High-Dimensional Visual Vocabularies for Image Retrieval

João Magalhães[1], Stefan Rüger[1,2]

[1]Department of Computing
Imperial College London
South Kensington Campus
London SW7 2AZ, UK

[2]Knowledge Media Institute
The Open University
Walton Hall
Milton Keynes MK7 6AA, UK

(j.magalhaes@imperial.ac.uk, s.rueger@open.ac.uk)

ABSTRACT

In this paper we formulate image retrieval by text query as a vector space classification problem. This is achieved by creating a high-dimensional visual vocabulary that represents the image documents in great detail. We show how the representation of these image documents enables the application of well known text retrieval techniques such as Rocchio tf-idf and naïve Bayes to the semantic image retrieval problem. We tested these methods on a Corel images subset and achieve state-of-the-art retrieval performance using the proposed methods.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting methods.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Semantic image indexing, vector space model.

1. INTRODUCTION

When retrieving images by keyword, algorithms are required to, a priori, index images with a set $\mathcal{W}$ of query keywords. Such algorithms semantically analyse images and assign probabilities to the keywords. In this paper we show how, by carefully creating a high-dimensional visual vocabulary, we enable the application of text categorization algorithms to index images. Unlike Duygulu et al. [1], Lavrenko et al. [3] and others, who have explored the idea of visual vocabularies based on image segmentation, our approach creates a high-dimensional visual vocabulary based on tiled regions and their colour and texture. This approach is similar to a bag-of-words approach in image processing when considering a large set of words.

Section 2 describes how to create the high-dimensional visual vocabulary that enables the application of text search techniques such as Rocchio text classifiers (Section 3), and naïve Bayes text classifiers (Section 4). Section 5 presents experimental results.

2. HIGH-DIM. VISUAL VOCABULARY

In the classic vector space model each document is represented as a vector $\vec{d}$ in a multidimensional space, where each dimension corresponds to the frequency of a given term t_i from a vocabulary of T terms. In our formulation of semantic image retrieval as a vector space model we create a visual vocabulary where each term corresponds to a set of homogenous visual characteristics (colour and texture features). Since we are going to use a single vocabulary to represent all images, we need a set of visual terms that is able to represent them. Thus, we need to check which visual characteristics are more common in the dataset. For example, if there are a lot of images with a wide range of blue tones we require a larger number of visual terms representing the different blue tones. This draws on the idea that to learn a good high-dimensional visual vocabulary we would benefit from examining the entire dataset to look for the most common set of colour and texture features.

The way we build the high-dimensional visual vocabulary is by clustering the entire dataset and representing each term as a cluster. We follow the approach presented in [4], where the entire dataset is clustered with a hierarchical EM algorithm using a Gaussian mixture model. This approach generates a hierarchy of cluster models that corresponds to a hierarchy of vocabularies with a different number of terms.

The only difference between our formulation and the traditional vector space model is that we use $P(t_i \mid d)$ instead of the classic term frequency $TF(t_i \mid d)$. This is equivalent because all documents are represented by a high-dimensional visual vocabulary of length T and $P(t_i \mid d) \cdot T \approx TF(t_i \mid d)$.

3. ROCCHIO TF-IDF CLASSIFIER

The Rocchio classifier is a classifier based on tf-idf, initially proposed as a relevance feedback algorithm [6] but also used in the area of text classification [2]. Both keywords and documents are represented as vectors, and the closer a document $\overrightarrow{d_j}$ is to the keyword vector $\overrightarrow{w_j}$ the higher is the similarity between the document and the keyword in the vector space.

A document *j* is represented as $\vec{d}_j = (d_1,...,d_T)$, where each dimension is the probability of term t_i in the document multiplied by the term's inverse document frequency $IDF(t_i)$,

$$d_i = P(t_i \mid d) \cdot IDF(t_i).$$

The inverse document frequency is defined as the logarithm of the inverse of the probability of a term over the entire collection $\mathcal{D}$,

$$IDF(t_i) = \log\left({}^{1}\!/\!_{P(t_i \mid \mathcal{D})} \right).$$

A keyword is then represented by a vector $\overrightarrow{w_j}$, computed as the difference of the averages of the vectors wrt positive example documents $\mathcal{D}_{w_j}$ and negative example documents $\mathcal{D}_{\overline{w_j}}$,

$$\overrightarrow{w_j} = \frac{1}{\left|\mathcal{D}_{w_j}\right|} \sum_{\vec{d} \in \mathcal{D}_{w_j}} \frac{\vec{d}}{\left\|\vec{d}\right\|} - \frac{1}{\left|\mathcal{D}_{\overline{w_j}}\right|} \sum_{\vec{d} \in \mathcal{D}_{\overline{w_j}}} \frac{\vec{d}}{\left\|\vec{d}\right\|}.$$

For retrieval scenarios, documents are ranked according to their similarity to the queried keyword. We use the cosine similarity measure as a baseline,

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

$$\mathrm{sim}\left(\overrightarrow{w_j},\vec{d}\right)=\frac{\sum_{i=1}^{T} w_{ji}\cdot d_i}{\sqrt{\sum_{i=1}^{T}\left(w_{ji}\right)^2}\cdot\sqrt{\sum_{i=1}^{T}\left(d_i\right)^2}}.$$

Experiments in section 5 illustrate the application of the Rocchio classifier to semantic image retrieval.

4. NAÏVE BAYES CLASSIFIER

The naïve Bayes text classifier results from the direct application of Bayes law and from the use of strong independence assumptions between terms in a document,

$$P\left(w_j \mid d\right)=\frac{p\left(w_j\right)p\left(d=\{t_1,\ldots,t_T\}\mid w_j\right)}{\sum_j p\left(d\mid w_j\right)}. \qquad (1)$$

A document can be represented as an event model of term presence or term count, leading to the choice of a binomial or multinomial model respectively, see [5]. We choose the multinomial distribution as the binomial distribution is too limiting given the probabilistic nature of our terms.

Using the multinomial model of events, we express "the number of times a term occurs in a document" as $P\left(t_i \mid d\right)\cdot T$. Since T, the vocabulary size, is quite large we achieve enough granularity to represent the presence of visual keywords as integers. Given this, the probability of a document d given a keyword w_j is expressed as a multinomial over all terms,

$$P\left(d \mid w_j\right)=P(|d|)|d|!\prod_{i=1}^{T}\frac{P\left(t_i \mid w_j\right)^{P(t_i|d)\cdot T}}{\left(P\left(t_i \mid d\right)T\right)!},$$

where $P(|d|)$ is the probability of document d having length $|d|$. Note that when plugging the multinomial distribution into Equation (1) the term $1/\left(P\left(t_i \mid d\right)T\right)!$ is canceled. In addition to this, all documents are assumed to have the same length, meaning that T, $|d|!$ and $P(|d|)$ are constants, which allow us to discard them and keep only the proportionality relation

$$P\left(d \mid w_j\right)\propto\prod_{i=1}^{T}P\left(t_i \mid w_j\right)^{P(t_i|d)\cdot T}.$$

Now, we are left with the task of computing the probability of a term t_i for a given keyword w_j

$$P\left(t_i \mid w_j\right)=\frac{\sum_{d_k\in\mathcal{D}_{w_j}}P\left(t_i \mid d_k\right)}{\sum_{i=1}^{T}\sum_{d_k\in\mathcal{D}_{w_j}}P\left(t_i \mid d_k\right)}.$$

At this point the complete expression of the multinomial naïve Bayes model can be written as

$$P\left(w_j \mid d\right)=\frac{P\left(w_j\right)\prod_{i=1}^{T}P\left(t_i \mid w_j\right)^{P(t_i|d)\cdot T}}{\sum_{w'\in\mathcal{W}}P(w')\prod_{i=1}^{T}P\left(t_i \mid w'\right)^{P(t_i|d)\cdot T}}.$$

In retrieval scenarios, documents are ranked according to their probability for the queried keyword. In annotation scenarios, documents are labelled with the set of keywords that maximize

$$\mathrm{w}(d)=\arg\max_{w_j\in W}P\left(w_j \mid d\right).$$

When formulating naïve Bayes in the log-odds space,

$$\log\frac{P\left(w_j \mid d\right)}{P\left(\overline{w_j} \mid d\right)}=\log\frac{P\left(w_j\right)}{P\left(\overline{w_j}\right)}+\sum_{i=1}^{T}P\left(t_i \mid d\right)\log\frac{P\left(t_i \mid w_j\right)}{P\left(t_i \mid \overline{w_j}\right)},$$

we see that it is a linear model, and in annotation problems we avoid decision thresholds.

5. EXPERIMENTS AND DISCUSSION

We used the same set of 4500 Corel images for training and 500 images for testing as in [1, 3]. Each image is annotated with 1 to 5 keywords from a set of 179 keywords. For each single keyword we used both classifiers to rank all test images and then computed the corresponding Average Precision. The mean of all AP (MAP) versus the vocabulary size is depicted on Figure 1. As was expected, precision increases with the vocabulary size. Curve irregularities are caused by the way the visual vocabulary is obtained, see [4] for details. The Rocchio tf-idf and naïve Bayes classifiers achieve a maximum of 22.7% and 25.2% MAP respectively, which is comparable to the 24.8% obtained in [3].

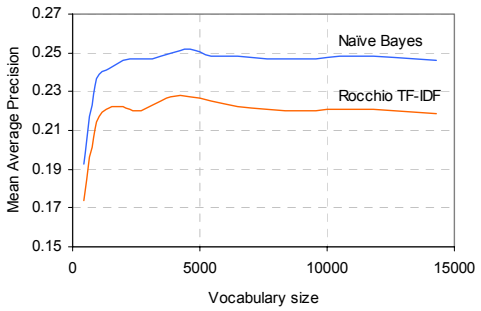

Figure 1 – Retrieval accuracy for 1 keyword.

This experiment shows that simple text based techniques, such as tf-idf, the Rocchio classifier and naïve Bayes, can be successfully applied to semantic image retrieval when we use a high-dimensional visual vocabulary, opening the door to the application of other text based techniques.

6. REFERENCES

[1] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," ECCV, Copenhagen, Denmark, 2002.

[2] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," ICML, Nashville, US, 1997.

[3] V. Lavrenko, R. Manmatha, J. Jeon, "A model for learning the semantics of pictures," NIPS, Vancouver, Canada, 2003.

[4] J. Magalhães, S. Rüger, "Logistic regression of generic codebooks for semantic image retrieval," CIVR Phoenix, AZ, USA, 2006.

[5] A. McCallum, K. Nigam, "A comparison of event models for naive Bayes text classification," AAAI Workshop on Learning for Text Categorization, 1998.

[6] J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System: Experiments in Automatic Text Retrieval*, G. Salton, Ed.: Prentice-Hall, 1971, pp. 313-323.