Fusion of tf.idf weighted bag of visual features for image classification

Christophe Moulin, Cécile Barat, Christophe Ducottet
Université de Lyon, F-69003, Lyon, France
Université de Saint-Étienne, F-42000, Saint-Étienne, France
CNRS UMR5516, Laboratoire Hubert Curien
{christophe.moulin, cecile.barat,ducottet}@univ-st-etienne.fr

Abstract

Image representation using bag of visual words approach is commonly used in image classification. Features are extracted from images and clustered into a visual vocabulary. Images can then be represented as a normalized histogram of visual words similarly to textual documents represented as a weighted vector of terms. As a result, text categorization techniques are applicable to image classification. In this paper, our contribution is twofold. First, we propose a suitable Term-Frequency and Inverse Document Frequency weighting scheme to characterize the importance of visual words. Second, we present a method to fuse different bag-of-words obtained with different vocabularies. We show that using our tf.idf normalization and the fusion leads to better classification rates than other normalization methods, other fusion schemes or other approaches evaluated on the SIMPLIcity collection.

1 Introduction

Classification of images is a challenging problem of many image processing and computer vision applications. The problem of image representation is crucial to the effectiveness of the classification system. This problem is often handled by computing low level features, which are processed with a classifier engine for inferring high-level information about the image.

In [1][19], histograms of color, texture and edge directions features have been used for various categorization problems. Although they are computationally effective, histograms provide a global information and are a crude representation of the image content. Region-based approaches have been proposed [6][2]. They consist in segmenting an image into regions and computing features on each of them. The image representation is then the collection of all these local descriptors. These methods are robust to partial occlusion, but are sensitive to inaccurate segmentation.

The trend in image classification is towards the use of patch-based representations. A patch is a small subimage centered on a pixel and characterized by its local visual properties. Patches can be sampled densely [5][3], randomly [20] or detected with various detectors [9][18]. They are characterized using some descriptors such as SIFT [8], color [22] or MPEG features [18]. A vocabulary is learned over all patches of the images of a collection. Patches are grouped into clusters according to a similarity measure. Each cluster gives a visual word that represents the local pattern shared by the patches within this cluster. To represent an image, patches are mapped into visual words leading to a 'bag-of visual words'. A 'bag-of-visual words' is a normalized histogram of visual words used as feature vector in the classification task. This representation comes from texton methods in texture analysis [4] and is analogous to bag-of-words representation of text documents [16]. As a result, techniques for text categorization are applicable to image classification.

Our work exploits the bag-of-visual words approach. We study different image representation choices, including the bag-of-word normalisation (i.e. the normalisation of the visual words histogram) and the combination of different bag-of-words describing a same image with respect to specific characteristics (color, shape or texture). Our contribution is twofold. First, we use a suitable Term-Frequency and Inverse Document Frequency weighting, a highly effective technique in document retrieval, to represent images and we compare with other weighting strategies presented by Nowak *et al.* in [11] and Yang *et al.* [23]. Second, we propose an approach to fuse different bag-of-visual words and show that the fusion of specific bag-of-visual words is an effective fusion strategy compared to the use of each bag-of-words separately or the use of early or late fusion schemes. Through our study, we use multiscale dense sampling to select patches which offers a better coverage of the image content than keypoint detectors [11]. The presented work is de-

978-1-4244-8027-2/10/$26.00

veloped in conjunction with the GDR ISIS research group [1] involved in indexation and multimedia information retrieval (Theme B, axis 3).

The paper is organized as follows. In section 2, we discuss related work and our contribution. In section 3, we detail the improvements we propose. Experimental settings and results are given in section 4 and 5. We conclude in section 6.

2 Related work and our contributions

There have been many works using the bag-of-visual words representation for image classification [23][3]. Several studies have also been conducted to analyze the influence of the different parameters involved in the bag-of-words process. These parameters include the choice of the patch detector and descriptor [9], the clustering algorithm for the vocabulary creation [3], the number of patches per image, the number of visual words, the histogram normalization process [11] and the classifier engine to use [1]. Here, we only focus on recent work on patches sampling, histogram normalization and descriptors fusion.

Recent work in the field of image categorization indicates that the best performance are obtained using dense sampling [3], especially when considering large vocabularies [11].

In the bag-of-words approach, an image is represented using an histogram of visual words which counts the number of occurrences of the different visual words in the given image. To enhance the discriminating power of the visual words, several histogram normalization strategies have been discussed in the literature. In [11], Nowak *et al.* compare the use of an unnormalized histogram with two normalization methods which binarize the histogram. The first method consists in returning a vector indicating the presence or absence of words. The second one adaptively selects a binarization threshold for each word by maximizing the mutual information between a word and a class. In [23], the authors work with keypoints and propose to weight the visual words using a Term Frequency- Inverse Document Frequency approach. Each weight is computed to enhance the importance of a word in an image by multiplying tf and idf, where tf is the term frequency that characterizes the frequency of a word in the given image. The idf is the inverse document frequency that quantifies the importance of a word over the corpus of images. The weight is high when the word is frequent in the image but rare in the others. The authors conclude that binary visual words methods are as effective as tf.idf ones. However, this tf.idf approach hides different formulations providing different classification results. Authors do not specify the one they use. Our first contribution is to show that applying the efficient Okapi formula leads to better classification results than binary or regular tf.idf methods.

Another question of interest is how to enhance several local properties of patches, such as color, texture or shape. Most of the methods resort to early fusion. They consist in stacking the different measured descriptors into a unique vector to describe each patch. Thank to this fusion only one fused modality is used to represent a document. The vocabulary is then learned from all vectors. In contrast, late fusion methods, which started with the representation of multimedia documents, fuse the results of different classifiers working on each type of features [17]. We propose an intermediate approach. It consists first in defining several vocabularies characteristic of color, shape or other feature. Second, given an image, a bag-of-word is computed for each vocabulary. The different bag-of-words are combined before classification. This approach is shown to provide better classification results than using each feature separately or using early or late fusion schemes.

3 The proposed improvements

This section describes the image representation model we propose based on the use of one or several tf.idf weighted bag-of-words. After presenting our choices for patches sampling, description and vocabulary creation , we detail our two proposed improvements : the tf.idf weighting process and the fusion of bag-of-visual words resulting from different vocabularies.

3.1 Sampling, description and vocabulary construction

As recently dense sampling gave good results for object recognition [12], we choose to adopt this approach. Let us define a patch as a squared region with scale s, i.e. the side length of the square, centered at a given pixel of the considered image. We define a dense sampling as a regular sampling of patches spatial position in space and regular sampling of patches scale, authorizing spatial overlapping between patches. The spatial overlapping intends to make the sampling independant of the spatial position of the scene content. Scales are taken as multiples of the finest scale $s_0 = 12 \times 12$, such as $s = is_0, i = 1, 2, 3, \ldots$. At scale $s = is_0$, patches are sampled every i pixels in order to keep the overlap rate between patches constant. We randomly select only n patches among all to describe an image.

Once patches are extracted, we consider two types of descriptors to enhance their color and texture information. For color information (MM), we transform the RGB components of the patch into normalized components defined as $\frac{R}{R+G+B}$, $\frac{G}{R+G+B}$ and $\frac{R+G+B}{3\times255}$. This color space presents

[1] http://gdr-isis.org/

two main advantages. First, it makes the first two variables independent of the third one representing the brightness. Second, it is very easy to compute. From the normalized components of a patch, we compute 6 features equal to the mean and the standard deviation of the values. The second description (SM) of patch is the well known SIFT descriptor based on histograms of gradient orientation [8].

For each description, we learn a visual vocabulary V applying a k-means algorithm over all the computed patches. We get k clusters of features whose means represent k visual words, k being the size of the visual vocabulary. Local patches of images are mapped to their closest visual words using the euclidean distance. An image is then described as a $tf.idf$ normalized histogram of visual words.

3.2 Tf.Idf histogram normalization

Term weighting is a key method in the context of text classification. As in the vector space model introduced by Salton *et al.* to represent text document [16], we represent an image d_i as a vector of weights. Let $D = \{d_1, \ldots, d_i, \ldots, d_{|D|}\}$ be the image documents of the collection and $V = \{v_1, \ldots, v_j, \ldots, v_{|V|}\}$ be the visual vocabulary created using a bag of visual words approach. As explained in Section 2, the weights $w_{i,j}$ are calculated by multiplying $tf_{i,j}$ and idf_j. Let's remind that a $tf_{i,j}.idf_j$ weight is high when the visual word v_j is frequent in the image d_i but rare in the others.

Several formulations exist to calculate these $tf_{i,j}$ and idf_j, but one of the most efficient is the okapi one proposed by Robertson *et al.*[14]. We apply a modified version implemented in the lemur software[2] proposed by [24]:

$$tf_{i,j} = \frac{k_1 n_{i,j}}{n_{i,j} + k_2(1 - b + b\frac{|d_i|}{d_{avg}})}$$

where $n_{i,j}$ is the occurrence of the word v_j in the image d_i, $|d_i|$ the number of visual words used to represent image d_i, d_{avg} the average number of visual word per images in the collection D. k_1, k_2 and b are three constants.

In our experiments, as the number of words per image is the same for each image, $|d_i|$ equals to d_{avg}. The tf formula can be simplified as:

$$tf_{i,j} = \frac{n_{i,j}}{n_{i,j} + 1}$$

with k_1 and k_2 equal to 1.

$$idf_j = \log \frac{|D| - |\{d_i | v_j \in d_i\}| + 0.5}{|\{d_i | v_j \in d_i\}| + 0.5}$$

It can be noticed that the idf term of this formula can be possibly negative, which has been discussed in [13]. This happens when a term appears in more than half of the documents. We choose to floor the idf values to 0.

3.3 Fusion using different bag-of-words

Images are represented with two specific vocabularies enhancing color and texture information. As a result, we get two bag-of-words per image. The question is how to combine these two bag-of-words in order to better describe the image content and improve the classification results. As recalled in section 2, several strategies exist for fusion. As in [18], we apply a simple merging fusion of our two specific $tf.idf$ weighted histograms. Instead of giving each bag-of-words as feature vector to the classifier, we give the new created unique vector. The fusion relevancy will be presented in the following, comparing results obtained using the modalities separately, those obtained with an early and a late fusion method and ours obtained with the fused vector of both modalities.

4 Experimental settings

The goal of this section is to specify our experimental parameters. We first present the collection used for the evaluation of our image representation. Second, we detail the values of the different parameters involved in the bag-of-words construction. Finally, we give some information about the classifier engine used and the evaluation measure.

4.1 Description of the collection

We choose to perform experiments on the SIMPLIcity collection[3] proposed by Wang [21] which contains 1000 images extracted from the COREL database. This collection is composed of 10 meaningful categories: *African people, beaches, buildings, buses, dinosaurs, elephants, flowers, food, horses and mountains*. Each category is composed of 100 images. All the images contain 384×256 pixels and are represented using bag of visual features. One image per category is presented on Figure 1.

This collection is often used to evaluate classification methods [15][10]. In the literature, classification rates vary between 70% and 86%.

4.2 Image descriptors

As said previously, to describe an image d_i, patches are extracted using dense sampling. At the finest scale ($s = 1$), a patch is extracted at each pixel and the patch size is equal to 12×12 pixels. Not all patches are retained, but only n

[2] http://www.lemurproject.com

[3] http://wang.ist.psu.edu/~jwang/test1.zip

Figure 1. Examples extracted from the SIMPLIcity collection.

ones which are randomly selected to represent d_i. In our experiments, n varies from 500 to 5000.

For both the SM and MM descriptors, a vocabulary is created applying a k-means algorithm. The value k defines the size of the vocabulary. We work with k equals 1000 and 5000, which correspond to suitable values as explained in [11].

Each image is finally represented as a normalized histogram. We compare the regular $tf.idf$ histogram normalization, the Okapi $tf.idf$ one and the binary normalization. In the latter case, the $tf.idf$ value is simply replaced by 1.

4.3 Classification with SVM and evaluation measure

A L2-regularized logistic regression is chosen for the classification [7]. We perform a leave-one-out cross-validation (LOOCV) to evaluate the classification performance. In this approach, a single image from the collection is used for the test and the remaining images are taken as the training data. This process is repeated such that each image in the collection serves once as the test data. Leave-one-out cross-validation is expensive from a computational point of view as the training process is repeated for each test image.

Several classification experiments are performed in order to analyze the influence of the different histogram normalizations (tfidf or binary). We also aim to study the advantage of using a fused bag-of-words combining different local descriptors rather than a monomodal bag-of-words or an early or late fusion scheme. The common classification rate is a number easy to interpret and widely used in the context of image classification. It corresponds to the number of images correctly classified divided by the total number of classified images.

5 Results

Two main results are presented below. Firstly, we compare $tf.idf$ and binary normalizations for the two sizes of vocabulary and several numbers of visual words per image, as the study conducted in [11]. Secondly, we present improvements of the classification rate using a simple merging fusion of the different bag-of-features vocabularies.

5.1 Comparison of tf.idf and binary normalization

Figure 2 allows to understand the influence of the histogram normalization on classification rate with respect to the vocabulary size, the descriptor used and the number of patches selected per image.

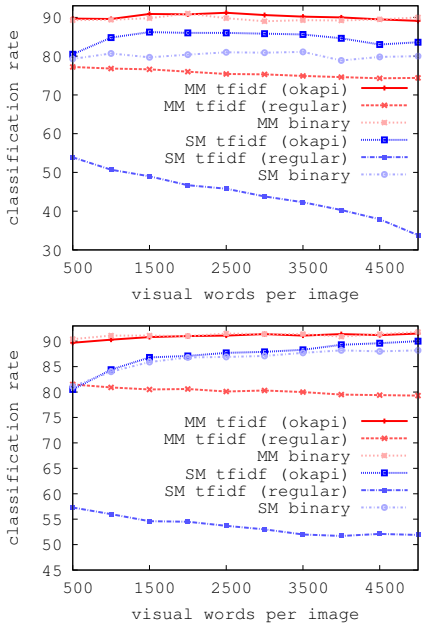

Figure 2. tf.idf versus binary normalization using a vocabulary of 1000 (left) and 5000 (right) words.

Several observations can be made. First, it is clear that we obtain better classification rates using the MM color descriptor than the SM texture descriptor, whatever the vocabulary size and the number of patches are. Second, for a vocabulary size of 1000, classification rates are almost constant with respect to the number of patches, while for a vocabulary size of 5000, classification rates increase with the number of words, whatever the normalization approach is. Third, we observe that Okapi $tf.idf$ normalization performs better that the regular $tf.idf$ one. The gap is particularly significant for the SM descriptor. Fourth, we can see that for both sizes of vocabulary and considering the SM descriptor, better results are obtained with the Okapi $tf.idf$ normalization, especially in the case of the size equal to 1000, where the mean gain is equal to 5%: the classification rate is roughly 80% with binary normalization and 85% with the $tf.idf$ normalization. Considering now the MM descriptor, Okapi $tf.idf$ normalization results are comparable to the binary ones. Nevertheless, they are slightly higher which is significant at such high rates of classification. As a result, we can say that the Okapi $tf.idf$ normalization improves the results in most of cases. Highest rates of classification are obtained with a vocabulary of 5000 words and high number of patches. Moreover, our representation method and parameters allow to reach classification rates around 90%, which is largely superior to those seen in the literature.

5.2 Fusion of different vocabularies

As said previously, our aim is to show a simple merging fusion of the different bag-of-features vocabularies enables to improve the results obtained using the modalities separately. Our fusion approach is also compared to early and late fusion ones. The early fusion consists in concatenating the vector of MM features and the vector of SM features. A $k-means$ is then applied to create a single visual vocabulary of 1000 or 5000 words and classification is perfomed using SVM. The late fusion combines outputs of 2 SVM classifiers trained independently from MM and SM feature vectors. For each input image, the decision is made for the class with the highest distance to the separation plane with the other classes.

Figure 3 presents the classification rates obtained with MM or SM $tf.idf$ bag-of-words and our fused $tf.idf$ bag-of-words. They are plotted as a function of the number of visual words and with respect to the two sizes of vocabulary. It is obvious from Figure 3 that the fusion of the two vocabularies improves significantly the classification rates. Whatever the number of visual words per image (from 500 to 5000 words) and the size of the vocabulary (1000 or 5000), the fusion of our two specific vocabularies roughly improves the best results obtained with the MM descriptor

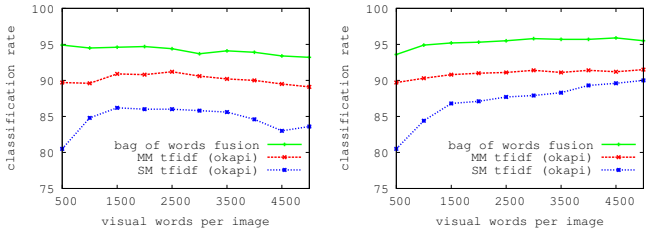

Figure 3. fusion results using a vocabulary of 1000 (left) and 5000 (right) words.

from 90% to 95%. For the vocabulary of 1000 words, we observe a decrease of classification rates when the number of words per image gets above 2500. It seems to be due to the fact that the number of words per image becomes high compared to the size of the vocabulary. Then, it becomes likely to find every visual words in an increasing number of images, reducing the interest of the idf weighting and the quality of image representation.

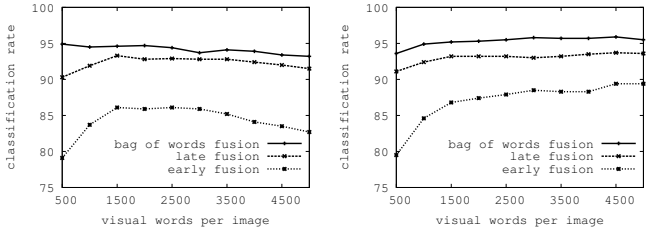

Figure 4. Comparison of three fusion schemes.

Figure 4 compares the results obtained with the three fusion schemes. The fused bag-of-words approach we propose performs better than the two other early and late methods. The early fusion returns the worst results. It can be explained by the fact that in this approach, the size of the feature space is increased, which spreads the data further apart and makes them very sparse. Distance measurements become increasingly meaningless. Clustering algorithms, such as k-means, struggle with high dimensional data.

6 Conclusion

In this article, we proposed an image representation model based on bag-of-words representation and $tf.idf$ normalization, which is well-suited to image classification. Two different bag-of-features vocabularies based on color and texture information were computed. The choice of an appropriate okapi-based tf.idf normalization process lets us to get high classification rates on the SIMPLIcity collection (up to 95%). These rates are most often superior to

those obtained with the binary normalization, and are significantly above to those encountered in the literature. Furthermore, we showed that fusing vocabularies by a vector merging approach significantly improves the classification rates.

For future work, we aim to add other oriented textual techniques used in classical text categorization, such as feature selection or feature extraction. Other fusion approaches can also be considered.

7 Acknowledgements

This work was partly supported by the LIMA project (*http://liris.cnrs.fr/lima*) and the SATTIC project (*http://labh-curien.univ-st-etienne.fr/aaaa/sattic.php*), a research program funded by the French National Research Agency (ANR).

References

[1] O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055, 1999.

[2] C. Fredembach, M. Schröder, and S. Süsstrunk. Region-based image classification for automatic color correction. In *Proc. IS&T 11th Color Imaging Conference*, pages 59–65, 2003.

[3] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society.

[4] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1265–1278, 2005.

[5] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[6] J. Li, J. Z. Wang, and G. Wiederhold. Irm: integrated region matching for image retrieval. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 147–156, New York, NY, USA, 2000. ACM.

[7] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton methods for large-scale logistic regression. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 561–568, New York, NY, USA, 2007. ACM.

[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 & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[10] M. Mouret, C. Solnon, and C. Wolf. Classification of images based on Hidden Markov Models. In *IEEE Workshop on Content Based Multimedia Indexing*, pages 169–174, 2009.

[11] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Computer Vision ECCV 2006*, pages 490–503, 2006.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[13] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24, New York, NY, USA, 1997. ACM.

[14] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at trec-3. In *Text REtrieval Conference*, pages 21–30, 1994.

[15] J. Ros, C. Laurent, and G. Lefebvre. A cascade of unsupervised and supervised neural networks for natural image classification. *Lecture Notes in Computer Science*, 4071:92, 2006.

[16] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[17] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM.

[18] E. Spyrou, H. L. Borgne, T. Mailis, E. Cooke, Y. Avrithis1, and N. OConnor. Information-Based Color Feature Representation for Image Classification. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*, volume 3697, pages 847–852, 2005.

[19] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.

[20] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 281, Washington, DC, USA, 2003. IEEE Computer Society.

[21] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. In *VISUAL '00: Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 360–371, London, UK, 2000. Springer-Verlag.

[22] S. Wang and A. Liew. Information-Based Color Feature Representation for Image Classification. In *IEEE International Conference on Image Processing, 2007. ICIP 2007*, volume 6, pages 353–356, 2007.

[23] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206, New York, NY, USA, 2007. ACM.

[24] C. Zhai. Notes on the lemur tfidf model. Technical report, Carnegie Mellon University, 2001.