The Selection of Useful Visual Words in Class-Imbalanced Image Classification

Sutasinee Chimlek*, Part Pramokchon, Punpiti Piamsa-nga*****
*Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Thailand
** Department of Computer Science, Faculty of Science, Maejo University, Chiang Mai, Thailand
***Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

Article Info

Article history:

Received Jul 21, 2015
Revised Oct 31, 2015
Accepted Nov 16, 2015

Keyword:

Bag of visual words
Chi-square statistics
Feature selection
Image classification
Mutual information

ABSTRACT

The bag of visual words (BOVW) has recently been used for image classification in large datasets. A major problem of image classification using BOVW is high dimensionality, with most features usually being irrelevant and different BOVW for multi-view images in each class. Therefore, the selection of significant visual words for multi-view images in each class is an essential method to reduce the size of BOVW while retaining the high performance of image classification. Many feature scores for ranking produce low classification performance for class imbalanced distributions and multi-views in each class. We propose a feature score based on the statistical t-test technique, which is a statistical evaluation of the difference between two sample means, to assess the discriminating power of each individual feature. The multi-class image classification performance of the proposed feature score is compared with four modern feature scores, such as Document Frequency (DF), Mutual information (MI), Pointwise Mutual information (PMI) and Chi-square statistics (CHI). The results show that the average F1-measure performance on the Paris dataset and the SUN397 dataset using the proposed feature score are 92% and 94%, respectively, while all other feature scores do not exceed 80%.

Corresponding Author:

Punpiti Piamsa-nga,
Department of Computer Engineering, Faculty of Engineering
Kasetsart University, Bangkok, Chatujak, Bangkok, 10900, Thailand.
Email: pp@ku.ac.th

1. INTRODUCTION

Due to the efficiency and effectiveness of using a bag of visual words (BOVW), which was proposed by Sivic and Zisserman [1], it became very well-known in the fields of image retrieval and classification, e.g., PASCAL [2] and SUN [3]. The BOVW is used to represent local features and descriptors, along with geometry verification, which is motivated by an analogy, with the 'bag-of-words' representation for text categorization. There are publications [4]-[7] about visual content representation using the BOVW due to it being a promising method for visual content classification, annotation, and retrieval. The BOVW model of images may be classified in a class on the basis of visual word histograms. Visual words are obtained by clustering in the descriptor space [1], where all patches covered by one visual word represent the same part in images. Each image is represented using BOVW, no longer being suitable as a large number of images. Furthermore, multi-class image classification is useful to organize a large number of images, which are increasing significantly. The supervised learning process is used to produce a classifier using a pre-defined number of classes based on BOVW [9]. A major problem of image classification using BOVW is high dimensionality, most features of which are usually irrelevant, and the amount of data exceeds what can be stored in available memory. The size of BOVW has a tremendous impact on the classification performance [10]. Therefore, the selection of significant visual words for each class is an essential method to reduce the size of BOVW while retaining the high performance of image classification.

In general, feature selection approaches are used in image classification to reduce the dimension of the feature space and improve the efficiency and precision of the classifier [10]-[12]. These approaches aim to select efficient useful features from the original feature space according to some evaluation criteria. The feature ranking-based approach [13]-[15] is a well-known filter-based feature selection for handling a very huge number of features. In this approach, each feature is evaluated by a scoring measurement. All features are sorted in descending order; then, a small set of high-score features is kept as an optimal feature set and the rest of them are ignored. The feature ranking-based approach is simple, efficient and independent of types of classifiers; hence, it has been widely used in image classification. There are many efficient and effective feature scores based on measurement of the relevance of each individual feature to the class, such as Document Frequency (DF) [13], Pointwise Mutual information (PMI) [10], Mutual information (MI) [10], Chi-square statistics (CHI) [13], etc. Most scoring functions are based on visual word occurrence frequency in each class of the dataset, having an imbalanced distribution in reality. As a result, these feature scores also cause low performance. The class-imbalanced distribution that arises from the ranking-based selection excessively considers visual words that strongly relate to large classes (called majority classes) and tends to ignore visual words in small classes (called minority classes) [16], [17]. It is a basic notion that a visual word whose occurrence frequency in an image of a specific class is higher than that of other classes is desirable because it contains higher information and has more discriminating power than others.

Therefore, we apply and extend the t-score technique [18], which is based on the statistical t-test technique, to compute feature scores for multi-class-imbalance text classification. The t-score is based on the idea that features may discriminate particularly well between two classes if occurrence frequencies from both classes are significantly different. We use the t-score to estimate the discriminating power of each feature. The higher the score a feature has, the more relevance there is to discriminate a specific class from the others. We applied this to combine the class-specific score for a visual word that has three included t-scores: Max t-score, Averaged t-score, and Weighted Averaged t-score.

However, the images in each class have different appearance features from the text in each class because the images in each class can be taken from multiple views, as shown in Figure 1. Therefore, there are subclasses in each image class that represent each view. Each subclass has specific visual words. Thus, we proposed the t-score for image classification, which is effective in imbalanced distribution classes and multi-view images in each class. In this paper, we present a t-score method to further reduce the amount of BOVW stored from each image sub class, while still maintaining strong classification performance.

The rest of the paper is organized as follows. In Section 2, we present related works. In Section 3, we present the feature score for visual word selection in a subclass. We describe our experiments and the results are discussed in Section 4; we conclude in Section 5.

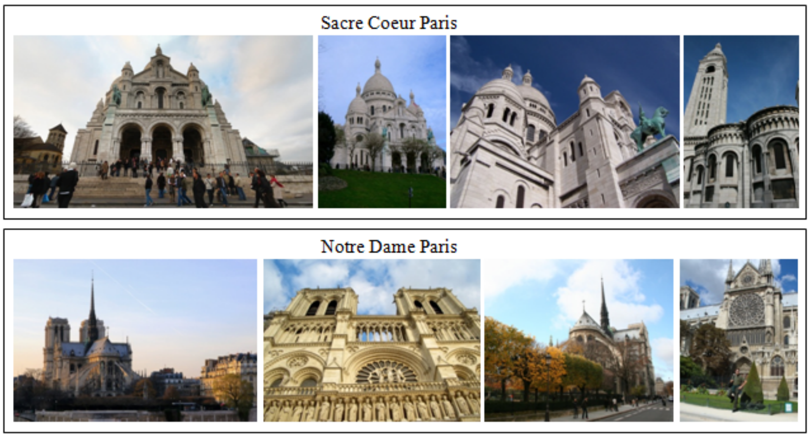

Figure 1. The example images of multiple views

2. RELATED WORK

2.1. Bag of Visual Word (BOVW)

The BOVW method is the state-of-the-art approach, which dominates in image classification and retrieval for large databases [1]. The methods that produce BOVW include the following three steps: feature extraction, feature quantization and BOVW generation. Feature extraction detects several local patches in each image and represents the patches as numerical vectors. Many interest point detectors [19], [20] and descriptors [21], [22] are proposed for use. The most used feature extraction in the bag-of-words model is the Scale-invariant feature transform (SIFT) descriptor [23]. The SIFT descriptor calculates the edge gradient in eight orientations for each of the tiles in the grid, thus resulting in a 128-dimension vector for each image. The SIFT descriptor has the ability to handle intensity, rotation, scale and affine variations.

Feature quantization produces a “visual word vocabulary” (analogous to a word dictionary). A visual word vocabulary represents similar patches. One simple method is to perform k-means clustering over all the vectors [1]. The visual words are then defined as the centers of the learned clusters. The number of clusters is the visual word vocabulary size (analogous to the size of the word dictionary). Each patch in an image is mapped to a certain visual word. The final step, BOVW generation, is performed to convert vector represented patches to visual words (analogous to words in text documents), which also represent each image by the histogram of the visual words.

2.2. Image Classification

Among supervised learning techniques, Bayesian classifiers [24] and Support Vector Machines (SVM) [24] are widely used. In image retrieval and classification, current visual word vocabulary sizes range from small, typically 1 K [25], to large, 1 M words [26]. Because of the computational and storage requirements, large visual word vocabularies are difficult to manage in real world scenarios that involve very large databases. Therefore, a method for visual word vocabulary reduction is imperative. There are several methods that try to reduce the visual word vocabulary significantly while keeping retrieval or classification performance constant.

The common methods proposed to reduce the visual word vocabulary try to keep visual words that frequently appear in the dataset [27], [28]. In contrast, Turcot and Lowe [29] use geometrically selected visual words, which are appropriate for constructing a reduced visual word vocabulary. However, for this technique, the reduced visual word vocabulary size depends on the geometric properties of the dataset images, which requires more computing. To solve the geometric constraint, we propose a selection of visual words from a pool of visual words that repeatedly appear in each subclass, which are robust against multi-views of the same scene or object.

2.3. Feature Scoring Function

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. Therefore, feature score is an important technique for reducing the visual word vocabulary size. This method measures the relevance between each visual word and the class by analyzing general characteristics of the training examples, such as information, dependency, distance, consistency, etc. A high-score visual word has useful features for classifying. There are several feature selection methods widely used in image classification, such as DF, MI, PMI, CHI, etc. We experiment using four feature scoring criteria used in image categorization.

2.3.1. Document Frequency (DF)

DF is the number of images in which a visual word occurs. The visual words with small DF are usually non-informative for category prediction. We choose visual words with DF above a predefined threshold.

2.3.2. Mutual Information (MI)

MI has been used as a criterion for feature scoring in image classification. It can be used to characterize both the relevance and redundancy of features and measure the dependence between two random features. This measure calculates the number a visual word contributes to making the correct classification decision on image class c.

The MI between a visual word vw and a class label c can be calculated by using the following equation:

$$\mathrm{MI}(vw,c)=\sum_{vw\in\{0,1)}\sum_{c\in\{0,1\}} p(vw,c)\log\left(\frac{p(vw,c)}{p(vw),p(c)}\right) \qquad (1)$$

We use the average of MI as MI_{avg} (vw*)* of K image classes in the dataset to compute Eq. 2.

$$MI_{avg}(vw)=\frac{1}{K}\sum_{i=1}^{K} MI(vw,c_i) \qquad (2)$$

2.3.3. Pointwise Mutual Information (PMI)

PMI is related to MI, referring to single events, whereas MI refers to the average of all possible events. It is used to measure the association between a visual word vw and a class label c, as in Eq. 3.

$$PMI(vw,c)=\log\left(\frac{p(vw,c)}{p(vw),p(c)}\right) \qquad (3)$$

We use the average of PMI as PMI_{avg} (vw) of K image classes in the dataset to compute Eq. 4.

$$PMI_{avg}(vw)=\frac{1}{K}\sum_{i=1}^{K} PMI(vw,c_i) \qquad (4)$$

2.3.4. χ^2Statistics (CHI)

The χ^2 test is used to test the independence of two events in statistics. Thus, we use it to test whether the occurrence of a specific visual word and the occurrence of a specific class are independent in image classification. Thus, we rank the quantity for each visual by their score. Let $\chi^2(vw,c_i)$ be the CHI between a specific visual word vw and label of an image class ci. We use the average of K image classes in the dataset as $\chi^2(vw)= \frac{1}{K}\sum_{i=1}^{K}\chi^2(vw,c_i)$.

The results in [10] indicated that CHI significantly outperforms MI, PMI and DF. However, all those feature scores were incapable of image classification for class imbalanced datasets and different appearance visual words in each class. Those scores are more interested in the large classes than small classes. Therefore, informative visual words in large classes have a higher chance to be selected than visual words in small classes, whose performances were also shown to be low. Furthermore, those scores omitted variation of the visual words in the same scenes and objects, which is important to improve performance for image classification.

Therefore, we propose a feature score of visual words from a class imbalanced dataset and different views of scenes and objects. Our approach aims to take these variations into account in the visual word vocabulary reduction process.

3. RESEARCH METHOD

The proposed feature score is based on the assumption that we can evaluate the difference of occurrence frequency of visual words between two specific image classes and other image classes as a consequence of the potential to use this difference as the feature score.

We solve visual word selection for different views of scenes and objects in each image class by grouping images in each class in the preliminary process. Therefore, we use cluster analysis to group a set of images in each class such that images in the same group have a more similar view than those in other groups. This operator performs clustering using the Expectation Maximization (EM) algorithm [30]. EM clustering is performed to estimate the means and standard deviations for each cluster to maximize the likelihood of the observed data and attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters.

Let D be the set of image data and $D=\{D_1,D_2...D_{|L|}\}$, where $l \in L$; a visual word vocabulary $V = \{vw_1,\ldots,vw_{|V|}\}$ be the set of representatives to D; $C=\{c_1,c_2...c_k\}$ be the set of K image classes of all images in dataset D; each image class c_i be grouped with the EM algorithm to subclass$c_{i,j}$, where $c_i=\{c_{i,1},c_{i,2},...,c_{i,j}\}$.

The images are converted into visual word occurrences by the Term Frequency and Inverse Document Frequency (tf-idf) weights [10]. The tf-idf weight can represent the semantic content of the images. A high weight in tf-idf is achieved by a high visual word frequency and a low image frequency of the visual word in the image dataset; the weights hence tend to filter out non-informative visual words.

The tf-idf weight is defined as $tfidf(vw_i, d_j) = tf(vw_i, d_j) \times idf(vw_i)$. The term frequency (tf) is the frequency of a visual word in an image. We use the normalized frequency of visual word vw_i of image d_j, defined as $tf(vw_i,d_j)=\frac{f_{vw_i,d_j}}{\sum_i f_{vw_i,d_j}}$, where f_{vw_i,d_j} is the number of occurrences of visual word vw_i in image

d_j. An inverse document frequency (idf) factor is incorporated, which decreases the weight of terms that occur very frequently in the image set and increases the weight of visual words that occur rarely. The idf is defined as $\text{idf}(vw_i) = \frac{|D|}{n_{vw_i}}$, where n_{vw_i} is the number of images in which visual word vw_i occurs. The tf-idf weights are normalized by cosine normalization, defined as $\text{tfidf}(vw_i, d_j) = \frac{\text{tfidf}(vw_i, d_j)}{\sqrt{\sum_{i=1}^{|V|}(\text{tfidf}(vw_i, d_j))^2}}$.

Hence, let $w(vw_i, d_j)$ be the tf-idf weight of visual word vw_i in image d_j. An image dj is represented as a feature weight vector $\overline{d_j} = [w(vw_1, d_j), \ldots, w(vw_{|V|}, d_j)]$.

We use the statistical t-test technique, which is a commonly used method for statistical evaluation of the difference between two samples means [31], to solve the class imbalance. The t-test can be used to determine whether both dataset sizes are tremendously unequal through analysis means, standard deviations and the assumption that both distributions are normal and both variances are unequal.

The t-test technique determines the significant difference of means of the tf-idf weight of a visual word between a specific subclass and other subclasses. The basic idea is that a visual word, whose mean tf-idf weight among the image in a specific subclass is significantly higher than that of other subclasses, is a highly discriminative visual word because it contains higher information about a specific subclass. This is a proposed t-test score, which is defined as follows:

$$\text{tscore}_{\text{subclass}}(vw_i, c_{k,j}) = \frac{\overline{w}(vw_i, c_{k,j}) - \overline{w}(vw_i, c'_{k,j})}{S_{\text{subclass}}} \quad (5)$$

where $\overline{w}(vw_i, c_{k,j})$ is the sample mean of the tf-idf weight of visual word vw_i of an image in a specific subclass $c_{k,j}$ in class c_k, where $c_{k,j} \in c_k$ and $c_k \in C$, and $\overline{w}(vw_i, c'_{k,j})$ is the sample mean of the tf-idf weight of visual word vw_i of an image in the other subclass, $c'_{k,j} = C - c_{k,j}$. S is the standard deviation of the two subclasses, which is calculated as follows:

$$S_{\text{subclass}} = \sqrt{\frac{S^2(vw_i, c_{k,j})}{N_{c_{k,j}}} + \frac{S^2(vw_i, c'_{k,j})}{N_{c'_{k,j}}}} \quad (6)$$

where $S^2(vw_i, c_{k,j})$ is the standard deviation of the tf-idf weight of visual word vw_i of an image in a specific subclass $c_{k,j}$ in class c_k, and $S^2(vw_i, c'_{k,j})$ is the standard deviation of the weight of vw_i in the other subclass. $N_{c_{k,j}}$ is the number of images in subclass $c_{k,j}$ and $N_{c'_{k,j}}$ is the number of images in the other subclass.

The high t-test score of visual word vw_i indicates higher discriminating power due to it having a statistically significant difference in the occurrence frequency in a specific subclass $c_{k,j}$, compared with the other subclass.

The t-test score is locally specified with respect to a specific subclass $c_{k,j}$. To globally assess the value of a visual word vw_i in each class c_k, we solve the following equation:

$$\text{tscore}(vw_i, c_k) = \sum_{j=1}^{|c_k|} \text{tscore}_{\text{subclass}}(vw_i, c_{k,j}) \quad (7)$$

We calculate three t-test scores in three alternate ways to combine the class with a specific score: the Max-tscore ($\text{tscore}_{\max}(vw_i)$) purposes to merit the maximum significance of a visual word occurring in one class against other classes; the Average-tscore ($\text{tscore}_{\text{avg}}(vw_i)$) uses equal weights of all classes, with an inattentive number of images belonging to it; the Weighted Average-tscore ($\text{tscore}_{\text{wavg}}(vw_i)$) applies the average of the F1 measure of each class, with its weight varying with its size. The three t-test scores are defined as follows:

$$\mathrm{tscore}_{\max}(vw_i) = \max_{k=1}^{|C|} \mathrm{tscore}(vw_i, c_k) \tag{8}$$

$$\mathrm{tscore}_{\mathrm{avg}}(vw_i) = \frac{\sum_{k=1}^{|C|} \mathrm{tscore}(vw_i, c_k)}{|C|} \tag{9}$$

$$\mathrm{tscore}_{\mathrm{wavg}}(vw_i) = \sum_{k=1}^{|C|} P(c_k) \times \mathrm{tscore}(vw_i, c_k) \tag{10}$$

4. RESULTS AND ANALYSIS

4.1. Dataset

We use two datasets to experiment the feature scoring of visual words and its use in image classification: the Paris dataset [32] and the SUN397 dataset [33].

The Paris dataset contains 6,300 high resolution (1024 × 768) images obtained from Flickr by querying the associated text tags for famous Paris landmarks, such as “Eiffel Tower Paris” or “Louvre Paris”. This dataset consists of 12 landmark scenes in Paris. Each landmark scene has different numbers of images ranging from as few as ~150 images for ‘‘Eiffel Tower’’ to ~1400 images for “General Paris”. The distribution of images in classes is imbalanced. Average, Std. dev., and Coefficient of Variance (CV) of the number of images of the classes are 525, 311.68 and 1.51, respectively. The example images from this dataset are shown in Figure 2.

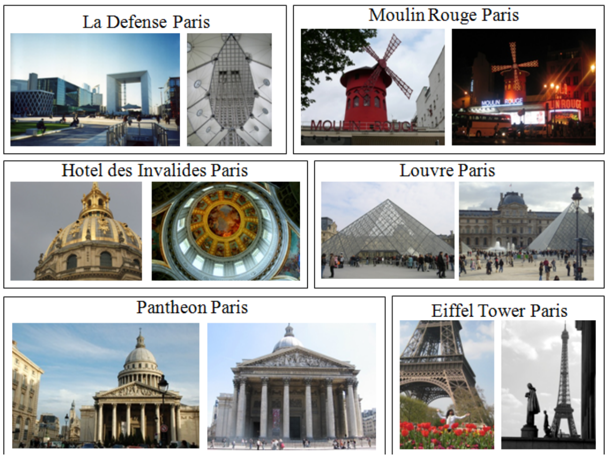

Figure 2. The example images of the Paris dataset

The SUN397 dataset includes the extensive Scene UNderstanding (SUN) database, which contains 397categories and 130,519 images (200 x 200). Examples of categories include “abbey”, “grotto”, “ossuary”, “salt plain”, “signal box”, “sinkhole”, “sunken garden” and “winners circle”. Each category has a different number of images, ranging from as few as ~100 images to ~2300 images. The distribution of images in classes is imbalanced. Average, Std. dev., and Coefficient of Variance (CV) of the number of images of the classes are 271.79, 251.62 and 1.50, respectively. The example images from this dataset are shown in Figure 3.

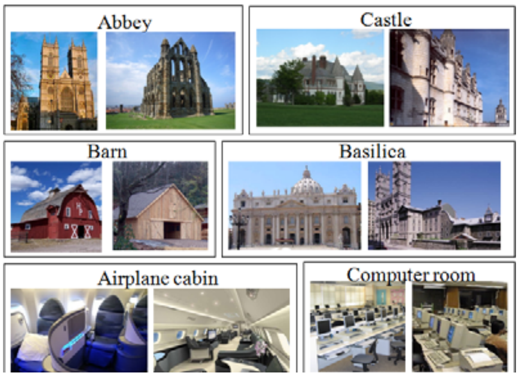

Figure 3. The example images of the SUN397 dataset

4.2. Vocabulary Size

We examine to select the best size of visual word vocabulary for each dataset; choosing the right vocabulary size involves the performance of each image. We experiment using binary BOVW in each image. Let $BOVW(vw_i,d_j)$be a bag of visual word vw_i of image d_j. If visual word vw_i occurs in an image d_j, then $BOVW(vw_i,d_j)=1$; otherwise, $BOVW(vw_i,d_j)=0$. We use a 100-visual word vocabulary to 50,000-visual words vocabulary in BOVW to classify the Paris dataset and the SUN397 dataset using the Support Vector Machines (SVM). Figure 4 shows the relationship between the classification performance and the size of the visual-word vocabulary. We use the F1-measure to evaluate the classification performance. The optimal visual word vocabulary size is an approximately 5,000-visual words vocabulary for the Paris dataset and an approximately 20,000-visual words vocabulary for the SUN397 dataset. Therefore, we examine feature score techniques on the optimal visual word vocabulary size for each dataset.

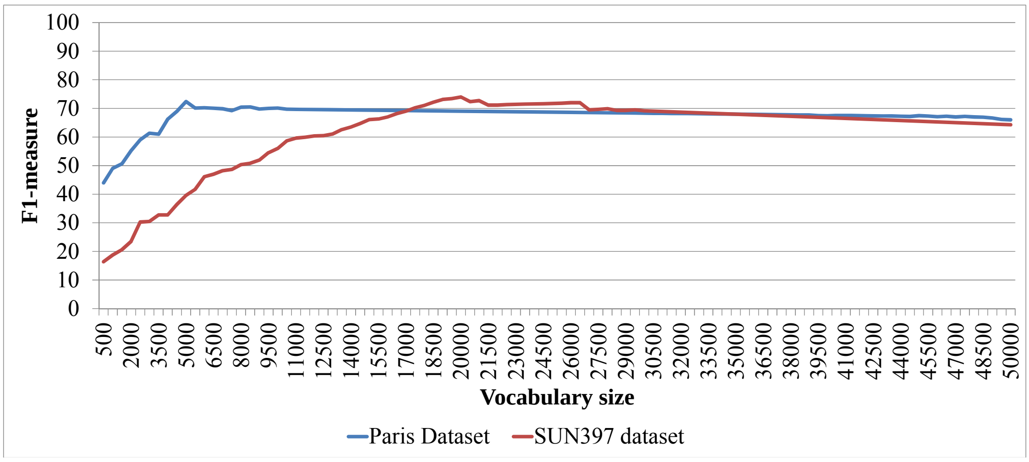

Figure 4. The classification performance for different sizes of the visual-word vocabulary on two datasets

4.3. Classifier

The performance of image classification is used to evaluate the effectiveness of feature selection. We use two classifiers in the experiment, which consist of SVM with a linear kernel and Naïve Bayes. SVM finds the maximum margin hyper plane between two classes by using the training data and applying an optimization technique. The decision boundary is defined by a sub-set of the training data, the so-called support vectors. SVM with a linear kernel has shown good generalization performance and is robust on highly dimensional image classification. Although the Naïve Bayes classifier suffers from lower accuracy compared to the SVM classifier, it makes it easy to estimate the probability that a sample belongs to a particular class. The experiment is performed using the LibSVM package and Naïve Bayes package of WEKA [34], with the default values of parameters.

4.4. Evaluation

We use the F1 measure to aggregate the performance of multiple classifiers. It examines both the precision and recall of the test to compute the score: precision is the number of correct positive results divided by the number of all positive results, and recall is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. Calculation of the F1 measure is defined as:

$$F1=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}} \tag{11}$$

4.5. Results

4.5.1. Performance on SVM and Naïve Bayes

The number of subclasses grouped using the EM algorithm is in the range from two subclasses to five subclasses for the two datasets. We compare the performance of image classification via 10 feature score methods: 1) Document Frequency (DF), 2) Mutual information (MI), 3) Pointwise Mutual information (PMI), 4) Chi-square statistics (CHI), 5) Max-tscore, 6) Average-tscore, 7) Weighted Average-tscore, 8) Max-tscore with subclass (Max-tscore-sub), 9) Average-tscore with subclass (Average-tscore-sub), 10) Weighted Average-tscore with subclass (Waverage-tscore-sub).

Figure 5 to Figure 8 show the F1 measure results, which were classified by using Support Vector Machines and Naïve Bayes and are used on the Paris dataset and SUN379 dataset, respectively.

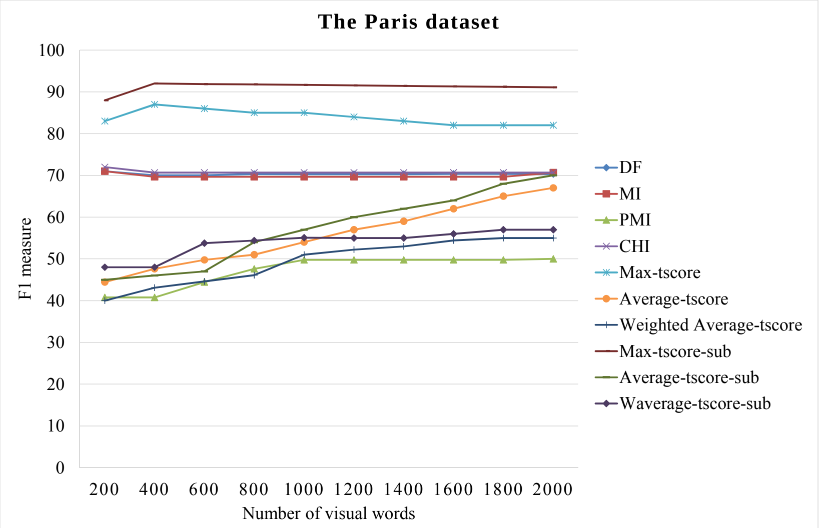

Figure 5. F1 measure result using SVM on the Paris dataset

The Paris dataset (Figure 5) shows that the F1 performances of SVM using MAX-tscore-sub and MAX-tscore are superior to those of other feature scores. Figure 6 shows the F1 measure of Naïve Bayes, where MAX-tscore-sub outperforms the other feature scoresand Weighted Averaged-tscore shows results that are slightly lower than those of the other feature scores; PMI shows very low performance for image classification. The results of Weighted Averaged-tscore are the worst because it is an average of the tscore and the weight of each class varies with its size. However, Averaged-tscore is not so. The F1 measure of SVM based on Max-tscore-sub is the highest (92.00) when the number of selected features is 400, and the F1 measure of Naïve Bayes using Max-tscore-sub reaches maximum (89.00) when the number of selected visual words is 600.

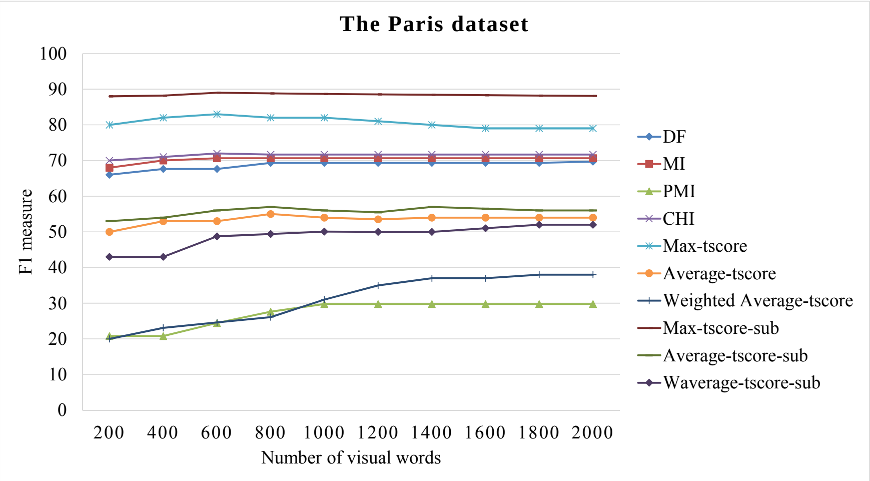

Figure 6. F1 measure result using Naïve Bayes on the Paris dataset

For both classifiers on the SUN397 dataset, Max-tscore-sub offers the best F1 measure results. The F1 measure of Max-tscore-sub using SVM is highest (94.00) when the number of selected features is 1800, as shown in Figure 7, and the F1 measure of Max-tscore-sub using Naïve Bayes is highest (91.00) when the number of selected visual words is 2400, as shown in Figure 8.

Therefore, Max-tscore-sub has significant classification performance on the two datasets and with the two classifiers because the feature score aims to maximize the significance of feature relevance in one subclass in each class against other classes rather than using the average of the feature score for all classes. Thus, the feature score is suitable for situations where the class sizes are quite different and multi-views exist in each class.

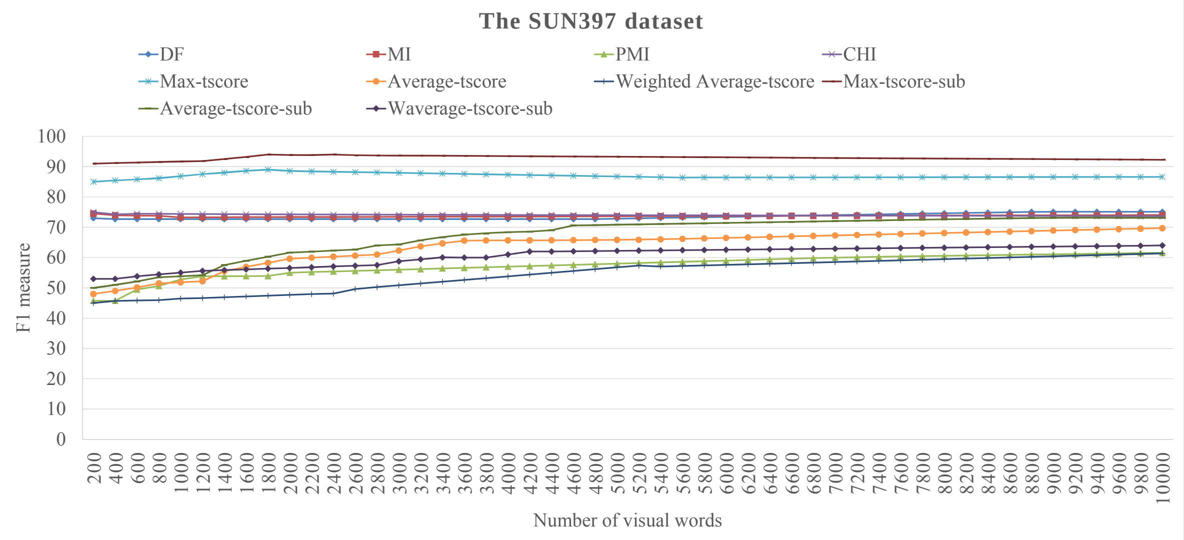

Figure 7. F1 measure result using SVM on the SUN379 dataset

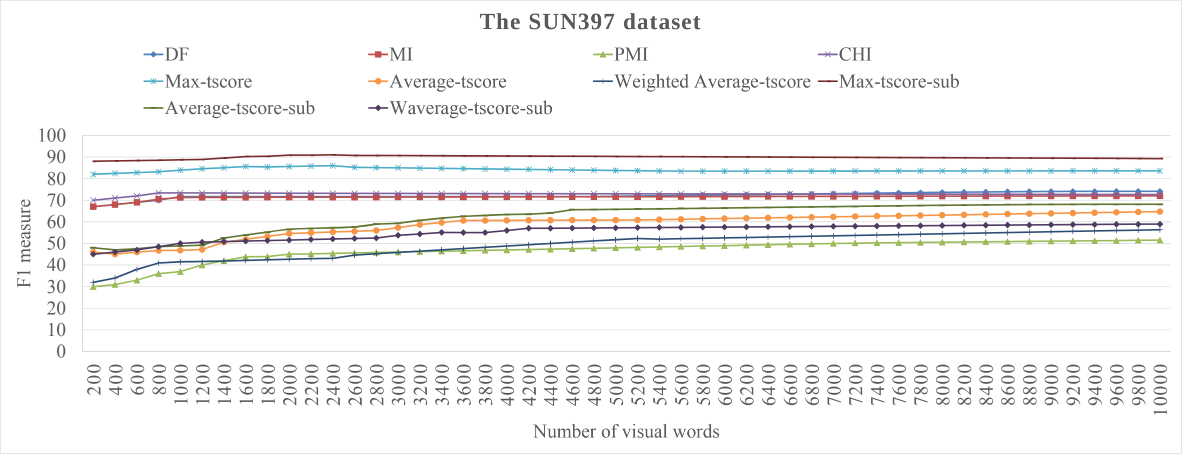

Figure 8. F1-measure result using Naïve Bayes on the SUN379 dataset

4.5.2. Quality of Selected Features

We divide the two datasets into three groups of classes based on size: majority, moderate, and minority. Figure 9 and Figure 10 show the averages of F1 measures of the Paris dataset and SUN397 dataset when various feature scores are used to consider the effect on the classification of majority, moderate and minority classes. We compare the Max-tscore-sub with other feature scores and use the SVM classifier.

However, our experiment shows that the Max-tscore-sub clearly outperforms all feature scores under all classes, especially very small size classes (minority). We conclude that our score increases the F1 measure of the minority class without sacrificing the F1 measure of the majority class.

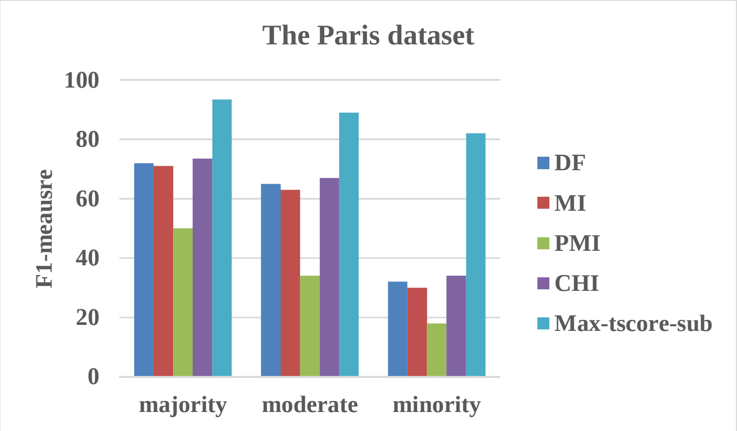

Figure 9. The F1 measure result of the SVM classifier for each group class on the Paris dataset when various feature scores are used

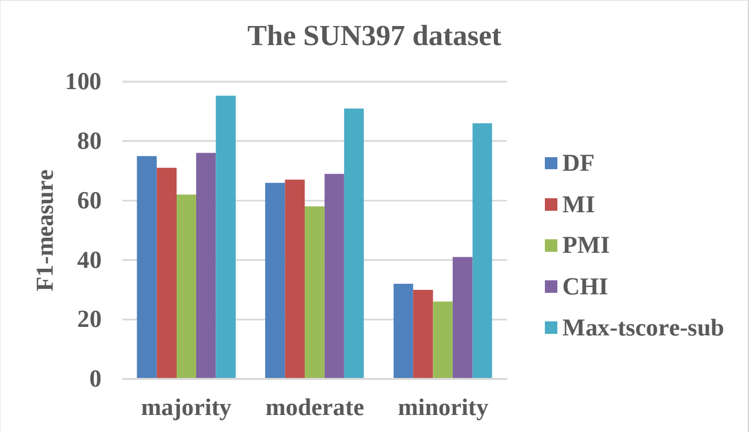

Figure 10. The F1 measure result of the SVM classifier for each group class on the SUN379 dataset when various feature scores are used

5. CONCLUSION

We proposed three feature scores for ranking-based feature selection in imbalanced multi-class image classification and multi-views in each class. The proposed feature score can reduce the amount of BOVW stored from each image subclass while still maintaining strong classification performance. We proposed feature scores based on the statistical t-test technique, which evaluates the difference of means of visual word occurrence frequencies between subclasses in each class and between other classes as a usefulness measure of individual visual words. The t-test is used to determine whether two class sizes are equal. It assumed that both distributions are normal and that both variances are unequal. Moreover, its function normalized the mean difference value by the common standard deviation of two classes. Therefore, the t-test score is insensitive to the imbalanced class distribution. Experiments showed that one of the proposed scores, Max-tscore-sub, which is based on maximizing the significance of visual word relevance for all subclasses in each class and all classes, has high performance on the two datasets. Moreover, the F1 measure results of minority classes are improved. The proposed feature score can be used instead of the other feature scores for imbalanced sizes of classes.

ACKNOWLEDGEMENTS

This research is supported by a grant from the Kasetsart University Research and Development Instituteand Ministry of Science and Technology, Thailand.

REFERENCES

[1] J. Sivic, et al, “Video Google: a text retrieval approach to object matching in videos”, in *Ninth IEEE International* Conference on Computer Vision, 2003, pp. 1470-1477.

[2] M. Everingham, et al, “The Pascal Visual Object Classes Challenge: A Retrospective”, *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98-136, 2015.

[3] J. Xiao, et al, “SUN database: Large-scale scene recognition from abbey to zoo”, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3485-3492.

[4] H. Lei, et al, “Learning group-based dictionaries for discriminative image representation”, *Pattern Recognition*, vol. 47, no. 2, pp. 899- 913, 2014.

[5] Q. Tian, et al, “Building descriptive and discriminative visual codebook for large-scale image applications”, *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 441-477, 2011.

[6] D. Yi, et al,“Video Object Matching Based on SIFT and RotationInvariant LBP”, *TELKOMNIKA*, vol. 11, no. 10, pp. 5876-5883, October 2013.

[7] Y. Xie, et al, “A Novel Specific Image Scenes Detection Method”, *Multimedia Tools Appl.*, vol. 74, no. 1, pp. 105-122, Jan. 2015.

[8] X.C. Lian, et al, “Probabilistic models for supervised dictionary learning”, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2305-2312.

[9] B. Fernando, et al, “Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation”, *Pattern Recognition*, vol. 45, no. 2, pp. 897-907, Feb 2012.

[10] J. Yang, et al, “*Evaluating bag-of-visual-words representations in scene classification*”, In Proceedings of the international workshop on Workshop on multimedia information retrieval (MIR '07), 2007, pp.197-206.

[11] S. Li, et al, “An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine”, *Knowledge-Based Systems*, vol.24, no.1, pp. 40-48, 2011.

[12] E. I Sela, et al, “Feature Selection of the Combination of Porous Trabecularwith Anthropometric Features for Osteoporosis Screening”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 1, pp. 78-83, 2015.

[13] Y. Yang, et al, "*A Comparative Study on Feature Selection in Text Categorization*", In Proceedings of 14th International Conference on Machine Learning, 1997, pp. 412-420.

[14] S. Samanta, et al, “*A Fast Supervised Method of Feature Ranking and Selection for Pattern Classification*”, In Proceedings of International Conference on Pattern Recognition and Machine Intelligence (PreMI), 2009, pp. 80-85.

[15] S. Fakhraei, et al, “Bias and stability of single variable classifiers for feature ranking and selection”, *Journal Expert Systems with Applications, ACM*, vol.41, no.15, pp. 6945-6958, 2014.

[16] P. Soda, “*A Hybrid Approach Handling Imbalanced Datasets*”, In Proceeding of 15th International Conference Image Analysis and Processing, 2009, pp.209-218.

[17] J.P. Hwang, et al, “A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function”, *Expert Systems with Applications*, vol.38, no.7, 2011.

[18] P. Pramokchon, et al, "A feature score for classifying class-imbalanced data", In Prodeeding of International Computer Science and Engineering Conference (ICSEC), 2014, pp.409-414.

[19] K. Mikolajczyk, et al, “A comparison of affine region detectors”, *International Journal of Computer Vision*, vol.65, no.43-72, 2005.

[20] T. Tuytelaars, et al, “Local invariant feature detectors: a survey”, *Foundations and Trends in Computer Graphics and Vision*, vol.3, no.3, pp.177-280, 2008.

[21] T. Tuytelaars, et al, “Matching widely separated views based on affine invariant regions”, *International Journal of Computer Vision*, vol.59, no.1, pp. 61-85, 2004.

[22] K. Mikolajczyk, et al, “A performance evaluation of local descriptors”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.10, pp.1615-1630, 2005.

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

[24] G. Csurka, et al, “Visual categorization with bags of keypoints”, In *ECCV: Workshop on Statistical Learning in Computer Vision*, 2004.

[25] J.R.R. Uijlings, et al, “*What is the spatial extent of an object?*”, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp.770-777.

[26] J. Philbin, et al, “*Object retrieval with large vocabularies and fast spatial matching*”, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[27] G. Schindler, et al, “*City-scale location recognition*”, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp.1-7.

[28] S. Zhang, et al, “*Descriptive visual words and visual phrases for image applications*”, In Proceedings of the ACM International Conference on Multimedia, 2009, pp.75-84.

[29] P. Turcot, et al, “*Better matching with fewer features: The selection of useful features in large database recognition problems*”, In Proceedings of the 17th ACM international conference on Multimedia, 2009, pp.75-84.

[30] R.A. Redner, et al, “Mixture Densities, Maximum Likelihood and the Em Algorithm”, *SIAM Review*, vol. 26, no. 2, pp. pp. 195-239, 1984.

[31] A.C. Tamhane et al, Statistics and Data Analysis: Prentice Hall, 2000.

[32] J. Philbin, et al, “The Paris Dataset”, Department of Engineering Science, University of Oxford, http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/, accessed Aug. 20. 2014.

[33] SUN dataset, http://groups.csail.mit.edu/vision/SUN/, accessed Aug. 27. 2014.

[34] Weka, Machine Learning Group at the University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/, accessed Sep. 27. 2014.

BIOGRAPHIES OF AUTHORS

Sutasinee Chimlek, she received her B.Sc. in Computer Science from Chiang Mai University in 1996 and her M.Sc. in Information Technology from King Mongkut’s Institute of Technology Ladkrabang in 2001. She is currently lecturer in department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand. Her research interests include pattern analysis and multimedia processing.

Part Pramokchon, he received his B.Eng. and M.Eng. in Computer Science from Chiang Mai University in 1999 and2003, respectively and his D.Eng. in Computer Engineering at Kasetsart University (KU)in 2014. He is currently lecturer in department of Computer Science, faculty of Science, Maejo University, Chiang Mai, Thailand. His research interest is pattern recognition and multimedia processing.

Punpiti Piamsa-nga, he received his B.Eng. and M.Eng. in Electrical Engineering from Kasetsart University (KU) in 1989 and 1993, respectively and his D.Sc. in Computer Engineering from George Washington University in 1999. He is currently an associate professor in computer engineering at KU. His research interest is pattern recognition and multimedia processing.