results in an average of 490 keypoints per image in TRECVID, and 1,416 keypoints per image in PASCAL. For each corpus, we use the k-means clustering algorithm to cluster a pool of 1,000,000 randomly sampled keypoints into a visual-word vocabulary of any chosen size. The cluster memberships of the remaining keypoints are found using a KD-tree based fast nearest-neighbor search algorithm.

It is interesting to see how the visual words are distributed in an image corpus. Particularly, we want to know whether their distribution satisfies Zipf's law, which is followed by natural languages. Zipf's law says that the frequency of any (visual) word is roughly inversely proportional to its rank in the frequency table. We choose the TRECVID corpus for this study due to its huge size and diversified content. Under vocabularies of various sizes, we plot the frequency of visual words in TRECVID against their frequency rank in a log-log scale in Figure 2. Note that a Zipfian distribution must be a straight line in such scale. Despite the vocabulary size, we see that every distribution curve starts as a straight line up to a certain point, after which the curve plunges. This shows that, except for those with extremely low frequency, the visual words basically satisfy Zipf's law. We suspect that the extremely rare words are either noises in images or artifacts of the clustering algorithm, which produces very small clusters.

In Figure 2, the slope of a curve indicates how steep (unbalanced) the distribution is. For comparison, we draw an imaginary line to mimic the distribution of a English vocabulary. Obviously, the curves of visual words are not as steep as that of English words, showing that they are distributed more evenly than English words. What is less obvious but equally interesting is that the curve gets steeper as the vocabulary size increases. To see this, we compute the exponent parameter in Zipf's distribution, where larger exponents indicate steeper curves. We find the exponent increasing from 0.190 for a 200-d vocabulary, to 0.289 for a 5,000-d vocabulary and 0.32 for a 80,000-d vocabulary (The exponent is approximately 1 for English words). This shows that a larger vocabulary has a more unbalanced distribution of visual words.

6. EXPERIMENT RESULTS

We study the performance of image classification with different visual-word representations generated using the techniques discussed in Section 4. The TRECVID corpus is partitioned into a training set of 15-hour footage (15,745 keyframes) and a test set of 18-hour footage (13,507 keyframes). We guarantee that each set has a balanced mixture of data from different channels, and temporally adjacent frames are never assigned to both sets since they are too similar. The PASCAL corpus has been pre-divided into a training, validation, and test set, and we use the first two sets for training and the third for testing.

The classification is conducted in an "one-against-all" manner. Using the Support Vector Machines (SVM), we build 20 binary classifiers for the 20 semantic concepts in TRECVID, and 4 binary classifiers for the 4 object categories in PASCAL, where each classifier is for determining the presence of a specific concept or object. We use average precision (AP) to evaluate the result of a single classifier, and mean average precision (MAP) to aggregate the performance of multiple classifiers. Note that the state-of-the-art classification performance on TRECVID is around 0.2 to 0.3 in MAP since the classification is very difficult on this challenging corpus.

For comparison, we evaluate the classification performance on TRECVID using two conventional features, a 225-d color moment feature based on 5×5 regions, and 48-d Gabor texture feature. The MAP is 0.250 for the color moment and 0.182 for the Gabor feature. They are lower than the MAP achieved by the best visual-word feature (1000-d vocabulary, 3×3 region), which is 0.291. In fact, this best visual-word feature is comparable to the combination of color moment and Gabor feature, which achieves 0.292 MAP. More importantly, combining the three features together further improves the MAP to 0.349. This shows that visual-word feature is very effective for the image classification and complementary to the conventional image features. Since this paper focuses on the representation choices of visual-word feature, we leave the details of this comparison to our technical report [16]. We did not repeat this comparison on PASCAL, where keypoint features are clearly more effective since they are used by most top-performing methods.

6.1 Vocabulary size

Figure 3 shows the relationship between the classification performance and the size of the visual-word vocabulary. We use binary features ("bxx" in Table 1) without spatial information or feature selection. Both the linear and RBF kernel are used in SVM. For the RBF kernel, different choices of the gamma parameter are tried and the best result is reported.

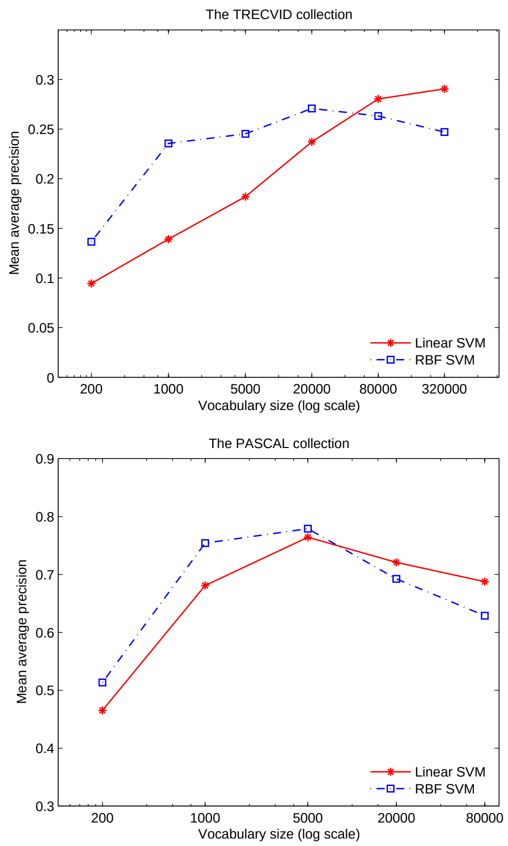

Figure 3: The classification performance at different vocabulary sizes on two corpora. (Note that the x-axis is in log scale.)

It is clear that the vocabulary size has a tremendous impact on the classification performance. On both corpora, as the vocabulary size increases from 200 to over 80,000, the performance first rises dramatically, peaks at certain points, and after that either levels off or drops mildly. The optimal vocabulary size is around 20,000 to 80,000 for TRECVID, and around 5,000 for PASCAL, both comparable to the size of a typical text vocabulary which is around thousands or tens of thousands. The difference can be explained by the fact that the keypoints in the smaller PASCAL corpus are not as widely-spread as those in TRECVID, and therefore demands fewer clusters (visual words). Although the optimal vocabulary size is clearly corpus dependent, this experiment suggests to search for the optimal one among relatively large vocabularies.

Another interesting observation comes from the comparison between the two kernels of SVM. For small vocabularies, the RBF kernel has a clear advantage over the linear one, but this advantage is reversed after the peak performance is reached. This suggests that the visual words in a small vocabulary are highly correlated, but become more independent and gain the property of linear separability as the vocabulary gets larger. When the visual words are independent, kernels that consider inter-feature correlations (e.g., RBF) have no advantage over linear kernels and may perform poorly due to overfitting.

Corpus	Whole Vocab	Percent of removed words				
		0.5%	1%	3%	5%	10%
TRECVID	0.280	0.279	0.278	0.275	0.273	0.267
PASCAL	0.778	0.778	0.777	0.775	0.773	0.771

Table 2: The classification performance after removing the most frequent visual words

6.2 Stop word removal

Do the most frequent visual words function like "stop words"? We approach this problem by examining the classification performance using vocabularies without the most frequent visual words, where the word frequency is computed from each corpus. As shown in Table 2, removing the most frequent words causes a small but steady decrease of performance on both corpora. This shows that these frequent visual words are unlikely stop words, since removing stop words should improve the classification performance. While it is premature to say there are no visual stop words, we show that eliminating the most frequent visual words is not desirable, which is against the claim in [14].

6.3 Weighting schemes

Now we move on to the problem of weighting schemes. Table 3 summaries the classification performance using visual-word features weighted by the 5 weighting schemes in Table 1. We use no spatial partitioning or feature selection in this experiment, but vary the vocabulary size to study their relationship with weighting schemes.

First, we focus on the comparison between the binary ("bxx") and *tf* feature ("txx") to see whether the counts of visual words are more informative than their presence or absence. It is only when the vocabulary size drops to 200 that the *tf* features consistently outperform the binary features. For larger vocabularies, the *tf* features are (slightly) worse than the binary features in most settings. This observation can be explained from two aspects. For one thing, as the vocabulary gets larger, the count of most visual words is either 0 or 1 and therefore *tf* features are not much different from binary features. On the other hand, the count information can be noisy. Suppose a certain visual word is typical among "building" images. An image containing 100 of this visual word is not necessarily more likely to be a "building" than an image containing only 20 of this visual word, but a classifier trained from the first image can be misled by the high count and classify the second image as "non-building".

Next, we examine the impact of the *idf* factor by comparing the performance of "txx" and "tfx". There is no consistent benefit of using *idf*, as "tfx" (which includes *idf*) is better than "txx' in about half of the settings but worse in the other half. We attribute this to the fact that a discriminative classifier like SVM can implicitly weight features to achieve maximum data separation, a presumably better weighting strategy than the heuristic *idf* method.

Finally, we have contradicting observations between the two corpora regarding the normalization factor. In PASCAL, "txc" (normalized) consistently outperforms "txx" (unnormalized), and "tfc" (normalized) outperforms "tfx" (unnormalized) in all but one setting. However, in TRECVID the

Corpus	Vocabulary size	Linear SVM					RBF SVM				
		bxx	*txx*	*txc*	*tfx*	*tfc*	*bxx*	*txx*	*txc*	*tfx*	*tfc*
TRECVID	200	0.095	**0.152**	0.109	0.147	0.110	0.137	**0.167**	0.112	0.130	0.108
	1,000	0.139	0.162	0.137	**0.183**	0.142	**0.235**	0.202	0.141	0.161	0.128
	5,000	0.183	0.178	0.150	**0.205**	0.153	**0.245**	0.224	0.141	0.194	0.145
	20,000	**0.237**	**0.228**	0.185	0.225	0.188	**0.271**	**0.278**	0.163	0.216	0.184
PASCAL	200	0.465	0.680	0.605	0.639	**0.693**	0.513	0.670	**0.742**	0.619	0.686
	1,000	**0.681**	0.677	0.677	**0.690**	**0.683**	**0.754**	0.639	0.751	0.618	0.722
	5,000	**0.764**	0.738	0.745	0.740	0.745	**0.777**	0.708	0.737	0.757	0.734
	20,000	**0.721**	0.682	0.708	0.682	0.711	**0.683**	0.642	**0.690**	0.528	**0.682**

* Weighting: *bxx = binary, txx = tf, txc = tf + normalization, tfx = tf + idf, tfc = tf + idf + normalization*

Table 3: The classification performance (MAP) on TRECVID and PASCAL corpus under different weighting schemes and vocabulary sizes. The bold font indicates the top performers in each setting.

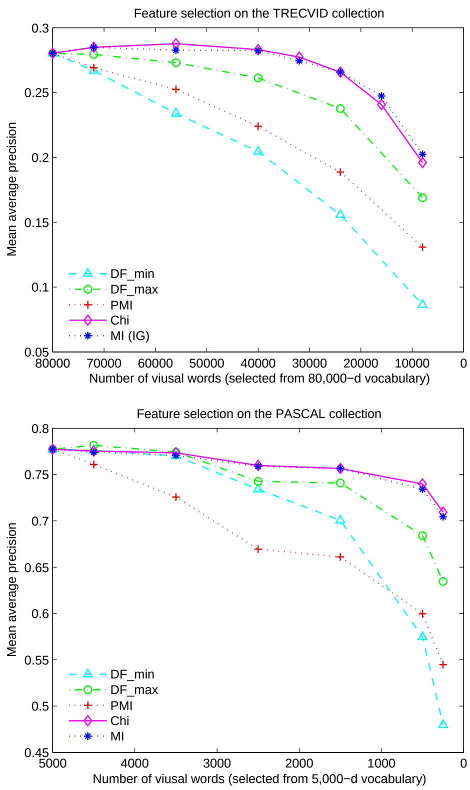

Figure 4: Classification performance under vocabularies pruned using various feature selection criteria.

un-normalized features are always better than their normalized counterparts. A plausible explanation is that, PASCAL has images of various sizes, and its classification performance benefits from the normalization factor which eliminates the difference on image sizes. This is not the case with TRECVID, which contains video frames of identical size; instead, normalization hurts the performance by suppressing the information on the number of keypoints in each video frame.

6.4 Feature selection

We examine feature selection techniques on the best vocabulary for each corpus, i.e., a 80,000-d vocabulary for TRECVID and a 5,000-d vocabulary for PASCAL. The 5 feature selection criteria discussed in Section 4.4 are compared, which are *DF-max*, *DF-min*, *CHI*, *MI*, and *PMI*. We reduce the vocabulary size to several percentages of its original size (90%, 70%, ..., 10%) by removing the most uninformative words determined by each criterion, and evaluate the classification performance in each setting. The results are shown in Figure 4.

We see that when effective criteria like *MI* and *CHI* are used, there is only minimum loss of MAP when the vocabulary is cut by half. When the vocabulary is reduced by 70%, the MAP has dropped merely by 5%, but after that it drops at a much faster rate. This shows that feature selection is an effective technique in image classification. In comparison, in text categorization a vocabulary can be reduced by up to 98% without loss of classification accuracy [18], which implies that the percentage of uninformative terms in text documents is much larger than in images.

Among different feature selection methods, *CHI* and *MI* are top performers on both corpora, followed by *DF_max*, while the performance of *DF_min* and *PMI* are lower than the others. This order is basically consistent with that in the text categorization [18][1]. The fact that *DF_max* is significantly better than *DF_min* implies that frequent visual words widely spread among images are more informative than rare visual words in terms of discriminative power. This is (again) consistent with the finding in text categorization that frequent words other than stop words are more informative than rare words [18]. It also partially explains why the feature selection can be done more aggressively on text documents than on images. As shown in Figure 2, the distribution of text words is much more uneven than that of visual words, which means there is a larger percentage of un-informative rare words to be eliminated from a text vocabulary.

6.5 Spatial information

The importance of spatial information can be seen by comparing the classification performance between the plain visual-word features and the region-based ones. We examine 4 ways of partitioning image regions, including 1×1 (the whole image), 2×2 (4 regions), 3×3 (9 regions), and 4×4 (16 regions). Figure 5 shows the classification performance on both cor-

[1] By definition, *MI* in this paper is equivalent to *IG* in [18], while *PMI* here is equal to *MI* in that paper.

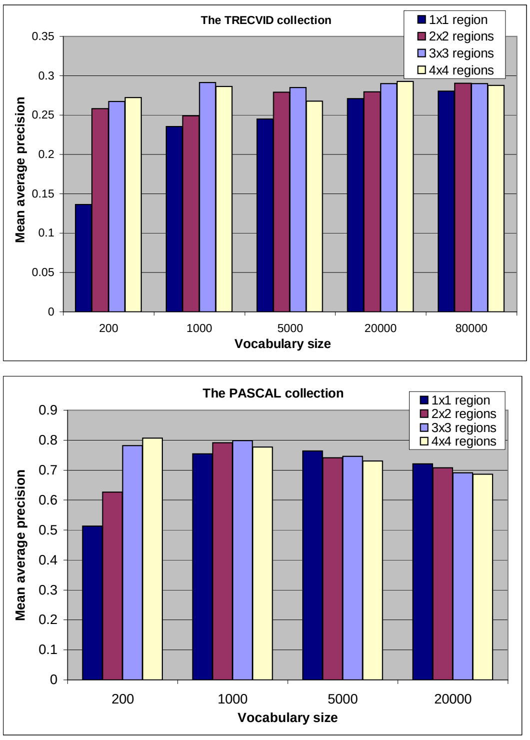

Figure 5: Classification performance of region-based features computed from different spatial partitions

pora using different spatial partitions and vocabulary sizes. For each setting, we experiment with both the linear and RBF kernel of SVM, and the performance of the better one is reported.

We see that the spatial information substantially improves the classification performance when the vocabulary is small. With a 200-d vocabulary, as the partition changes from 1×1 to 4×4, the MAP doubles on TRECVID and increases by 60% on PASCAL. However, as the vocabulary size increases, the spatial information is of little help in TRECVID and hurts the performance in PASCAL. This shows the contribution of large vocabulary and spatial information are not orthogonal. Overall, using a small vocabulary (e.g., 200-d or 1,000-d) with 3×3 or 4×4 partition is a good combination which achieves top or close-to-top performance and is less expensive than using a larger vocabulary.

7. CONCLUSION

Bag-of-visual-word is an effective image representation in the classification task, but various representation choices w.r.t its dimension, weighting, and word selection has not been thoroughly examined. In this paper, we have applied techniques used in text categorization, including term weighting, stop word removal, feature selection, to generate various visual-word representations, and studied their impact to classification performance on the TRECVID and PASCAL collections. This study provides an empirical basis for designing visual-word representation that is likely to produce superior classification performance.

The analogy between visual words in images and words in documents opens up opportunities for migrating techniques of information retrieval (IR) to solve problems in image and video data. Given the success on the classification task, we plan to apply IR techniques to image and video search based on the bag-of-visual-words representation. While there has been some pilot works on this direction [14, 20], a thorough study of this approach is missing. More interesting future work is to build "visual language models" that describe the distribution of visual words in images. Such visual language models provide a generative view of images, and can be used for image retrieval and classification using existing language modeling techniques for IR [12]. We can even build bigram or trigram type of models of visual words to capture the spatial relationships of adjacent keypoints, which could be more powerful in terms of describing complex image content.

8. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Series/Addison Wesley, 1999.

[2] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. of the 7th Int'l Conf. on Information and Knowledge Management*, pages 148–155, 1998.

[3] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM Int'l Conf. on Image and Video Retrieval*, 2007.

[4] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proc. of the 10th European Conf. on Machine Learning*, pages 137–142. Springer-Verlag, 1998.

[5] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[7] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[9] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[11] M. R. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, J. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. In *IBM Research Technical Report*, 2005.

[12] J. M. Ponte and W. B. Croft. A language modeling

approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an Int'l Journal*, 25(5):513–523, 1988.

[14] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of 9th IEEE Int'l Conf. on Computer Vision, Vol. 2*, 2003.

[15] A. Smeaton and P. Over. Trecvid: Benchmarking the effectiveness of infomration retrieval tasks on digital video. In *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.

[16] J. Yang and A. Hauptmann. A text categorizaton approach to video scene classification using keypoint features. In *CMU Technical Report*, 2006.

[17] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proc. of the 22nd Annual int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 42–49, 1999.

[18] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *14th Int'l Conf. on Machine Learning*, pages 412–420, 1997.

[19] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: An in-depth study. In *Technical report, INRIA*, 2005.

[20] W. Zhao, Y.-G. Jiang, and C.-W. Ngo. Keyframe retrieval by keypoints: Can point-to-point matching help? In *Proc. of 5th Int'l Conf. on Image and Video Retrieval (CIVR)*, pages 72–81, 2006.