PLSA ON LARGE SCALE IMAGE DATABASES

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany

Malcolm Slaney
Yahoo! Research
Santa Clara, CA 95054
USA

ABSTRACT

The web and image repositories such as Fickr™ are the largest image databases in the world. There are billions of images on the web, and hundreds of million high-quality images in image repositories. Currently, these images are indexed based on manually-entered tags and individual and group usage patterns. In this work we a exploring a third information dimension: image features. We are exploring probabilistic latent semantic analysis in order to infer which visual patterns describe each object. We wish to build models that connect words and image features, and use content features and tags to better find similar images.

Index Terms— large scale image retrieval, probabilistic semantic analysis.

1. INTRODUCTION

ATTENTION: THIS IS A DUMMY SUBMISSION. The final paper will be submitted to Shih-Fu Chang before the internal deadline of Oct.- 31st. On the last page there is a rough outline the final paper will be about.

2. FORMATTING YOUR PAPER

All printed material, including text, illustrations, and charts, must be kept within a print area of 7 inches (178 mm) wide by 9 inches (229 mm) high. Do not write or print anything outside the print area. The top margin must be 1 inch (25 mm), except for the title page, and the left margin must be 0.75 inch (19 mm). All *text* must be in a two-column format. Columns are to be 3.39 inches (86 mm) wide, with a 0.24 inch (6 mm) space between them. Text must be fully justified.

3. PAGE TITLE SECTION

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

4. TYPE-STYLE AND FONTS

To achieve the best rendering both in the proceedings and from the CD-ROM, we strongly encourage you to use Times-Roman font. In addition, this will give the proceedings a more uniform look. Use a font that is no smaller than nine point type throughout the paper, including figure captions.

In nine point type font, capital letters are 2 mm high. If you use the smallest point size, there should be no more than 3.2 lines/cm (8 lines/inch) vertically. This is a minimum spacing; 2.75 lines/cm (7 lines/inch) will make the paper much more readable. Larger type sizes require correspondingly larger vertical spacing. Please do not double-space your paper. True-Type 1 fonts are preferred.

The first paragraph in each section should not be indented, but all following paragraphs within the section should be indented as these paragraphs demonstrate.

5. MAJOR HEADINGS

Major headings, for example, "1. Introduction", should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (".") after the heading number, not a colon.

5.1. Subheadings

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

5.1.1. Sub-subheadings

Sub-subheadings, as in this paragraph, are discouraged. However, if you must use them, they should appear in lower case (initial word capitalized) and start at the left margin on a separate line, with paragraph text beginning on the following line. They should be in italics.

6. PRINTING YOUR PAPER

Print your properly formatted text on high-quality, 8.5 x 11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (12 mm) empty at the BOTTOM of the page and follow the top and left margins as specified. If the last page of your paper is only partially filled, arrange the columns so that they are evenly balanced if possible, rather than having one long column.

7. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

8. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Do not use any colors in illustrations.

9. FOOTNOTES

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

10. COPYRIGHT FORMS

You must include your fully completed, signed IEEE copyright release form when you submit your paper. We **must** have this form before your paper can be published in the proceedings. The copyright form is available as a Word file, a PDF file, and an HTML file. You can also use the form sent with your author kit.

11. REFERENCES

List and number all bibliographical references at the end of the paper. The references can be numbered in alphabetic order or in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [1].

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title," *Journal*, Publisher, Location, pp. 1-10, Date.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, *Book Title*, Publisher, Location, Date.

What is the starting point?
We have a database in size only few people have at their disposal, and we want to exploit this fact. In addition, we want to exploit that the database is partially tagged. However, we must use a technique that can tolerate a large fraction of noise/incorrect labels since based on visual similarity these tags might by large look incorrect.
[I have browsed through images with the same tags such as “Christmas”. Less than 5% of the images have anything to do with Christmas. I have identified a few tags where the associated images show a common theme. They are in the table at the end. And, yes, I agree that ‘Christmas’ just specifies the time, when the images were shot, not their visual content.]

Premise:
pLSA is a very promising technique to identify concepts in data. So far -- in the image domain -- pLSA has only been applied to small data sets up to a few thousand images. We expect that some aspects that are ignored on small image databases such as the derivation of representative visual words might become important on large image sets. Also, some findings and proposed algorithms for small image data sets will not hold on large database sets. In this paper we present our initial findings on a database with more than a million images.

We focus exclusively on improving image retrieval based on image similarity as perceived by humans. Thus we will employ the following query paradigm and evaluation scheme:

Query paradigm & Evaluation / Performance metric:
First we select 12 distinct categories from our database (see table below). In each category we selected “randomly” 5 representative queries images (60 in total)

#	OR list of tags	# of image
1	wildlife animal animals cat cats	30477
2	dog dogs	26119
3	bird birds	21284
4	flower flowers	28819
5 6	graffiti sign signs (graffiti sign signs)	23318 14489 (36628)
7	surf surfing	30001
8	Night	34001
9	food	
10	Building buildings	17303
11	Goldengate goldengatebridge (+bridge bridges)	24364 (35637)
12	baseball	12390
	TOTAL SUM	188476

We use different techniques to return the top 20 most similar images. Similarity is purely judges by humans. No rules about what constitutes visual similarity are given to the subjects. 20 test people have to rank the results of the various techniques. In other words, each test subject gets the printouts of the top 20 most similar images (tiled 5 by 4 on one sheet of paper) for each retrieval method and must bring the retrieval results (i.e., the printouts) into an order from best to worst. We compute one combined score over all test queries (60 in total) to assign a single performance number to each algorithm: the average rank position.

As baseline technique we use (a) the tags + random selection and (b) color coherence vectors (CCVs). This is compared to plain vanilla pLSA and pLSA with active learning.

<u>Exp. 1 – Visual Words</u>
Given: C_N = # of categories; C = $\{c_i\}$ = set of categories (currently 12 categories); W_N = # of visual words

Goal: Derive set W = $\{w_j\}$ = set of visual words; $|W| == W_N$

Approaches:
W_N visual words are needed. We investigate three ways to determine visual words:

(a) Derive (W_N/C_N) visual words by means of K-means clustering within each category using KN sample features → result: $W_N = C_N * (W_N/C_N)$ visual words in total
(b) Select C_N times randomly KN sample features from the set of all features. Apply K-means clustering to each set of KN samples to derive (W_N/C_N) visual words → result: $W_N = C_N * (W_N/C_N)$ visual words in total.
(c) Select randomly W_N sample features from the set of all features → result: W_N visual words in total.

Based on the result of the performance metric between (a) and (b) we can decide whether tags provide useful information for deriving visual words. Based on (c) compared to (a) and (b) we can decide whether K-means clustering is really worth the effort. We use pLSA as the retrieval technique in all three experiments.

Reasoning behind experiments (c):
Is K-means clustering on large databases necessary? At the extreme we can postulate that as the size of the database grows, the feature vectors will be uniformly distributed. Thus, if 1 million samples are randomly selected from a uniform distribution and then clustered into e.g., 1K clusters, the result should statistically not differ from selecting directly randomly 1K feature vectors as cluster centers (= visual words).
This is something we can test and is very important in practices. Clustering is the slowest part in the learning algorithm.

Intuitively, I believe that there is still some structure because images created by humans are biased and thus the features should not be totally uniformly distributed. In that sense first selecting a larger set and cluster them should help to find common visual words and avoid using two visual words for the same thing. Thus, with the same number of words, a larger diversity is captured through clustering. But how many are need if W visual words are requested? Do I need 10 * W or 100 * W or only 5*W feature vectors.
→ Create graph where # of input features for clustering vs. performance is plotted.

Using the tags:
Does performance improve if the visual words are chosen not by just randomly sampling the features space, but by extracting them

from labeled subset of images? For instance, if a total of 2400 visual words are required, 200 for each distinct category could be generated and combined to from the 2400 visual words.

By how much does it improve? If this works it shows that even largely incorrect labels/tags carry information and thus improves results.

Exp. 2 – pLSA

Given: W_N = # of visual words; W = $\{w_j\}$ = set of visual words = $\{W^c\}$ = $\{w^c_j\}$

Goal: Perform visual similarity retrieval using P(z|d) to compute similarity score (= select images with the most similar concept distribution)

Approaches:

(a) Create term-document matrix based on W; learn pLSA; retrieve similar documents based on P(z|d) similarity (plain vanilla pLSA) (Euclidian or cosine similarity metric?)
(b) Compare pLSA to using p(w|d) directly. Is there a performance difference? (Euclidian or cosine similarity metric?)
(c) Interpret P(z|d) as a feature vector and apply active learning with support vector machine. Allow 3 rounds of feedback with just 20 images (5x4) each. Then evaluate the result.
(d) Baseline methods: tags + randomly selected images
(e) Baseline methods: use color coherence vectors (CCVs)

The active learning approach will be very simple to add since we have the code for it. Is should significantly improve retrieval. Something users are interested in.