RECHERCHE D’IMAGES ET CLASSIFICATION PAR MOTS VISUELS ET DESCRIPTEURS FLOUS

THÈSE PRÉSENTÉE À LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET DE LA RECHERCHE EN VUE DE L’OBTENTION DE LA MAÎTRISE ÈS SCIENCES EN INFORMATIQUE

WASSIM BOUACHIR

DÉPARTEMENT D’INFORMATIQUE
FACULTÉ DES SCIENCES
CAMPUS DE MONCTON
UNIVERSITÉ DE MONCTON

AVRIL 2010

COMPOSITION DU JURY

Président du jury	**Jalal Almhana** *Dr. / Prof. Université de Moncton, Moncton*
Examinateur externe	**Gheorghe Marcel Gabrea** *Dr. / Prof. Université du Québec (ÉTS), Montréal*
Examinateur interne	**Éric Hervet** *Dr. / Prof. Université de Moncton, Moncton*
Directeur de thèse	**Mustapha Kardouchi** *Dr. / Prof. Université de Moncton, Moncton*
Codirecteur de thèse	**Nabil Belacel** *Dr. / Agent de recherche, CNRC, Moncton* *Prof. associé, Université de Moncton*

REMERCIEMENTS

Pour commencer, je tiens à exprimer ma profonde gratitude au Prof. Mustapha Kardouchi, mon directeur de thèse, qui m’a permis de m’engager dans cette spécialité. Qu’il soit vivement remercié pour l’encadrement qu’il m’a offert tout au long de ma maîtrise, pour tout le temps qu’il m’a accordé et pour les soutiens financiers et moraux dont j’ai bénéficié auprès de lui.

Mes remerciements s’adressent tout autant au Dr. Nabil Belacel qui n’a épargné le moindre effort dans la codirection de ma thèse. Je le salue pour ses précieuses interventions et pour l’aide scientifique et financière qu’il m’a accordée.

Je tiens également à dire un grand merci à chaque membre du jury, le Prof. Éric Hervet, le Prof. Gheorghe Marcel Gabrea et le Prof. Jalal Almhana, qui ont eu l’amabilité d’examiner ce travail de recherche.

Il m’est très agréable d’exprimer aussi ma reconnaissance à mes enseignantes et enseignants, qui ont joué un rôle fondamental dans ma formation. Que mes professeures et professeurs au Département d’informatique de l’Université de Moncton trouvent ici l’expression de ma gratitude infinie pour leur disponibilité perpétuelle et pour avoir enrichi ma réflexion.

LISTE DES TABLEAUX

LISTE DES FIGURES

TABLE DES MATIÈRES

TABLE OF CONTENTS

RÉSUMÉ

L'approche « Sac de mots visuels » (*Bag of Visual Words*), dite aussi « Sac de caractéristiques » (*Bag of Features*), décrit une image par un ensemble de descripteurs locaux en utilisant un histogramme. Chaque composante de cet histogramme représente l'importance d'un motif visuel (appelé mot visuel) dans l'image. Bien que cette méthode d'indexation par le contenu ait été largement utilisée pour la recherche d'images et la classification, des choix de représentation cruciaux – tels que les schémas de pondération – n'ont pas fait l'objet d'études approfondies dans la littérature. En se basant sur les points caractéristiques SIFT (*Scale Invariant Features Transform*) et un modèle flou de représentation, ce travail apporte des améliorations par rapport aux implémentations connues de l'approche, afin de créer des signatures plus robustes pour les images et mieux refléter les poids des mots visuels. Dans un premier temps, nous proposons une méthode d'indexation pour la recherche d'images par le contenu. Cette tâche consiste à chercher dans une collection d'images celles qui ressemblent le plus à une image requête. Ensuite nous traitons le problème de reconnaissance de la catégorie à laquelle une image appartient. À cette fin, nous avons utilisé un classifieur Bayésien naïf en appliquant la méthode d'indexation proposée. Qu'il s'agisse de recherche par le contenu ou de catégorisation, les résultats expérimentaux démontrent que le schéma de pondération proposé est plus performant que les techniques de pondération classiques.

.

Mots-clés: Sac de mots visuels, Recherche d'images par le contenu, Classification d'images, Assignation floue, Schémas de pondération, Réseau Bayésien naïf.

ABSTRACT

The Bag of Visual Words (or Bag of Features) approach describes an image as a set of local descriptors using a histogram. Each bin of the histogram represents the importance of a visual pattern (called visual word) in the image. This indexing method has been frequently used for image search and classification, but crucial representation choices – such as the weighting schemes – have not been thoroughly studied in existing works. In this work, we present an improved implementation for the Bag of Visual Words approach. Our implementation is based on SIFT (Scale Invariant Features Transform) keypoints and uses a Fuzzy model as an alternative to known weighting schemes, in order to reflect the real weights of visual words. First, we propose an indexing method for content based search task that aims to retrieve a large collection of images and returns a ranked list of objects in response to a query image. On the other hand, we consider the problem of recognizing the semantic category of an image. For this purpose, we apply the proposed indexing method and use a naive Bayesian classifier. The conducted experiments demonstrate that the Fuzzy weighting scheme outperforms the existing term weighting techniques for image search as well as for image categorization.

.

Key words: Bag of Visual Words, Content Based Image Search, Image classification, Fuzzy assignment, Weighting schemes, Naïve Bayesian Network.

AVANT-PROPOS

« *Without publication, science is dead.* »
Gerard Piel

Cette thèse englobe les travaux élaborés dans le cadre de mon projet de recherche pour l'obtention du diplôme de Maîtrise ès sciences en informatique. Elle suit un format par articles et comprend deux chapitres dont chacun traite un problème typique du domaine de la « vision par ordinateur » (*computer vision*), à savoir : la recherche d'images par le contenu et la catégorisation d'images. Chacun des deux chapitres correspond à un article publié ou soumis pour publication. Le lecteur remarquera qu'ils présentent des points communs et que chaque chapitre peut être lu de façon indépendante, mais il retrouvera la cohérence de l'ensemble de la recherche dans l'introduction générale et la conclusion qui soulignent la complémentarité des deux travaux.

L'introduction et la conclusion sont conçues en langue française spécifiquement pour cette thèse. Cependant, les deux chapitres de l'étude sont rédigés dans la langue des revues ciblées et ont été ainsi reproduits en anglais mot pour mot, tout en respectant les exigences de la Faculté des Études Supérieures et de la Recherche concernant le format.

La première contribution s'intitule « *Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation* ». Elle aborde, comme son nom l'indique, le problème d'indexation d'images et la recherche basée sur le contenu. Ce travail a été accepté et présenté à « *SITIS'09: the International Conference on Signal Image Technology and Internet Based Systems (sponsored by IEEE)* » qui s'est tenue en décembre 2009 à Marrakech, Maroc.

Dans le deuxième travail, nous nous sommes basés sur la méthode présentée dans le premier article pour indexer les images, en proposant un modèle adapté pour traiter le problème de classification et particulièrement la catégorisation des scènes. L'article porte le titre « *Fuzzy Indexing for Bag of Features Scene Categorization* » et a été soumis le 14 février 2010 à «*ICISP 2010 : the International Conference on Image and Signal Processing* » qui se déroulera en juillet 2010 à Trois-Rivières, Québec, Canada.

J’ai entamé ce travail par une recherche bibliographique portant sur l’indexation d’images par le contenu, en explorant deux familles de méthodes : les méthodes par descripteurs globaux et celles par caractéristiques locales. Cette phase m’a permis d’identifier les approches les plus efficaces dans la littérature et particulièrement le modèle de « Sac de mots visuels » qui s’inscrit dans la deuxième famille et qui a fait l’objet de plusieurs recherches récentes. Une étude approfondie des implémentations connues de cette approche a permis de découvrir les limites des travaux existants et de proposer par conséquent des idées d’améliorations. Les résultats obtenus après les travaux d’implémentation ont constitué la ligne directrice qui a motivé la rédaction des articles cités ci-dessus. En effet, les études expérimentales ont confirmé les hypothèses a priori en démontrant l’efficacité de la méthode proposée, aussi bien pour la recherche d’images par le contenu, que pour la classification.

Cette démarche a été structurée par le Prof. Mustapha Kardouchi (mon directeur de thèse et second auteur dans les articles) et le Dr. Nabil Belacel (mon codirecteur de thèse et troisième auteur). Mes coauteurs m’ont guidé tout au long de la réalisation de ma thèse en m’aidant depuis la définition du sujet jusqu'à la publication des résultats.

Introduction générale

L’évolution rapide des technologies d’acquisition et d’échange d’images numériques s’est accompagnée d’un progrès impressionnant de la capacité de stockage physique. Ainsi, il est devenu indispensable pour les individus, aussi bien que pour les organisations, de disposer des techniques et applications de la « vision par ordinateur », permettant un accès efficace aux quantités énormes d’images disponibles en ligne.

Dans ce travail, on s’intéresse à deux problèmes de la « vision par ordinateur » qui constituent, depuis les années 90, les centres d’intérêt de beaucoup de chercheurs du domaine: la recherche par le contenu et la catégorisation. La recherche par le contenu consiste à trouver dans une collection d’images, les images les plus similaires à une image requête, alors que pour la catégorisation, l’objectif est de trouver la catégorie à laquelle appartient une image.

Plutôt que de compter sur les métadonnées introduites par les humains – telles que les titres et les mots clés – pour rechercher des images ou reconnaître leurs catégories, la recherche par le contenu et la classification reposent sur l’analyse et l’extraction automatique des caractéristiques visuelles à partir des pixels, afin d’indexer les collections d’images. Étant donné le fossé qui existe entre les objets du monde réel et leur représentation sur des images, la problématique de base est : comment traduire la similarité sémantique entre les images en une similarité visuelle?

1. Problématique

Les premiers travaux d’indexation utilisent des descripteurs globaux pour représenter l’image par un vecteur numérique reflétant des propriétés physiques tels que l’histogramme de couleurs [1], la cooccurrence de couleurs [2] et les filtres de Gabor [3]. Bien que ces caractéristiques aient montré de bonnes performances dans ces travaux, un descripteur global a l’inconvénient de ne pas représenter les concepts locaux exprimés dans l’image.

Le modèle « Sac de mots visuels » (en anglais *Bag of Visual Words* ou *Bag of Features*) est une approche récente d'indexation par caractéristiques locales [4]. L'indexation d'une base d'images par cette approche revient à extraire des caractéristiques locales de toute la collection afin de construite un vocabulaire visuel unique formé par des motifs locaux appelés « mots visuels ». Chaque image de la collection est ensuite décrite par une signature sous forme d'histogramme, où chaque composante correspond à un mot visuel et la valeur associée représente son poids dans l'image.

La simplicité de ce modèle et sa généralisation pour tout type d'images en font une approche largement utilisée, aussi bien pour la recherche que pour la catégorisation. Toutefois, nous avons constaté dans la littérature une multitude de choix d'implémentation et plusieurs facteurs régissant l'efficacité de chaque étape, tels que le choix des descripteurs locaux, la méthode de création du vocabulaire, la taille du vocabulaire, la méthode de calcul des poids des mots visuels, la mesure de similarité entre signatures d'images (pour la recherche d'images), le modèle de classification (pour la catégorisation)…etc. Ce travail de recherche explore tous ces facteurs en soulevant deux problématiques principales : 1) le schéma de pondération et 2) la méthode de classification.

1.1 Problématique liée aux schémas de pondération

L'approche « Sac de mots visuels » est fortement inspirée du modèle « Sac de mots » pour la représentation des documents textuels, où chaque document est décrit par les poids correspondants aux mots du vocabulaire [5]. Cette analogie est reflétée par le titre du premier travail qui utilise l'approche textuelle pour l'indexation visuelle : "*Video Google: A Text Retrieval Approach to Object Matching in Videos*" [4]. Par conséquent, le modèle textuel et le modèle visuel partagent plusieurs propriétés, dont le schéma de pondération. Ce dernier détermine la nature des poids qu'on associe aux mots visuels, tels que la présence ou l'absence des mots, leurs fréquences, ou tout autre poids, avec l'hypothèse que plus le poids d'un mot visuel est élevé, le mieux il décrit l'image.

Bien que ces représentations d'images et de texte reposent sur la même forme, il existe des différences fondamentales entre les deux. En effet, le vocabulaire, les mots, leur signification et leurs fréquences sont des concepts typiquement différents selon le modèle. Le problème est donc de trouver le schéma de pondération le plus informatif en indexation d'images. Comme nous allons le montrer dans ce travail, les schémas de pondération du texte ne garantissent pas la meilleure représentation pour les images.

1.2. Problématique liée au modèle de classification

La classification (ou la catégorisation) est une méthode d'analyse de données qui vise à regrouper un ensemble d'observations en classes (ou catégories) homogènes [6]. Dans ce travail, on s'intéresse à la classification supervisée des images ce qui revient à apprendre à partir d'images indexées dont on connaît les classes, un modèle qui permet de prédire l'appartenance d'une nouvelle image à une des classes connues a priori. Dans la phase d'apprentissage, deux difficultés importantes sont rencontrées: les modèles de classification complexes garantissent généralement une bonne reconnaissance, mais nécessitent souvent un temps d'apprentissage important avec les bases de données volumineuses. Par ailleurs, l'approche « Sac de mots visuels » génère des données de grande dimension, ce qui augmente davantage le temps d'apprentissage et complique la recherche de corrélation entre les données.

Le choix du bon classifieur influe considérablement sur le temps d'apprentissage et l'efficacité de la reconnaissance. L'enjeu est donc de trouver le modèle de classification le mieux adapté à l'approche d'indexation, en permettant d'établir un bon compromis entre temps d'apprentissage réduit et taux élevé de classification correcte.

2. Objectifs

L’objectif de ce travail de recherche est d’étudier l’approche « Sac de mots visuels » pour l’indexation d’images et de l’appliquer à la recherche par le contenu et à la classification. Cette recherche permettra de proposer des améliorations par rapport aux implémentations connues, puis de valider l’implémentation proposée en la comparant aux autres méthodes. Le but est d’optimiser les performances de la recherche d’images et de la classification, tout en maintenant la simplicité de l’approche.

CHAPITRE I

IMPROVING BAG OF VISUAL WORDS IMAGE RETRIEVAL: A FUZZY WEIGHTING SCHEME FOR EFFICIENT INDEXATION

AMÉLIORATION DE LA RECHERCHE D'IMAGES PAR L'APPROCHE « SAC DE MOTS VISUELS » : UN SCHÉMA FLOU DE PONDÉRATION POUR UNE INDEXATION EFFICACE

Wassim Bouachir, Mustapha Kardouchi et Nabil Belacel

Texte original de l'article publié le 2 décembre 2009 dans la conférence «*SITIS'09: the International Conference on Signal Image Technology and Internet Based Systems (sponsored by IEEE)*».

Abstract

Recent works on Content Based Image Retrieval rely on Bag of Visual Words to index images. Analogically to the Bag of Words approach used in text retrieval, this model allows describing an image as a bag of elementary local features called visual words. As a result, an image is represented by a vector of weights, where each weight corresponds to the importance of a visual word in the image. The choice of local features and the weighting scheme are very important to perform image retrieval. The existing weighting schemes are mostly migrated from text retrieval domain and don’t take into account fundamental differences between textual words and visual words. In this paper, a novel approach based on Scale Invariant Features Transform (SIFT) features and a new weighting scheme is proposed. The proposed scheme uses a fuzzy representation to index images with a more robust signature. Experimental results with the Coil-100 image database demonstrate that the proposed method produces better performance than known term weighting representations.

Index Terms: Bag of visual words, Content based image retrieval, Fuzzy assignment, Weighting schemes, SIFT.

I.1 Introduction

Content Based Image Retrieval (CBIR) is the computer vision application based on visual contents that aims to organize images in response to a query. This application differs from traditional retrieval systems based on keywords to search images. The fundamental problem in CBIR is how to transform visual contents into distinctive features for dissimilar images, and into similar features for images that look alike. On the other hand, the main problem is how to represent the semantic contents with features in order to index images.

Many different approaches for CBIR have been proposed in the literature. Swain and Ballard [1] were the first to use color histograms features to describe images. Since, many other authors introduced other features like texture [7] or colorimetric moments [8].

These descriptors allow a quite efficient retrieval in many cases, but fail in precision, because global features lose most of local information expressed in the image.

Recent approaches propose to use local features to describe interest regions in the image. The idea is to detect interesting local patches, represent the patches as numerical vectors and consider images as subsets of these basic elements. Finally, a single signature is computed for each image which allows comparing images by measuring the similarity between signatures. Bag of Visual Words (BoVW) or Bag of Features is one of the most popular frameworks to describe images as sets of elementary local features. Based on local descriptors, this approach is analogous to the bag of words representation for text document in terms of form and semantics. The description of an image collection by BoVW requires three main steps: detecting and describing interest regions, building a visual vocabulary and indexing images. Therefore, an image is described by a vector of weights computed according to a weighting scheme. Each weight in the vector represents the importance of a visual word in the image.

Since an image is described by its visual words like a text document is described by its terms, we have seen the use of term weighting techniques directly migrated from text retrieval domain [9] without considering the differences between text and images. In fact, the text words vocabulary contains the terms that appear in the text corpus so that the document term vector is constructed naturally by finding the term of each word according the language grammar and semantic. For images, a visual word is a numerical vector, and the visual vocabulary is the output of vector quantization. Thus a BoVW of an image is obtained by finding for each local feature the most similar visual word based on a nearest neighbour search. Mapping local features into visual words in such a way may involve a loss of fidelity to visual content since two local features associated with the same visual word contribute in the same way to the construction of the image signature, whether they are identical or appreciably different.

The aim of this work is to propose a method that keeps simplicity and efficiency of the BoVW approach and generates a more realistic image signature, taking into account differences between textual words and visual words. Our approach uses SIFT

[10], [11] to extract local features and introduces a new weighting scheme based on a fuzzy model.

This paper is organized as follows: section 2 describes the BoVW indexation system and reviews most popular weighting schemes. In Section 3, we present the proposed weighting scheme. Section 4 provides detailed experimental results. Finally, section 5 concludes the paper.

I.2 Bag of Visual Words model

The visual words model describes an image using a set of visual words called visual vocabulary. The vocabulary is obtained by clustering local features extracted from images where each resulting cluster is a visual word. In this model, an image is finally represented by a histogram, where each bin corresponds to a visual word and the associated weight represents its importance in the image. Thereby, the construction of the histogram requires three steps: 1) extracting visual features, 2) building a visual vocabulary and 3) indexing images.

I.2.1 Extracting visual features

Many approaches have been proposed to extract local features from images. In [12] and [13] the authors extract local patches using a regular grid. Other authors use also random sampling [14], [15] and segmentation methods [16]. Despite their simplicity, these methods don’t often use the most relevant information of an image. A more interesting approach is to extract keypoints. These keypoints are the centers of salient patches since they are generally located around the corners and edges. In our work, we use SIFT as keypoints detector and descriptor. SIFT features are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint [17]. A SIFT descriptor is a vector of 128 elements summarizing a local information in an image. SIFT features are extracted and will be used to build the visual vocabulary.

I.2.2 Building the visual vocabulary

Building the visual vocabulary means quantifying extracted local descriptors for a large sample of images. The vocabulary can be generated by clustering SIFT features using the standard k-means algorithm. The size of the vocabulary is the number of clusters and the clusters centers are the visual words. Each image in the database will be represented by visual words from this vocabulary.

I.2.3 Images indexing

Once the visual vocabulary is built, we index the images by constructing their BoVW signatures. An image BoVW signature requires finding the weight of each visual word from the vocabulary. Thereby, each image is represented by a histogram where the bins are visual words and the weights are their frequencies in the image.

I.2.3.1 Popular BoVW weighting schemes

Analogically to standard weighting schemes in text retrieval domain, the weight of a visual word is obtained by multiplying three factors:

- **Term Frequency *(tf)*.** The visual word is frequently mentioned in an image which suggests *tf* factor as a part of the weight.
- **Inverse Document Frequency *(idf)*.** This is a collection-dependent factor used to favour visual words found in a few images of the collection. The intuition is that *tf* weights visual words often occurring in a particular image, while *idf* down-weights those that often appear in the collection.
- **The normalization factor.** This component is introduced to treat equally all the images, because the number of keypoints varies according to the complexity of the image content.

Table 1 summarizes popular term weighting factors where they are named and described after the convention in [5].

Table I-1. Term weighting factors

Name	Value	Description
		Term frequency factor
b	1 or 0	Binary i.e. 1 for visual words present, 0 if not.
t	tf	Number of occurrence of the visual word.
		Collection frequency factor
x	1.0	No change in weight.
f	$log \frac{N}{n}$	Multiply by *idf* (N is the number of images in the collection, and n the number of images containing the visual word).
		Normalization factor
x	1.0	No normalization.
c	$\frac{1}{\sum w_i}$	Each visual word weight w_i is divided by the sum of the image weights.

For image search, we have seen the use of *term frequency-inverse document frequency* (*tfx)* [4], [18] and the count of visual words *(txx)* [19]. We have also seen the use of normalized term frequency *(txc)* [20] and binary weights *(bxx)* [21] for image classification. All these methods perform the nearest neighbour search in the vocabulary to map each keypoint to the most similar visual word. In the next section we present shortcomings of directly assigning a keypoint to its nearest neighbour to weight the visual words.

I.2.3.2 Drawbacks of existing representations

In [14], the authors studied empirically the impact of the weighting factors choice on image retrieval performance. They demonstrated that to choose the best weighting scheme, it’s necessary to consider image collection properties and the vocabulary size but we found no recommendations to make this choice.

Using term weighting schemes migrated from automatic text retrieval domain is not an optimal choice. In fact, the textual terms vocabulary is generated naturally by analyzing the text corpus, while the visual words vocabulary is the output of numerical vector quantization using the clustering algorithm. Furthermore, when constructing a bag of words vector for a text document, each word corresponds to a certain term of the vocabulary, e.g. the words “walks”, “walking” and “walked” would be counted in the entry of the term “walk”. A BoVW of an image is obtained in a different way by mapping keypoints to visual words. A similarity measure between numerical vectors is used and each keypoint is considered as its nearest visual word from the vocabulary. Indexing images in this way reduces the signature discriminative power. In fact, two keypoints may be assigned to the same visual word even if they are not equally similar to this word. As a consequence, they contribute in the same way in the construction of the image signature and the obtained value doesn’t reflect the real weight of the visual word in the image. Certainly, the more the vocabulary size is increased, the more this effect is opposed. But in this case two similar keypoints may be considered as two different visual words. In addition, the vocabulary would be less generalizable, noise sensitive and incurs longer processing time to perform retrieval.

Instead of using a text retrieval weighting scheme, we propose a more realistic approach to weight visual words with a fuzzy assignment.

I.3 The proposed Fuzzy representation

Suppose that $V = \{v_1, v_2, ..., v_i, \dots, v_k\}$ is the vocabulary formed by the k centers of clusters (visual words) obtained after vector quantization with k-means algorithm. Let p_j, $j \in \{1, 2, \dots, M\}$ be a SIFT local descriptor among M keypoints descriptors extracted from an image. We associate to p_j a fuzzy description considering all the vocabulary visual words. This represents the contribution of the keypoint in the weight of each visual word. A membership degree is defined using the fuzzy membership function of Fuzzy-C-Means algorithm [22]:

where U_{ij} is the contribution of the keypoint described by p_j in the weight of the visual word v_i and $m \in]1,\infty[$ is the degree of fuzziness. Thus, a fuzzy histogram is obtained and each bin represents the fuzzy weight of the corresponding visual word. Such a representation takes into account the similarity between the keypoint and each visual word from the vocabulary.

To illustrate this effect, let's consider only the two first SIFT elements among the 128 components. Figures 1, 2 and 3 represent the contribution of two local descriptors p_1 and p_2 in the weights of two visual words. In figure 2, p_1 and p_2 contribute in the same way to the weight of their nearest visual word even if they were not equally similar to this word (figure 1). By using the fuzzy assignment, the two keypoints contribute to the weights of both words, thus the distribution of weights is more equitable (figure 3). The parameter m $(1<m<\infty)$ controls the degree of fuzziness in the distribution of weights. Empirically, we found that m=1.1 is the best setting.

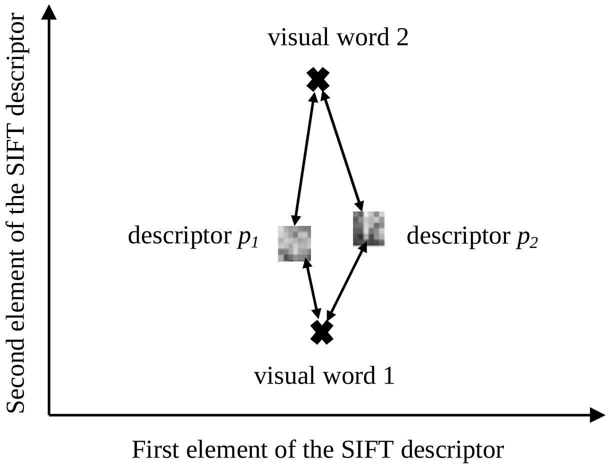

visual word 2
descriptor p_1
descriptor p_2
visual word 1
Second element of the SIFT descriptor
First element of the SIFT descriptor

Figure I-1. Similarity measurement before assigning keypoints to visual words

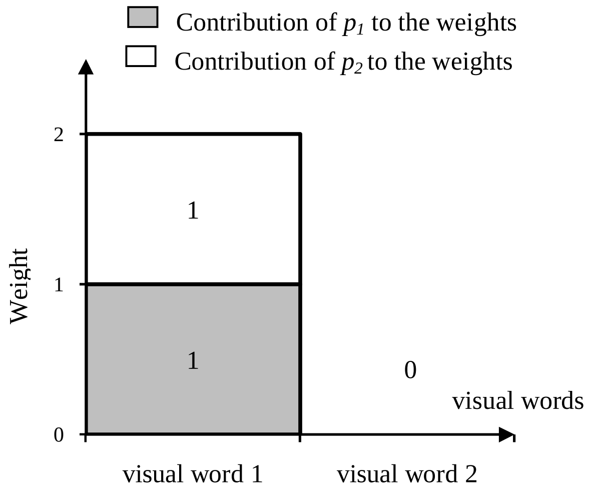


Figure I-2. Crisp assignment

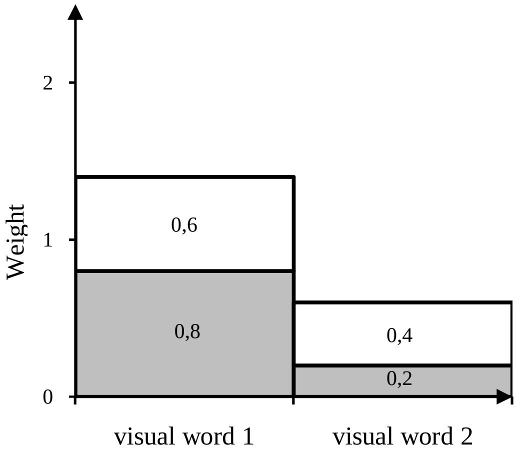


Figure I-3. Fuzzy assignment

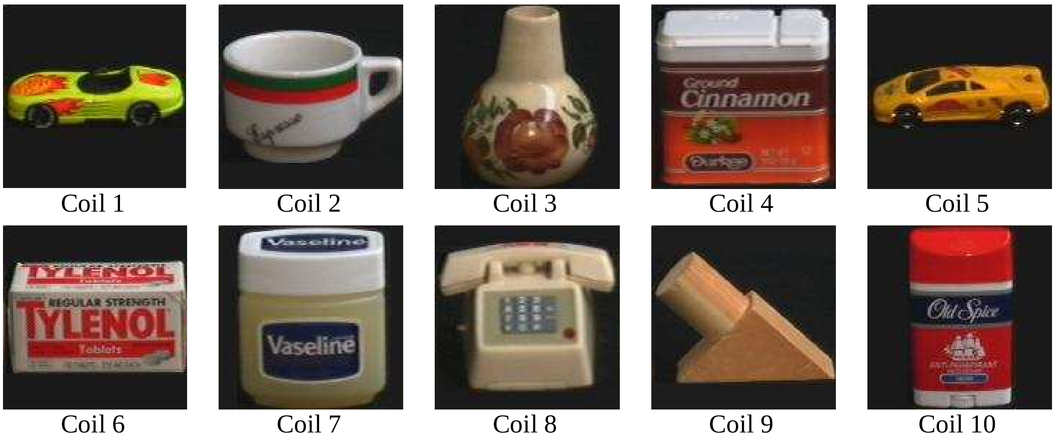
I.4 Experiments

I.4.1 Image collection

We used Coil-100[1] image database to evaluate the proposed approach and compare to the other weighting schemes. The Columbia University Coil-100 database

[1] Available at: http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

contains 7200 images of 100 different objects, where 72 images where taken for each object at 72 different viewpoints separated by 5°. Figure 4 shows ten different objects randomly selected for experiments.



Coil 1 Coil 2 Coil 3 Coil 4 Coil 5

Coil 6 Coil 7 Coil 8 Coil 9 Coil 10

Figure I-4. Sample images from Coil-100 database
(the images have the same size : 128x128)

We randomly selected 3000 images from the database to extract SIFT keypoints. Then we use the k-means clustering algorithm to cluster extracted local features into a visual vocabulary. For our experiments, we set the size of the vocabulary to 100 visual words. Since previous works used *term frequency-inverse document frequency* (*tfx)* [4], [18] and *term frequency (txx)* [19] for image search, we index the database in three different ways using *tfx*, *txx* and the proposed *Fuzzy weighting scheme*. Several queries are performed to compare the performances of these schemes and we used the Euclidian distance to compute the similarity between signatures.

I.4.2 Experimental results

In this section, we evaluate the proposed fuzzy weighting scheme. We use statistics *recall* and *precision* where *precision* is defined as the number of correctly retrieved images by a search divided by the total number of images retrieved, and *recall* denotes the number of images retrieved by a search divided by the number of images of the class that the target image belongs to. The *precision/recall* curve is obtained by varying the number of images returned by a query.

For each query, the recall and precision are computed for the 10, 20, 40, 60, 100 and 200 most similar images retrieved. Figures 5 and 6 plot the results of *precision* versus *recall* for two images Coil 10 and Coil 3, showing that the precision rate decreases as recall increases.

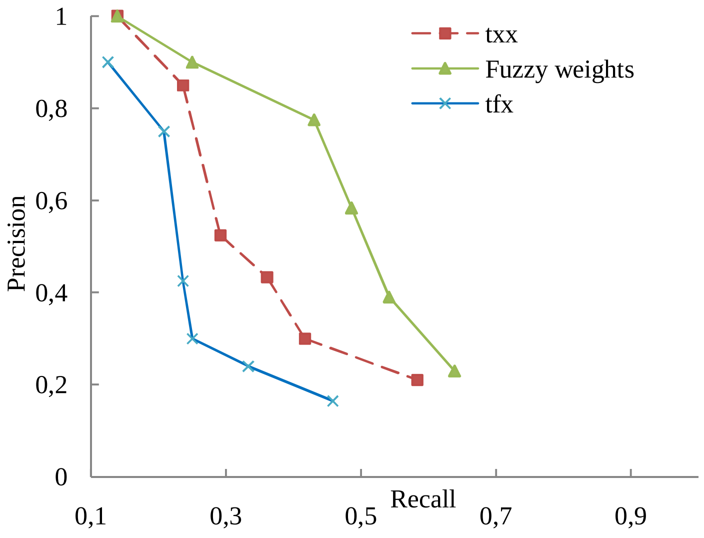


Figure I-5. Recall versus Precision: image Coil 10

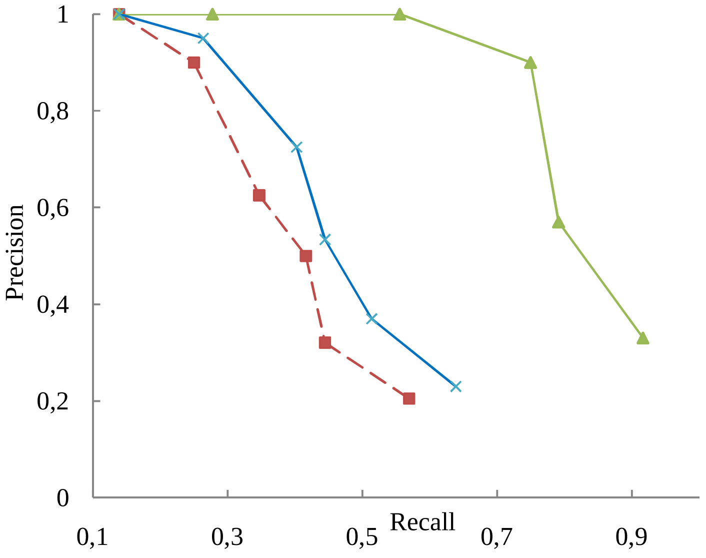


Figure I-6. Recall versus Precision: image Coil 3

In figure 5, *tfx* has the worst performance and the *Fuzzy weighting scheme* outperforms the others in all *recall/precision* points. For the image Coil3, it's clear that the *Fuzzy weighting scheme* outperforms significantly the other schemes, having a precision of 100% for the three first *recall/precision* points. Both *txx* and *tfx* curves are close but for this image *txx* had the worst performance.

To further compare the performance of various weighting schemes, we performed on each database 10 queries returning the 20 most similar images, using the images in figure 4 as targets. Table 2 presents the precision of the retrieval by using different schemes, and shows that the *Fuzzy weighting scheme* outperforms the others except for Coil 7 and Coil 9 having also the best average precision.

Table I-2. Recognition rates among 20 most similar images for different weighting schemes

Image	***txx***	***tfx***	***Fuzzy weight***
Coil 1	0,5	0,4	0,65
Coil 2	0,4	0,1	0,45
Coil 3	0,9	0,95	1
Coil 4	1	0,9	1
Coil 5	0,25	0,1	0,75
Coil 6	1	1	1
Coil 7	1	0,85	0,95
Coil 8	0,55	0,5	0,7
Coil 9	0,7	0,6	0,6
Coil 10	0,85	0,75	0,9
Average	**0,715**	**0,615**	**0,8**

We completed the measurements for the 10, 40, 60, 100 and 200 most similar images retrieved to plot, in figure 7, the average precision versus average recall for the ten images.

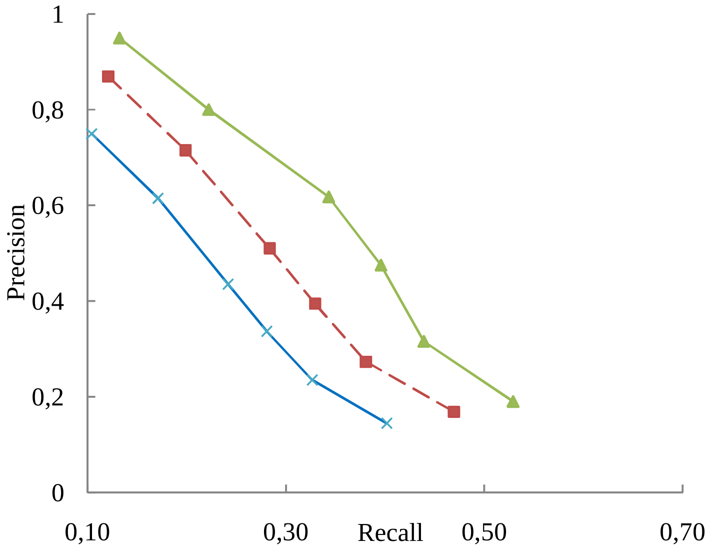


Figure I-7. Average recall versus average precision for ten queries on Coil-100

This figure shows that when indexing Coil-100 images using fuzzy weights, retrieval results are considerably better than those obtained with *txx* and *tfx*.

I.5 Conclusion

The Bag of Visual Words model has recently received a lot of attention owing to its simplicity and good practical performance. Despite the effectiveness of the BoVW signature, we have shown that using representation techniques from automatic text retrieval domain is not the optimal choice. To avoid this drawback, we define a fuzzy model, specifically for visual words instead of using known term weighting schemes. The proposed approach takes into account the fundamental difference between text and images and the experiments proved its superiority.

The BoVW approach can be improved by several other ways, such as using a more effective algorithm to create the visual vocabulary, taking into account the large number of keypoints and noisy data. We believe also that the color provides valuable information in keypoints description. Since SIFT descriptors use only gray scale information and don’t handle color, a very important source of distinction may be lost. Consequently, a further improvement of BoVW would occur by introducing the color information to describe keypoints. One other interesting direction for future work is to decompose the image signature into sub-histograms. Each one corresponds to a part of the described image. As a result, the BoVW signature is enriched by the information on the spatial relation between visual words.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the New Brunswick Innovation Foundation (NBIF). Authors would like to thank David Lowe for the SIFT feature code and the department of computer science at Columbia University for the Coil-100 database**.**

CHAPITRE II

FUZZY INDEXING FOR BAG OF FEATURES SCENE CATEGORIZATION

INDEXATION FLOUE POUR LA CATÉGORISATION DES SCÈNES AVEC L'APPROCHE « SAC DE CARACTÉRISTIQUES »

Wassim Bouachir, Mustapha Kardouchi et Nabil Belacel

Texte original de l'article soumis le 14 février 2010 à «*ICISP 2010 : the International Conference on Image and Signal Processing*».

Abstract

This paper presents a novel Bag of Features (BoF) method for image classification. The BoF approach describes an image as a set of local descriptors using a histogram, where each bin represents the importance of a visual word. This indexing approach has been frequently used for image search and classification, but crucial representation choices – such as the weighting schemes – have not been thoroughly studied in the literature. In our work, we propose a Fuzzy model as an alternative to known weighting schemes in order to create more representative image signatures. Furthermore, we use the Fuzzy signatures to train the Gaussian Naïve Bayesian Network and classify images. Experiments with Corel-1000 dataset demonstrate that the proposed weighting scheme outperforms known term weighting techniques in scene categorization.

Index Terms: Bag of Features, Image classification, Fuzzy assignment, Weighting schemes, Naïve Bayesian Network.

II.1 Introduction

The expansion of means for acquisition, storage and exchange is producing growing image databases. Managing and accessing such huge collections is becoming a field of great interest for computer vision researchers. In this work, we consider the problem of recognizing the semantic category of an image. For instance, we may want to classify a given image to one of these categories: *Building, Mountain, beach*, etc. This recognition task requires automatically analyzing, and transforming visual contents into representative features in order to index images.

BoF model is a recent indexing method that uses local descriptors to represent interest regions and consider images as sets of elementary features [4], [21]. The description of an image collection with this approach requires three main steps: detecting and describing interest regions, quantifying extracted local descriptors to build a visual vocabulary, and finally indexing each image by computing a signature that contains the weights of all visual words of the vocabulary. The weights are calculated according to a weighting scheme and each one represents the importance of a visual word in the image. The BoF framework was conceived analogically to the “Bag of Words” approach in text retrieval domain [5], [23], [24]. Consequently, computer vision researchers have been using text retrieval weighting schemes to compute the weights of visual words. Since there are fundamental differences between textual words and visual words, we aim to define a specific weighting scheme for BoF indexing using a Fuzzy model. Our method maintains simplicity and efficiency of the BoF approach, while producing a Fuzzy signature that reflects the real weights of visual words. Furthermore, we propose to train the Gaussian Naive Bayesian Network using the obtained Fuzzy weights and evaluate our method for scene classification.

The paper is organized as follows: the second section describes the BoF framework, the third one reviews the known weighting schemes and presents shortcomings of such representations. Sections 4 and 5 respectively present the proposed Fuzzy method and the used Naïve Bayes classifiers. Section 6 provides detailed experimental results, and section 7 concludes the paper.

II.2 BoF framework:

The BoF model describes each image using a set of visual words called visual vocabulary. The vocabulary is obtained by clustering local features extracted from images, where each resulting cluster is a visual word. An image is finally represented by a histogram. Each bin of this histogram corresponds to a visual word, and the associated weight represents its importance in the image. Thereby, the construction of the histogram requires three steps: 1) extracting visual features, 2) building a visual vocabulary and 3) indexing images.

II.2.1 Extracting local features

A very interesting approach for extracting local features is to detect keypoints. Those are the centers of salient patches generally located around the corners and edges. In our work, we detect and describe keypoints using Scale Invariant Features Transform (SIFT) [11] because of its reasonable invariance to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint [17]. In this step, SIFT keypoints are extracted, and each one is described by a vector of 128 elements summarizing a local information. The extracted features will be used to build the visual vocabulary.

II.2.2 Building the visual vocabulary

Building the visual vocabulary means quantifying extracted local descriptors for a large sample of images. The vocabulary can be generated by clustering SIFT features using the standard k-means algorithm. The size of the vocabulary is the number of clusters, and the centers of clusters are the visual words. Each image in the database will be represented by visual words from this vocabulary.

II.2.3 Image indexing

Once the visual vocabulary is built, we index each image by constructing its BoF signature. This requires finding the weight of the visual words from the vocabulary. Each image is described by a histogram, where the bins are the visual words and the corresponding values are the weights of the words in the image.

II.3 A review of weighting schemes for BoF indexing

II.3.1 Popular weighting schemes

Analogically to the text retrieval approach, the weight of a visual word is obtained by multiplying three factors explained below and detailed in table 1:

- **Term Frequency *(tf)*.** The visual word is frequently mentioned in an image.
- **Inverse Document Frequency *(idf)*.** This is a collection-dependent factor used to favour visual words found in a few images of the collection. The intuition is that *tf*

weights visual words often occurring in a particular image, while *idf* down-weights those that often appear in the collection.

- **The normalization factor.** This component is introduced to treat equally all the images, because the number of keypoints varies depending on the complexity of the image content.

Table II-1. Description of the term weighting factors [5].

Name	Value	Description
		Term frequency factor
b	1 or 0	Binary i.e. 1 for visual words present, 0 if not.
t	tf	Number of occurrence of the visual word.
		Collection frequency factor
x	1.0	No change in weight.
f	$log\,\frac{NC}{nv}$	Multiply by *idf* (*NC* is the number of images in the collection, and *nv* the number of images containing the visual word).
		Normalization factor
x	1.0	No normalization.
c	$\frac{1}{\sum w_i}$	Each visual word weight w_i is divided by the sum of the image weights.

For image search, we have seen the use of *term frequency-inverse document frequency* (*tfx)* [4], [18] and the count of visual words *(txx)* [19]. We have also seen the use of *txx* [25] and binary weights *(bxx)* [21] for image classification. Note that all of these methods perform the nearest neighbour search in the vocabulary to map each keypoint to the most similar visual word.

II.3.2 Drawbacks of existing representations

Using term weighting schemes migrated from text retrieval domain is not the optimal alternative. In fact, the textual terms vocabulary is generated naturally by analyzing the text corpus, while the visual words vocabulary is the output of numerical vector quantization using the clustering algorithm. Furthermore, when constructing a "Bag of Words" vector for a text document, the document term vector is obtained naturally, by finding in the vocabulary, the word stem in accordance with the language grammar and semantic. A BoF for an image is obtained in a different way by mapping keypoints to visual words. A similarity measure between numerical vectors is used and each keypoint is considered as its nearest visual word from the vocabulary. Indexing images in this way reduces the discriminative power of the signature. Two keypoints may be assigned to the same visual word even if they are not equally similar to this word. Consequently, they contribute in the same way to the construction of the image signature, and the obtained value does not reflect the real weight of the visual word. Certainly, the more the vocabulary size is increased, the more this effect is opposed. But in this case, two similar keypoints may be considered as two different visual words. In addition, the vocabulary would be noise sensitive, less generalizable, and incurs longer processing time to train the classifier. Instead of using a text retrieval weighting scheme, we propose a more realistic approach to weight visual words by using a Fuzzy assignment.

II.4 The Fuzzy representation

Suppose that $V = \{v_1, v_2, ..., v_i, \ldots, v_k\}$ is the vocabulary formed by the k centers of clusters (visual words) obtained after vector quantization with k-means algorithm. Let p_j, $j \in \{1, 2, \ldots, M\}$ be a SIFT local descriptor among M keypoints descriptors extracted from an image. We associate to p_j a Fuzzy description considering the whole vocabulary. This description represents the contribution of the keypoint in the weight of each visual word. For this purpose, a membership degree is defined using the Fuzzy membership function of Fuzzy-C-Means algorithm [22]:

$$U_{ij} = \frac{1}{\sum_{n=1}^{k} \left(\frac{||p_j - v_i||}{||p_j - v_n||} \right)^{\frac{2}{m-1}}} \qquad (1)$$

where U_{ij} is the contribution of the keypoint described by p_j in the weight of the visual word v_i , and m is the degree of fuzziness. Thus, a Fuzzy histogram is obtained and each bin represents the Fuzzy weight of the corresponding visual word. The main advantage of such representation is that it considers the similarity between the keypoint and each visual word from the vocabulary. To illustrate this effect, let us consider two different local descriptors p_1 and p_2 having the same closest cluster center. Figure 1 represents the difference between crisp and Fuzzy assignments.

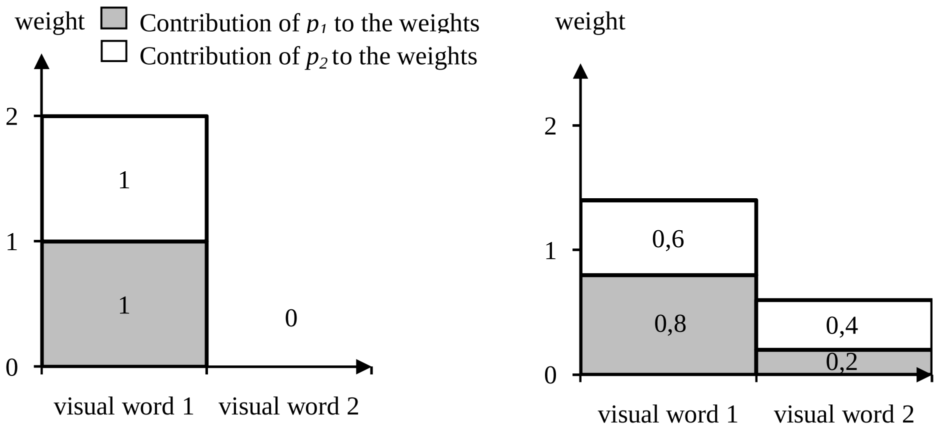


Figure II-1. Crisp assignment (left) versus Fuzzy assignment (right)

In the first histogram of figure 1, p_1 and p_2 contribute in the same way to the weight of their nearest visual word even if they are not equally similar to this word. By using the Fuzzy assignment, the two keypoints contribute to the weights of the two words, and thus the distribution is more equitable. The parameter m *(*$1<m<\infty$*)* controls the degree of fuzziness in the distribution of weights. Empirically, we found that m=1.1 is the best setting.

II.5 Categorization by Naïve Bayes

The Naïve Bayesian Network (NBN) has been widely used for bags of words text categorization because of its simplicity, learning speed and competitiveness with the state-of-the art classifiers [23], [24], [26], [27]. Consequently, it has also been used as a BoF image classifier [25]. The main idea of this model is to learn from a training set the conditional probability of each attribute given a class. The classification decision is taken by applying Bayes's rule:

$$P(C_i|X_n) = \frac{P(C_i)\,P(X_n|C_i)}{P(X_n)} \qquad (2)$$

where $P(C_i|X_n)$ is the probability of the category C_i given X_n (the BoF vector of an image I_n). $P(C_i)$ and $P(X_n)$ are respectively the prior probability of the class C_i, and the prior probability of obtaining the signature X_n for an image. The probability $P(X_n)$ is the same for all the classes, and therefore, it can be ignored without affecting the relative values of class probabilities. Finally, we consider the largest a posteriori score as the class prediction. This prediction is possible by making a strong independence assumption called the *naïve assumption*: the visual words of the vocabulary are conditionally independent given the class. The reason why NBN is able to work well with the BoF approach is that the conditional independence assumption is quite reasonable: if we know that an image belongs to a category, this is sufficient to specify what kind of visual words we will find in this image. Moreover, BoF approach uses high-dimensional attribute spaces where it is very difficult to estimate the correlation between attributes. Practically, attributes are seldom independent given the class, but it has been verified that the NBN performs well even when strong attribute dependences are present [28]. The other important aspect that motivated our classifier choice is its tolerance to learn parameters from different data types generated by different weighting schemes. In existing works, we have seen the use of *txx* [25] and binary weights *(bxx)* [21] for image classification. To compare the weighting schemes performance, we train two instances of NBN. The first learns its parameters from data produced by applying *bxx*, while the second uses *txx* data. Further, we use the Gaussian NBN to learn from the Fuzzy weights.

II.5.1 Conditional probabilities Estimation for Binary Weights

With *bxx*, the BoF vector of an image I_n is X_n=*(w*$_1$*,..., w*$_j$*..., w*$_k$*)* where w_j is the weight of x_j (the *j*th visual word in the vocabulary). The weight w_j is 1 if the word is present, and 0 if not. Given the *naïve assumption* explained above, the conditional probabilities for these binary attributes are computed from the frequencies by counting the number of occurrences of each possible attribute value with each class. Categorization is done by applying equation (2) after decomposing $P(X_n|C_i)$ into the product of the conditional probabilities learned for each attribute value:

$$P(C_i|X_n) = P(C_i)\prod_{j=1}^{k} P(w_j = v|C_i) \tag{3}$$

with $v \in \{0, 1\}$. Note that in order to avoid probabilities of zero, $P(w_j = v|C_i)$ are computed with Laplace smoothing:

$$P\left(w_j = v\middle|C_i\right) = \frac{\left(\#\, images\ of\ class\ C_i\ with\, w_j = v\right) +\ 1}{(\#\, images\ of\ class\ C_i) +\ 2} \tag{4}$$

II.5.2 Multinomial Naïve Bayes for *txx* Representation

The multinomial NBN has been widely used for text classification, where a document is represented by the set of stems occurrences [23], [24], [26], [29]. With *txx* features, the BoF vector contains the visual words counts so that we can model the classifier parameters using the multinomial distribution. During learning step, the classifier computes the relative visual words probabilities separately for each class as follows:

$$P\left(x_j\middle|C_i\right) = \frac{N_{ij} + 1}{N_i + k} \tag{5}$$

where N_{ij} is the count of the visual word x_j in all the training images belonging to class C_i, and N_i the count of all visual words in the training images belonging to C_i. Laplace estimator is used as well as in Equation (4) to avoid the zero probability problem. To categorize a new image I_n, the Naïve Bayes defines a multinomial distribution by using the vector of k probabilities $P\left(x_j\middle|C_i\right)$ for the corresponding class, and by using N_n, the number of visual words for that image. The classification is based on the relative frequencies w_{jn} of the visual words in I_n, by multiplying the class prior $P(C_i)$ by $P(X_n|C_i)$. The latter parameter is the probability of obtaining the signature X_n for an image belonging to C_i. This is calculated by using the multinomial mass function, and thus, we get the a posteriori classes score:

$$P(C_i|X_n) = P(C_i)\, N_n! \prod_{j=1}^{k} \frac{P(x_j|C_i)^{w_{jn}}}{w_{jn}!} \tag{6}$$

Note that we can delete the computationally expensive terms $N_n!$ and $w_{jn}!$ without any change in the results since neither depends on class C_i.

II.5.3 The Proposed Gaussian NBN

By using the Fuzzy weighting scheme, we obtain a BoF vector of real valued attributes that represent the Fuzzy weights of visual words. To model the conditional probabilities distributions, we assume that for a given class C_i, the Fuzzy weight of each visual word x_j is a normally distributed random variable with mean μ_{ij} and variance $\sigma_{ij}{}^2$. This model is based on the assumption that for the images belonging to same class, the weights of a visual word tend to cluster around the mean value. The a posteriori score of classes is then computed using Equation (3) with:

$$P(w_j = v|C_i) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}}\, e^{-\frac{(v-\mu_{ij})^2}{2\sigma_{ij}{}^2}} \tag{7}$$

where $\boldsymbol{v} \in [\mathbf{0}\ ;\ \infty[$.

II.6 Experiments

We explored the performance of the proposed method on the NBN categorization task conducted on Corel-1000 database[2]. Corel is a collection of about 60000 images created by the professor Wang’s group at Penn State University. Corel-1000 is a well known sub-collection that contains 1000 natural images divided into ten categories with 100 images per category.

We randomly selected 300 images to extract SIFT keypoints. Then, we use the k-means clustering algorithm to cluster the extracted local features into a visual vocabulary. For our experiments, we set the size of the vocabulary to 100 visual words. Since

[2] Available at: http://wang.ist.psu.edu/docs/related.shtml

previous works relied on binary weights *(bxx)* [2] and term frequency *(txx)* [10] for image classification, we applied these two schemes and the Fuzzy method to index the image collection in three ways. We divided the collection at random into two sets of images: 70% are used for training each of the three NBN instances and 30% are used for testing. The table 2 shows that when the Gaussian NBN was used with the Fuzzy weighting scheme, 60% of the images were correctly classified, and this was the best rate. With the multinomial NBN, 57% of the scenes were correctly recognized, whereas the binary weights classifier had the worst percentage (37%).

Table II-2. Classification rates.

Weighting scheme	**Percentage of correct classification**
bxx	37%
txx	57%
Fuzzy weights	60%

Corel-1000 is a very challenging collection because of the large number of classes and the high variability of poses and background even for images belonging to the same class. Nevertheless, the conducted experiments demonstrated that when the Gaussian NBN learns from Fuzzy weights, we can handle better difficult situations such as multiple objects in the scene and variable orientation as we can see in figure 2. This figure presents examples where scenes were well classified.

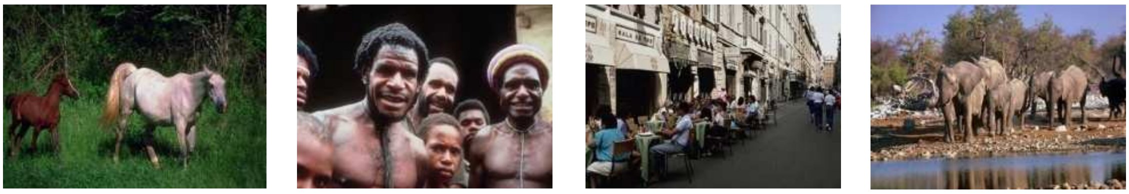


Figure II-2. Scenes correctly classified by the Gaussian NBN as: *Horse*, *Africa*, *Building*, *Elephant*.

The confusion matrix of the Gaussian NBN is given in table 3 where the diagonal elements show interesting correct classification rates for most of classes. It also shows two high rates obtained for the classes *Dinosaur* and *Flowers* (respectively 92% and 94%). The lowest rates are 41% and 40%, and were obtained respectively for the categories *Building* and *Mountain*. The last two percentages could be explained by the fact that these two categories are sharing objects with other classes. For example, 17% of *building* scenes were confused with the category *Bus* because many images from the latter contain also buildings.

Table II-3. The Gaussian NBN confusion matrix.

↓True classes	*Africa*	*Beach*	*Building*	*Bus*	*Dinosaur*	*Elephant*	*Flowers*	*Horse*	*Mountain*	*Food*
Africa	**46**	0	5	8	0	3	14	3	11	11
Beach	0	**45**	10	0	0	7	17	3	7	10
Building	10	10	**41**	17	0	7	3	0	7	3
Bus	4	0	4	**81**	0	0	4	0	4	4
Dinosaur	0	8	0	0	**92**	0	0	0	0	0
Elephant	0	16	6	0	3	**55**	3	13	3	0
Flowers	0	0	0	0	3	0	**94**	0	3	0
Horse	3	0	3	7	0	0	3	**67**	13	3
Mountain	8	4	0	8	4	4	12	0	**40**	20
Food	15	0	15	9	0	6	0	0	9	**47**

II.7 Conclusion

We presented a novel alternative to the known term weighting methods for BoF visual indexing, and we demonstrated that the classical text representation techniques are not a suitable choice for image classification. The proposed method relies on a Fuzzy model conceived for visual words, and the experiments proved its efficiency in NBN classification. The BoF indexing could be improved by several other ways, such as using a more effective algorithm to create the visual vocabulary. In fact, a more representative vocabulary would be generated by using a clustering algorithm that handles the large number of local descriptors and the presence of outliers.

On the other hand, SIFT descriptors use only gray scale information, while the color provides valuable information in keypoints description. This proposes a further improvement by introducing the color information to describe keypoints. One other interesting direction for future work would be to divide the image signature into sub-histograms. Each sub-histogram would correspond to a part of the described image. As a result, the BoF signature is enriched by the information on the spatial relation among visual words.

Conclusion générale

Les travaux de recherche développés dans cette thèse s’inscrivent dans le contexte de la recherche d’images par le contenu et la classification avec l’approche « Sac de mots visuels ». La simplicité et l’efficacité de cette approche découlent des avancées des travaux, aux cours des dernières années, sur les thèmes d’extraction de caractéristiques locales des images, de description locale et des techniques d’analyse et de groupement des données. À l’aide du *clustering*, les descripteurs extraits d’une collection d’images sont réduits à un ensemble représentatif de toute la base appelé le vocabulaire visuel. L’indexation d’une image se ramène par la suite à la représenter par un histogramme des poids des mots du vocabulaire.

Cette thèse de maîtrise a mis l’accent sur les facteurs régissant l’efficacité de l’indexation, de la recherche et de la classification des images par l’approche « Sac de mots visuels ». Pour cela, nous avons consacré plus de temps pour étudier les schémas de pondération et les techniques de classification supervisée appropriées.

La première étude nous a permis de démontrer que les schémas de pondérations retrouvés dans la littérature de l’approche sont migrés du domaine d’indexation des documents textuels et qu’ils ne garantissent pas une indexation efficace des images. Pour remédier à ces limites, nous avons proposé un modèle flou afin d’indexer les images par des signatures qui reflètent mieux les poids des mots visuels et qui permettent par conséquent une recherche par le contenu plus performante.

La catégorisation d’images a fait l’objet de la deuxième phase de cette recherche. Le but de cette phase était de trouver le modèle de classification le plus adapté à la méthode d’indexation proposée, tout en évitant les problèmes dus aux données volumineuses et de grande dimension. Dans ce sens, nous avons adopté un réseau Bayésien naïf avec une distribution gaussienne des paramètres. Ce classifieur repose sur l’hypothèse selon laquelle les mots visuels sont indépendants dans le contexte de la classe, d’où on ne tient pas compte de la corrélation entre ces mots. Par ailleurs, il a permis un apprentissage très rapide et un meilleur taux de classification par rapport aux instances connues du réseau.

Les extensions et les améliorations qui peuvent être ajoutées à l’approche étudiée sont nombreuses et interviennent dans chacune des trois étapes de l’indexation. La première perspective consiste à fusionner d’autres descripteurs avec SIFT pour mieux caractériser les images. En effet, ajouter d’autres informations telles que la couleur et la texture serait une piste intéressante vu l’absence de cette information dans le descripteur SIFT.

La deuxième perspective concerne la construction du vocabulaire visuel qui représente une étape très sensible puisque l’indexation des images en dépend fortement. À la base, il s’agit d’un problème de classification non supervisée (*clustering*) où le but est de grouper des données en grappes (*clusters*), sans connaissance a priori de leur nombre. Les travaux existants se sont basés sur l’expérimentation pour déterminer la taille du vocabulaire, qui peut varier de quelques dizaines à des centaines de mots, dépendamment du nombre d’images, de leur complexité, de leur résolution…etc. Nous pensons que la détermination théorique a priori du nombre optimal de mots pourrait être une voie intéressante pour des travaux futurs. Cela permettrait de créer un vocabulaire regroupant les mots visuels les plus caractéristiques de la base d’images.

La troisième direction serait d’ajouter l’information sur la distribution spatiale des mots visuels sur l’image. L’idée qu’on propose est de construire une grille sur l’image et d’associer un sous-histogramme de poids des mots à chaque bloc obtenu. Tous les sous-histogrammes d’une image peuvent être concaténés pour former une signature qui contient – à part les descripteurs locaux – la disposition spatiale entre eux.

RÉFÉRENCES

[1] M. J. Swain and D. H. Ballard, “Color Indexing,” International Journal of Computer Vision: 1991, pp. 11-32.

[2] P. Howarth and S. Rueger, “Evaluation of texture features for content-based image retrieval,” Proceedings of the international conference on image and video retrieval N^{o}3: 2004, pp. 326-334.

[3] P. Howarth, A. Yavlinsky, D. Heesch, and S. Rüger, “Visual features for content-based medical image retrieval,” Notebook of the Cross Language Evaluation Forum (CLEF) Workshop: 2004.

[4] J. Sivic and A. Zisserman, Video Google: “A Text Retrieval Approach to Object Matching in Videos,” Proceedings of the Ninth IEEE International Conference on Computer Vision, vol.2 : 2003, pp. 1470-1477.

[5] G. Salton and C. Buckley. “Term-weighting approaches in automatic text retrieval,” Information Processing and Management: an Int'l Journal: 24(5): 1988, pp. 513-523.

[6] C. Bouveyron, *Modélisation et classification des données de grande dimension: application à l’analyse d’images*, thèse doctorale, Université Grenoble 1, 2006.

[7] J. Ohm , F. Bunjamin, W. Liebsch, B. Makai, K. Müller, B. Saberdest and D. Zier, “A Visual Search Engine for Distributed Image and Video Database Retrieval Applications,” Proceedings of the Third International Conference on Visual Information and Information Systems: 1999, pp. 187-194.

[8] M. Stricker and M. Orengo, “Similarity of color images,” Proceedings of SPIE Vol. 2, Storage and Retrieval for Image and Video Databases: 1995, pp. 381-392.

[9] J. Yang, Y. Jiang, A. G. Hauptmann, C. Ngo, “Evaluating bag-of-visual-words representations in scene classification,” Proceedings of the international workshop on multimedia information retrieval: 2007, pp. 197-206.

[10] David G. Lowe, “Object Recognition from Local Scale-Invariant Features,” Proceedings of the International Conference on Computer Vision, vol.2: Corfu, September 1999, pp. 1150-1157.

[11] David G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” The International Journal of Computer Vision: 2004, pp. 91-110.

[12] L. Fei-Fei and P. Perona. “A Bayesian hierarchical model for learning natural scene categories,” Proceedings of IEEE International Conference Computer Vision and Pattern Recognition: 2005, pp. 524-531.

[13] J. Vogel and B. Schiele, “On Performance Characterization and Optimization for Image Retrieval,” Proceedings of European Conference on Computer Vision: 2002, pp. 51-55.

[14] M. Vidal-Naquet and S. Ullman. “Object recognition with informative features and linear classification,” Proceeding of IEEE International Conference on Computer Vision: 2003, pp. 281-288.

[15] R. Maree, P. Geurts, J. Piater, J. And L. Wehenkel, “Random subwindows for robust image classification,” Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition: 2005, pp. 34-40.

[16] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, “Matching words and pictures,” Journal of Machine Learning Research 3: 2003, pp. 1107–1135.

[17] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 27, Number 10: 2005, pp. 1615-1630.

[18] W. Zhao, Y. Jiang and C. Ngo, "Keyframe retrieval by keypoints: Can point-to-point matching help?" Proceedings of the 5th international Conference on Image and Video Retrieval: 2006, pp. 72-81.

[19] S. Newsam and Y. Yang, "Comparing Global and Interest Point Descriptors for Similarity Retrieval in Remote Sensed Imagery," Proceedings of the 15th International Symposium on Advances in Geographic Information Systems, 2007.

[20] Y. Yang and S. Newsam. "Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery," Proceedings of IEEE International Conference on Image Processing: 2008, pp. 1852-1855.

[21] E. Nowak, F. Jurie and B. Triggs, "Sampling strategies for bag-of-features image classification," Proceedings of the European Conference on Computer Vision: 2006, pp. 490-503.

[22] J. Bezdeck, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press Ed., New-York, 1981.

[23] A. Juan and H. Ney, "Reversing and Smoothing the Multinomial Naive Bayes Text Classifier," Proceedings of the 2nd Int. Workshop on Pattern Recognition in Information Systems: 2002, pp. 200-212.

[24] A. McCallum, K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," Proceedings of the AAAI-98 Workshop on Learning for Text Categorization: 1998, pp. 41-48.

[25] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," In Workshop on Statistical Learning in Computer Vision, ECCV: 2004, pp. 1-22.

[26] E. van Dyk and E. Barnard, "Naive Bayesian classifiers for multinomial features: a theoretical analysis," Proceedings of the Eighteenth Annual Symposium of the Pattern Recognition Association of South Africa: 2007, pp. 87-92.

[27] David D. Lewis, “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval,” Proceedings of the 10th European Conference on Machine Learning: 1998, pp. 4-15.

[28] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,” Machine Learning 29: 1997, pp. 103-130.

[29] S.-B. Kim, H.-C. Rim and H.-S. Lim, “A new method of parameter estimation for multinomial naive bayes text classifiers,” Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval: 2002, pp. 391-392.