Visual Words Selection
based on Class Separation Measures

Przemysław Górecki, Piotr Artiemjew, Paweł Drozda, Krzysztof Sopyła
Department of Mathematics
and Computer Sciences,
University of Warmia and Mazury,
Olsztyn, Poland
Email: pgorecki@matman.uwm.edu.pl, artem@matman.uwm.edu.pl,
pdrozda@matman.uwm.edu.pl, ksopyla@uwm.edu.pl

Abstract—Bag of Visual Words is one of the most effective image representations. One of the optimization methods for BoVW is the selection of the most informative visual words, which leads to more compact visual dictionaries and more accurate categorization. In this paper we investigate the problem of feature selection in the Bag of Visual Words framework. The main contribution is the presentation of two novel methods for visual word selection. The first one choses the features which are the best at separating one class from the rest (MFM1 one-vs-all). In the second method, the features which are the best at separating class pairs are selected (MSF6 one-vs-one). The effectiveness of the proposed methods is verified empirically on two different image datasets.

Index Terms—visual bag of words, feature selection, SVM

I. INTRODUCTION

In this paper we investigate the problem of visual word selection for image categorization. The Bag of Visual Words (BoVW) methods have become de facto a standard for semantic categorization of images [9], object localization, identification [9], and content based image retrieval [20], [12]. These methods have demonstrated their outstanding performance even in case of partial occlusions, clutter or viewpoint changes [19], [24], [17], [18].

In its simplest form, a bag of visual words describes an image as an orderless collection of local image patches. Typically, the patches are identified using a keypoint detector (i.e.: SIFT[16], SURF[5], BRIEF[6], ORB[21] or similar) and quantized into visual words by means of a clustering method (i.e. k-means). By mapping the patches into visual words the image can be represented as a histogram of visual words. An alternative keypoint identification strategy is based on sampling visual words from a dense grid of locations. In both cases, a histogram-based representation does not contain any information about the spatial layout of visual words in the image. Ignoring such information results in decreased classification performance.

Many extensions of the BoVW framework were proposed, which aimed at extending the image representation with information about the spatial relationship between visual words. Lazebnik et al. [15] introduced spatial pyramid matching. In this method, an image is recursively divided into a number of rectangular regions, and for each region an orderless feature representation is computed. In this way, an absolute arrangement of visual words is encoded in the spatial pyramid, which in turn allows for an effective matching between two images.

In some scenarios it is more efficient to augment the image representation with pairwise relations between neighboring visual words. In [23] the authors introduce spatial pyramid co-occurrences of visual words, which was motivated by Haralick’s work [13] on co-occurrences of pixel intensities. The combination of global spatial pyramid partitioning and local co-occurrences encodes both the absolute and relative layout of visual words.

In many cases, adding spatial information to an orderless BoVW increases the dimensionality of the feature space. For example, a spatial pyramid with 3 levels and 50 visual words results in feature vectors of size 4250 describing every image. Therefore, it is practical to reduce the dimensionality of the feature space. A more compact representation requires less storage space and less computational resources both for training and classification. In addition, an overall classification accuracy may improve if noisy dimensions are removed from the feature space.

This paper is organized as follows. Section II reviews current techniques for feature weighting and feature selection which are commonly used with Bag of Visual Words. In section III, we introduce a novel feature selection methods. The details of the experimental sessions validating our approach are presented in section IV. Section V concludes the paper.

II. RELATED WORK

The Bag of Visual Words approach derives from the Bag of Words model for text documents. Both frameworks use the concept of the dictionary of terms (features, visual words), but the visual dictionary is built by clustering local keypoint features in an unsupervised manner. The size of a visual dictionary is equal to the number of clusters obtained from the clustering process. In contrast, the vocabulary of text documents is relatively fixed.
There is a number of factors which can affect the performance of BoVW, among which there are feature (term) weighting and feature selection. These techniques are discussed below.

A. Feature weighting schemes

Term weighting methods have a significant impact on the quality of image classification. Three most common term weighting schemes are described below.

- **Term Frequency (TF)** indicates the number of times that the visual word (term) $t$ occurs in the image $d$. Another variant is the binary frequency, where $TF(d, t) = 1$ if $t$ occurs in $d$ and 0 otherwise.

- **Term Frequency - Inverse Document Frequency (TF-IDF)** reflects the importance of the visual word $t$ in the dataset $D$. The visual word that occurs in many images is less relevant that the one that only occurs in a few. This can be expressed using the following formula:

$$TFIDF(t, d) = TF(t, d) \log \frac{|D|}{|d : t \in d|}$$  \hspace{1cm} (1)

- **Normalization** is used to convert a feature vector into a unit length vector. This step improves the accuracy of image categorization and makes the feature vector independent of the number of keypoints located in the image.

B. Feature selection schemes

Yang et al. [8] studied the problem of feature selection for text categorization. The authors proposed five methods of calculating the term-goodness, followed by thresholding so that only the best terms were left in the dictionary. These methods are: Document Frequency thresholding (DF), Information Gain (IG), Mutual Information (MI), $\chi^2$ statistics (CHI) and Term Strength (TS). In [22] the authors used the Pairwise Mutual Information (PMI) in addition to DF, CHI, and MI. Similarly, Yo-Gang et al. applied DF, CHI, IG, and MI in [14] for visual word selection.

The details of the introduced selection criteria are summarized below.

- **Document Frequency (DF)** of a visual word is the number of images which contain this word. In text retrieval, it is common to remove very rare words, since they are uninformative. On the other hand, very frequent words are considered to be stop words, which should also be removed.

- **Information Gain (IG)** measures the dependence between a visual word $t$ and a class label $c_i$. This allows calculating an average information gain $IG_{avg}$ for each term $t$ in the following manner:

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

$$IG_{avg}(t) = \frac{1}{C} \sum_{i=1}^{C} IG(t, c_i)$$  \hspace{1cm} (3)

- **Mutual Information (MI)** is similar to IG and defined as follows:

$$MI(t, c_i) = \log \frac{P(t, c_i)}{P(t)P(c_i)}$$  \hspace{1cm} (4)

$$MI_{avg}(t) = \frac{1}{C} \sum_{i=1}^{C} MI(t, c_i)$$  \hspace{1cm} (5)

- **Pairwise Mutual Information (PMI)** is similar to MI and defined as follows:

$$PMI(t, c_i) = \frac{P(t = 1, c_i = 1)}{P(t = 1)P(c = 1)}$$  \hspace{1cm} (6)

$$PMI_{avg}(t) = \frac{1}{C} \sum_{i=1}^{C} PMI(t, c_i)$$  \hspace{1cm} (7)

- **$\chi^2$ statistics (CHI)** is another measure of correlation between a visual word $t$ and a class label $c_i$. Similarly to the previous criteria, $\chi^2_{avg} = \frac{1}{C} \sum_{i=1}^{C} \chi^2(t, c_i)$ can be computed to get an average score for $t$.

Previous studies [14], [22] have shown that IG and CHI are the most effective and IG slightly better than CHI. For this reason, we have compared our visual word selection methods with IG.

III. PROPOSED FEATURE SELECTION SCHEME

One of the key issues in BoVW is the selection of the most important visual words. Obtaining a more compact representation allows us to speed up the process of learning and classification. In addition, the classification accuracy may be improved if unimportant and noisy data is removed.

The methods for Visual Word selection have been applied in our previous studies related to DNA microarray classification [1], [2], [4], [11]. Microarrays, similarly to images with a BoVW representation, consist of a large number of features describing a single object. Our successful results presented in [1], [2], [4], [11] have motivated us to apply the same attribute selection methods to the image domain.

Without a loss of generality, we can denote a set of labeled images as the decision system in the form of triple $(U, A, d)$, where $U = \{ u_1, u_2, ..., u_n \}$ is the universe of objects (images), $A = \{ a_1, a_2, ..., a_m \}$ is the set of conditional attributes (visual words), and $d \notin A$ is the set of decision classes (image labels) $d : \{ c_1, c_2, ..., c_k \}$. For each attribute $a \in A$, we propose to obtain the rate of separation denoting the degree in which this attribute separates the objects of different decision classes. We propose to apply two different strategies for feature selection. The first one is based on the one-vs-all class separation strategy (MFM1)[4]. The second one is based on the one-vs-one class separation strategy (MSF6)[3].

A. One-vs-All strategy

Let us start with the details of the first method - MFM1 [4]. In this case, we have computed for all considered attributes $a$ and decision classes $c_i$ the rate $S^{c_i}(a)$ in which attributes
differentiate the central decision class $c_i$ from the other classes. The ratio is defined as follows:

$$S^{c_i}(a) = \frac{(C^a_i - \overline{C}^a_i)^2}{Z_{C^a_i}^2 + Z_{\overline{C}^a_i}^2}, \ a \in A.$$  

(8)

where,

$$C^a_i = \{a(u) : u \in U \text{ and } d(u) = c_i\}.$$  

(9)

$$\overline{C}^a_i = \frac{\sum a(u) : u \in U \text{ and } d(u) = c_i}{\text{card}\{C^a_i\}},$$  

(10)

$$\hat{C}^a_i = \frac{\sum a(v) : v \in U \text{ and } d(v) \neq c_i}{\text{card}\{U\} - \text{card}\{C^a_i\}},$$  

(11)

$$Z_{C^a_i}^2 = \frac{\sum a(u) \in C^a_i (a(u) - C^a_i)^2}{\text{card}\{C^a_i\}},$$  

(12)

$$Z_{\overline{C}^a_i}^2 = \frac{\sum a(v) \in U \setminus C^a_i (a(v) - \hat{C}^a_i)^2}{\text{card}\{U\} - \text{card}\{C^a_i\}}.$$  

(13)

After the rate of separation, $S^{c_i}(a)$ is computed for all genes $a \in A$ and all decision classes $c_i$; genes are sorted in the increasing order of $S^{c_i}(a)$,

$$S^{c_1}_1(a) > S^{c_2}_2(a) > ... > S^{c_k}_{\text{card}\{A\}}(a)$$

$$S^{c_2}_1(a) > S^{c_2}_2(a) > ... > S^{c_2}_{\text{card}\{A\}}(a)$$

$$...$$

$$S^{c_k}_1(a) > S^{c_k}_2(a) > ... > S^{c_k}_{\text{card}\{A\}}(a)$$

Finally, we choose for the experiments a fixed number of genes from the sorted list by means of the following procedure:

**Procedure**

**Input data**

$A' \leftarrow 0$

$\text{iter} \leftarrow 0$

**for** $i=1,2,...,\text{card}\{A\} \text{ do}$

**for** $j=1,2,...,k \text{ do}$

$S^{c_j}_i(a) = S^{c_j}_j(a)$

if $a \notin A'$ then

$A' \leftarrow a$

end if

$\text{iter} \leftarrow \text{iter} + 1$

if $\text{iter} = \text{fixed number of the best genes}$ then

**BREAK**

end if

**end for**

if $\text{iter} = \text{fixed number of the best genes}$ then

**BREAK**

end if

**end for**

return $A'$

The next algorithm (MSF6 [3]) has similar motivation to the previous one. However, we applied here $F$ statistics, extended on multiple decision classes, well known for separation of the two decision classes.

**B. One-vs-One strategy**

The MSF6 method is based on the strategy of feature selection by separating all pairs of decision classes (non-repeating combinations). In this case we propose to obtain the rate of separation of the feature $a \in A$ for pairs of decision classes $c_i, c_j$ where $i, j = 1, 2, ..., k$ and $i \neq j$ in the following way. We let,

$$F_{c_i,c_j}(a) = \frac{\text{MST}_R^{c_i,c_j}(a)}{\text{MSE}_{c_i,c_j}(a)}$$  

(14)

$$C^a_i = \{a(u) : u \in U \text{ and } d(u) = c_i\},$$  

(15)

$$C^a_j = \{a(v) : v \in U \text{ and } d(v) = c_j\}.$$  

(16)

$$\overline{C}^a_i = \frac{\sum a(u) : u \in U \text{ and } d(u) = c_i}{\text{card}\{C^a_i\}},$$  

(17)

$$\hat{C}^a_j = \frac{\sum a(v) : v \in U \text{ and } d(v) = c_j}{\text{card}\{C^a_j\}}.$$  

(18)

$$\overline{c}^{a}_{i,j} = \frac{\sum a(u) : u \in U \text{ and } (d(u) = c_i \text{ or } d(u) = c_j)}{\text{card}\{C^a_i\} + \text{card}\{C^a_j\}},$$  

(19)

$$\text{MST}_R^{c_i,c_j}(a) = \text{card}\{C^a_i\} \ast (\overline{C}^a_i - \overline{c}^{a}_{i,j})^2$$

$$+ \text{card}\{C^a_j\} \ast (\hat{C}^a_j - \overline{c}^{a}_{i,j})$$  

(20)

$$L = \sum_{l=1}^{\text{card}\{C^a_i\}} (a(u_l) - C^a_i)$$

$$M = \sum_{m=1}^{\text{card}\{C^a_j\}} (a(v_m) - C^a_j)$$

(21)

where $u_l \in C^a_i$, $l = 1, 2, ..., \text{card}\{C^a_i\}$, $v_m \in C^a_j$, $m = 1, 2, ..., \text{card}\{C^a_j\}$. After the rates of separation $F_{c_i,c_j}(a)$, are computed for all genes $a \in A$ and all pairs of decision classes $c_i, c_j$, where $i \neq j$ and $i < j$, genes are sorted in decreasing order of $F_{c_i,c_j}(a)$:

$$F^1_{c_{i_1},c_{i_2}} > F^2_{c_{i_1},c_{i_2}} > ... > F^{\text{card}\{A\}}_{c_{i_1},c_{i_2}},$$

where $i_1 \in \{1, 2, ..., k-1\}$ and $i_2 \in \{i_1 + 1, ..., k\}$. Finally, for the experiments we have chosen a fixed number of genes from the sorted list by means of the following procedure:

**Procedure**

**Input data**
The goal of the experimental session is to validate the proposed feature selection methods. This is achieved by selecting a subset of best features from the given dataset and classifying the reduced dataset. It is desirable that the classification accuracy for the reduced dataset is similar, or even better, compared to the original data. However, the accuracy will decrease if too few features are selected.

Our proposed feature selection methods are evaluated in two scenarios. In the first experiment, we use our in-house dataset of shoe images. It is divided into 5 categories, which are visually similar to each other, so this dataset is an example of a fine-grained categorization problem. Each image is represented as an orderless bag of words, computed from sparse SIFT.

In the second experiment, we use the Caltech101 dataset [10] and the spatial pyramid image representation introduced in [15]. This representation offers the features of much higher dimensionality, compared to the previous experiment.

For both cases we have tested our one-vs-one and one-vs-all feature selection methods by selecting a number of best features from the dataset, and by classifying it. The classification procedure consisted of 10-fold cross-validation using multi-class Support Vector Machines (SVM) [7].

A. Experiment 1 - Shoes dataset

The in-house shoes dataset consists of 200 shoe images divided into 5 distinctive visual categories containing 59, 20, 34, 29, and 58 images. Sample images for each category are shown in figure 1.

We have used the SIFT detector/descriptor extract all the keypoints from the dataset. Next, the descriptors were quantized into dictionaries of 500, 1000, 2500 and 5000 visual words. Given the vocabulary, we have created orderless BoVW features for each image. The feature vectors were pruned using our MFM1, MFM6, IG and MI methods and categorized using SVM with the \( \chi^2 \) kernel. The results of the classification are presented in figure 2. It can be noted that both MFM1 and MSF6 outperform standard IG and MI methods. For example, a dictionary of 500 visual words can be pruned to about 100 words using the MSF6 method without any loss in the accuracy, while IG and MI allow us to remove about 50% of the words without losing accuracy. In case of larger dictionaries, our methods allow us to remove even 95 per cent of visual words. For the size of about 5% of the original words the same classification accuracy is obtained, while IG and MI tend to slowly loose accuracy until 30% of the original size, when the accuracy drops rapidly.

B. Experiment 2 - Caltech101 dataset

As the name implies, the Caltech101 consists of over 8000 images divided into 101 categories. For this dataset, we have used a dense SIFT detector/descriptor in conjunction. In order to add the spatial information about the keypoints, we have used a spatial pyramid of height 3 with the dictionaries of sizes 50 and 200. Spatial pyramids greatly increase the dimensionality of the feature space. In case of a 3-level pyramid and 50 visual words, there are 4250 features. For 200 visual words, there are 17000 features. The goal of this experiment was to check if the features can be as effectively pruned as in case of an orderless BoVW. Similarly to the previous example, we have tested MFM1 and MSF6 for feature selection and SVM with pyramid matching kernel for classification. The results are presented in tables I and II. In both cases the dictionary can be pruned by half, without losing classification accuracy. Further pruning to 25% of the original dictionary size results in a small loss of accuracy (order of 1 – 2%). Moreover, MSF6 is more effective than MFM1 if more than 50% of visual words are removed. However, it is more computationally expensive than MFM1.

### IV. Experiments and Results

<table>
<thead>
<tr>
<th>Number of selected features</th>
<th>MFM1</th>
<th>MSF6</th>
<th>IG</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4250</td>
<td>67.72</td>
<td>67.72</td>
<td>67.72</td>
<td>67.72</td>
</tr>
<tr>
<td>4000</td>
<td>67.76</td>
<td>67.70</td>
<td>67.72</td>
<td>67.74</td>
</tr>
<tr>
<td>2000</td>
<td>66.79</td>
<td>67.47</td>
<td>66.35</td>
<td>65.12</td>
</tr>
<tr>
<td>1000</td>
<td>64.69</td>
<td>65.32</td>
<td>64.24</td>
<td>61.48</td>
</tr>
<tr>
<td>500</td>
<td>60.55</td>
<td>62.67</td>
<td>61.69</td>
<td>53.87</td>
</tr>
<tr>
<td>250</td>
<td>55.67</td>
<td>57.72</td>
<td>56.45</td>
<td>45.62</td>
</tr>
<tr>
<td>125</td>
<td>43.75</td>
<td>51.46</td>
<td>50.77</td>
<td>37.49</td>
</tr>
<tr>
<td>50</td>
<td>9.70</td>
<td>43.82</td>
<td>46.22</td>
<td>26.06</td>
</tr>
</tbody>
</table>

### TABLE I

Classification accuracy for the Caltech101 dataset and 50 visual words.
Fig. 1. Sample images from each of 5 shoe categories (a-e)

Fig. 2. Classification accuracy for Shoes dataset, pruned with MFM1, MSF6, IG and MI feature selection methods. a) for vocabulary size 500, b) 1000, c) 2500, d) 5000
In the context of gene selection, the SAM5 and SAM10 methods turn out to be the best, but in the context of visual word selection MFM1 and MSF6 methods are performing better than the algorithms based on Information Gain and Mutual Information. As concerning the future work, we plan to investigate other methods of visual word selection, in particular the extensions of the methods based on IG and MI.

V. Conclusions

In this paper we have investigated two novel methods of visual word selection previously used with success in the context of DNA Microarray gene selection - see [1], [2], [4], [11], [3]. In the context of gene selection, the SAM5 and SAM10 methods turn out to be the best, but in the context of visual word selection MFM1 and MSF6 methods turn out to be the most effective. A series of experiments have proven the effectiveness of our approach, it has also been shown that our methods are performing better than the algorithms based on Information Gain and Mutual Information. As concerning the future work, we plan to investigate other methods of visual word selection, in particular the extensions of the methods based on IG and MI.

Acknowledgements

The research has been supported by the grant N N516 480940 from The National Science Center of the Republic of Poland and grant 1309-802 from Ministry of Science and Higher Education.

References


