Concept Content Based Wikipedia WEB Image Retrieval using CLEF VCDT 2008

Zhongqiu ZHAO & Herve GLOTIN
Laboratoire des sciences de l'information et des systemes
UMR CNRS & Universite' Sud Toulon-Var France
glotin@univ-tln.fr, zhongqiuzhao@gmail.com

Abstract

One challenge for this Wikipedia task is the training of visual models. We propose in this paper to link each topics one or few visual concepts of the Visual Concept Detection (VCDT) CLEFimage08 task, even if three topics do not fit VCDT concepts. We use the same models and features than in our VCDT systems. We show that our visual IMG NOFB run is the second best model in this campaign for this run type. So it can be concluded that our VCDT visual concept partly fit this task. Moreover we show that even a simple boolean text analysis overcomes the best IMG NO FEEDBACK run, which has 0.0037 MAP, against 0.399 for our TXT NOFB text run. This emphases the fact that visual retrieval for Wiki task is very difficult.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Infor- mation Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Languages–Query Languages

General Terms

Measurement, Performance, Experimentation

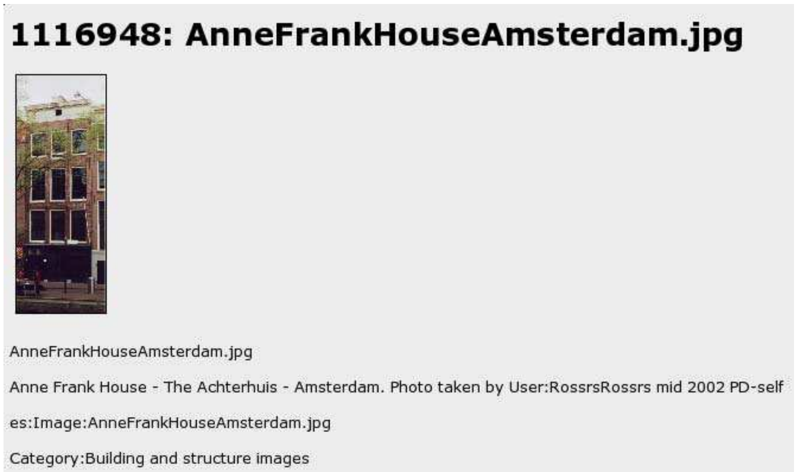
Keywords

Rank Fusion, Image Retrieval

1 Introduction

The ImageCLEF's 2008 wikipediaMM task provides a test bed for the system-oriented evaluation of visual information retrieval from a collection of Wikipedia images. The aim is to investigate retrieval approaches in the context of a larger scale and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs.

The given images are associated with unstructured and noisy textual annotations in English. This is an ad-hoc image retrieval task; the evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task and the ImageCLEFphoto task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e. topics are not known to the system in advance). The goal of the simulation is: given a textual query (and/or sample images and/or concepts) describing a user's



1116948: AnneFrankHouseAmsterdam.jpg

AnneFrankHouseAmsterdam.jpg

Anne Frank House - The Achterhuis - Amsterdam. Photo taken by User:RossrsRossrs mid 2002 PD-self

es:Image:AnneFrankHouseAmsterdam.jpg

Category:Building and structure images

Figure 1. An example image and its associated metadata

(multimedia) information need, find as many relevant images as possible from the wikipedia image collection.

The characteristics of the (INEX MM) wikipedia image collection allow for the investigation of the following objectives: how well do the retrieval approaches cope with larger scale image collections? How well do the retrieval approaches cope with noisy and unstructured textual annotations? How well do the content-based retrieval approaches cope with images that cover diverse topics and are of varying quality? How well can systems exploit and combine different modalities given a user's multimedia information need? Can they outperform monomodal approaches like query-by-text, query-by-concept or query-by-image? In the context of INEX MM 2006-2007, mainly text-based retrieval approaches have been examined. This task is done to attract more visually-oriented approaches and most importantly, multimodal approaches that investigate the combination of evidence from different modalities.

In this paper we present our strategy to retrieve relevant documents using a concept-content-based retrieval method in reusing the Visual Concept models we built for the CLEF VCDT 2008 task, without exclusively training the svms for this task. Then we test a very simple text analysis for a multimodal fusion. Results are showing that the visual process works badly, and does not contribute to the fusion.

2 Material

In 2008, ImageCLEF wikipediaMM used the image collection created and employed by the INEX Multimedia (MM) Track (2006-2007). This (INEX MM) wikipedia image collection contains 151,519 .jpeg and .png images that cover diverse topics of interest. Further information about the image collection can be found in Westerveld and al. [2].

Images were provided by wikipedia users. Each image is associated with user-generated alphanumeric, unstructured metadata in English. These metadata usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image. These descriptions are highly heterogeneous and of varying length. Figure 1 provides an example image and its associated metadata.

3 Using Visual Concepts Detection (VCDT) Clef08 models

A possible source of information to help participants in the retrieval tasks was, For each image, the classification scores for the 101 different MediaMill [1] concepts provided by University of

Table 1. The table of all the wiki concepts and the links with VCDT CLEF 2008 topics

WIKI CONCEPTS	**VCDT TOPICS**
'aircraft'	'sky'
'boat'	'water'
'building'	'buildings'
'car'	'road'
'cartoon', 'drawing_cartoon', 'drawing', 'graphics'	NONE
'charts'	NONE
'cloud'	'partly cloudy' or 'overcast'
'crowd'	'person'
'desert'	NO 'vegetation' and NO 'water' and NO 'beach' and NO 'buildings'
'dog'	'animal'
'female'	'person'
'grass'	'vegetation'
'hu_jintao'	'person'
'maps'	NONE
'mountain'	'mountains'
'outdoor'	'outdoor'
'people', 'people_marching'	'person'
'sports'	'person'
'studio'	'day'
'tree'	'tree'
'vegetation'	'vegetation'
'waterbody'	'water' and 'person'

Amsterdam (UvA). The UvA classifier is trained on manually annotated TRECVID video data and the concepts are selected for the broadcast news domain.

Anyway, visual concept information can be given by a visual concept training as specified in the VCDT CLEF task. We choose this last solution to get an idea of which level of visual information we can get from few visual concept classes.

There are 17 VCDT topics that correspond to most of the 75 concept-topics as depicted in Table 1. Only few wiki concepts are not described by VCDT : "map" and "charts" and "cartoons", this will decrease the global visual classification.

4 Visual Feature Extraction

An important step in content-based image retrieval (CBIR) system is the extraction of discriminant visual feature that are fast to compute. Information theory and Cognitive sciences can provide some inspiration for developping such feature.

Among the many visual features that have been studied, the distribution of color pixels in an image is the most common visual feature studied. The standard representation of color for content-based indexing in image databases is the color histogram. A different color representation is based on the information theoretic concept of entropy. Such entropic feature can simply equal the entropy of the pixel distribution of the image, as proposed in [3]. A more theoretical presentation of this kind of image entropy feature, accompanied by a practical description of its merits and limitations compared to color histograms, has been given in [4]. We propose [5,6] a new feature equal to the pixel 'profil' entropy. A pixel profil can be a simple arithmetic mean in horizontal (or vertical) direction. The advantage of such feature is to combine raw shape and texture representations in a low cpu cost feature. These feature, associated to mean and color std, reached the second best rank in the official ImagEval 2006 campaing (see www.imageval.org). For CLEF we extend these features using another projection to get the pixel profil. We use the harmonic mean of the pixel of each ligne or column. The idea is that the object or pixel region distribution, which is lost in arithmetic mean projection, could be partly catch by the harmonic mean. These two projections are then expected to give complementary and/or concept dependant informations. Details can be found in [5]. The extraction is summarized.

Let I be an image, or any rectangular subpart of an image.

For each normalized color ($L = R + G + B$, $r = R/L, andg = G/L$), we first calculate two orthogonal profils by the projections of the pixels of I. We consider two simple orthogonal projection axes : the horizontal axis X (noted Π_X), versus the vertical one Y (noted Π_Y). The projection operator is either the arithmetic mean (noted 'Ar', then the projection is noted Π_X^{Ar}), as illustrated in Figure 2, or the harmonic mean of the pixels on each column or each ligne of I (noted 'Ha', then we have Π_X^{Ha}).

Then, we estimate the probability distribution function (pdf) of each profil according to [7]. Considering that the sources are ergodic, we finaly calculate each PEF equal to the normalized entropy ($H(pdf)/log(\#bins(pdf))$). We detail below each steps of the PEF extraction.

Let be *op* the selected projection,
for each color of I of $L(I)$ lignes and $C(I)$ columns :

$\Phi_X^{op}(I) = \hat{pdf}(\Pi_X^{op}(I))$, over $nbin_X(I) = round(\sqrt{C(I)})$ bins,
where Π_X^{op} is the vertical projection with operator *op*,
$PEF_X(I) = H(\Phi_X^{op}(I))/log(nbin_X(I))$.

$\Phi_Y^{op}(I) = \hat{pdf}(\Pi_Y^{op}(I))$, over $nbin_Y(I) = round(\sqrt{L(I)})$ bins,
$PEF_Y(I) = H(\Phi_Y^{op}(I))/log(nbin_Y(I))$.

We add to these PEF_a the usual entropic feature :
$\hat{pdf}(I)$ = pdf of all the pixels of I over $nbin_{XY}(I) = nbin_X(I) * nbin_Y(I)$ bins,
$PEF_{.}(I) = H(\hat{pdf}(I))/log(nbin_{XY}(I))$.

And we finaly complete the PEF features by the usual mean and standard deviation of each normalized color of I.

We can calculate the PEF into three horizontal subimages (see Glotin Zhao VCDT CLEF papers for details). We note such PEF '='. We also calculate the PEF in three vertical subimages, we note these PEF '|||'.

For each, we have 3 bands and 3 different PEF for each of the 3 colors, plus their mean and variance, thus we have $3 * 3 * 3 + 3 * 3 * 2 = 45$ dimensions for '=' or for '|||' features. We note '+' the feature concatenation of '=' and '|||' features, which has then 90 dimensions. Considering the two mean type, the PEF concatenation without repetition of the mean and std color are quite compact with a total of 126 dimensions (= 2 (subimages type '=' or '|||') * 3 (bands by subimages type) * 3 (rgL) * 4 (=4 Π types = (X or Y) * (Ar or Ha)) + 1 (=$H(I)$) + 2 (= mean and std))).

5 Text Retrieval

Because of the very small length of the xml text data associated to the image, we selected almost 500 words defining the requests. We count these word occurences in each xml file. The text score for each xml file and each query is then simple the sum of theses occurences for each selected words of the query. This lead to a very fast text classification : for the whole 150 000 xml files only few minutes are necessary on a pentium IV. Our goal was not to produce a very efficient text retrieval system, but to compare our visual concept approach to the most basic text retrieval.

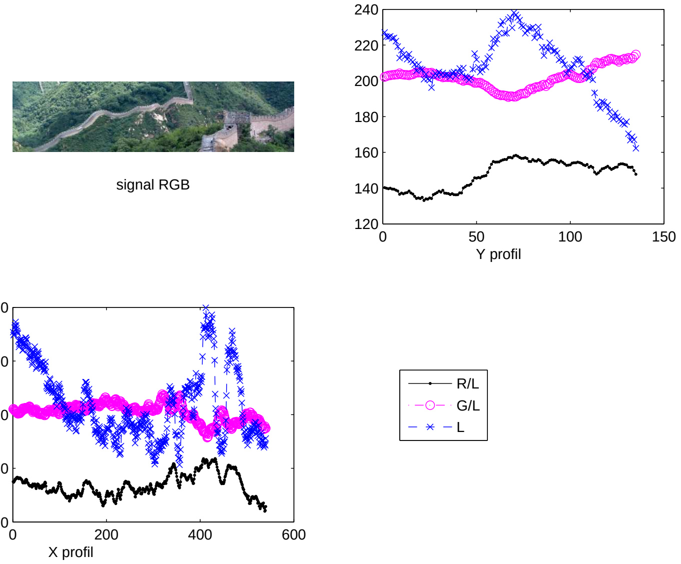


Figure 2: Illustration of the horizontal and vertical profils using simple arithmetic projection (or sum) of each normalized color $r = R/L, g = G/L, L = R + G + B$.

6 Experiments

The support vector machine (SVM) [8] first maps the data into a higher dimensional input space by some kernel functions, and then to learn a separating hyperspace to maximize the margin. Currently, because of its good generalization capability, this technique has been widely applied in many areas such as face detection, image retrieval, and so on. The SVM is typically based on an ε-insensitive cost function, meaning that approximation errors smaller than will not increase the cost function value. This results in a quadratic convex optimization problem. So instead of using an ε-insensitive cost function, a quadratic cost function can be used. The least squares support vector machines (LS-SVM) [9] are reformulations to the standard SVMs which lead to solving linear KKT systems instead. It is computationally attractive.

In our experiments, the RBF kernel

$$K(x_1 - x_2) = exp(-|x_1 - x_2|^2/\sigma^2)$$

is selected as the kernel function of our LS-SVM. So there is a corresponding parameter, σ , to be tuned. A large value of σ^2 indicates a stronger smoothing. Moreover, there is another parameter, γ, needing tuning to find the tradeoff between to stress minimizing of the complexity of the model and to stress good fitting of the training data points.

We trained 100 SVMs with different parameter values for each topic, and selected the best SVM using the validation set.

For Image or Text classification we do not make any Feedback nor Query Expansion techniques. So our runs are for NOFB type. In our experiments, we computed the average of the ranks of TXT and VISUAL. The process we adopt to implement the image retrieval in photo task is shown in Figure 3 and depicted as the following steps:

Step 1) According to the keywords of each topic, perform the text retrieval on the WIKI XML text database, and then get the rank result which is called Rank-Text.

Step 2) Split the VCDT labeled image dataset into 2 sets, namely training image dataset and

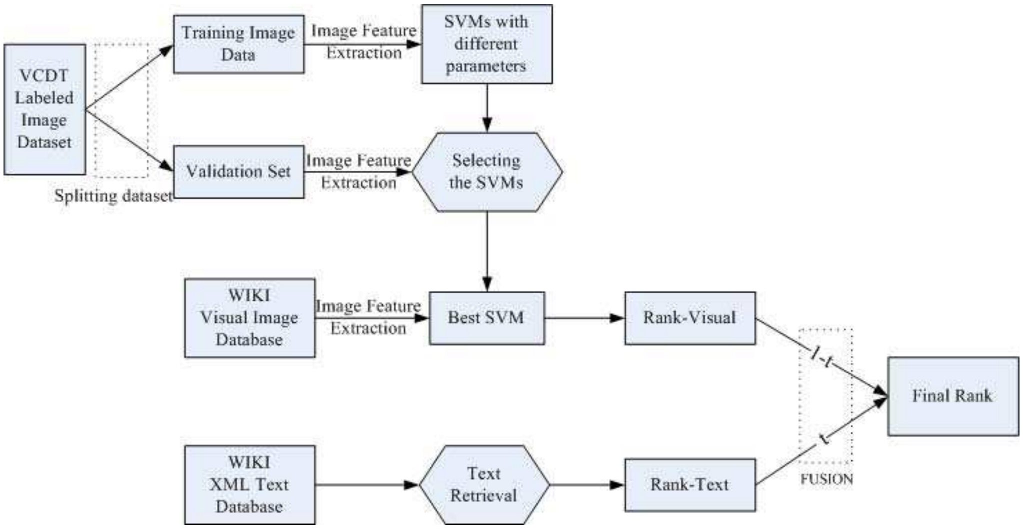


Figure 3. The Framework for Image Retrieval of Each Topic

validation set.

Step 3) Extract the visual features from the training image data using our extraction method; train and generate lots of SVMs with different parameters.

Step 4) Use the validation set to select the best one among the SVMs.

Step 5) Extract the visual features from the WIKI visual image database using our extraction method; use the best SVM as the tool to perform the image retrieval and produce the rank result called Rank-Visual.

Step 6) Merge Rank-Visual and Rank-Text into Final Rank using the weights, where 't' denotes the text ratio (t is fixed = 0.5).

7 Results

The results for the wikipediaMM task have been computed with the trec_eval tool (version 8.1). The submitted runs have been corrected (where necessary) so as to correpond to valid runs in the correct TREC format. The following corrections have been made: The runs comply with the TREC fomat as specified in the submission guidelines for the task When a topic contains an image example that is part of the wikipediaMM collection, this image is removed from the retrieval results, i.e., we are seeking relevant images that the users are not familiar with (as they are with the images they provided as examples). When an image is retrieved more than once for a given topic, only its highest ranking for that topic is kept and the rest are removed (and the ranks in the retrieval results are appropriately fixed).

The interpolated recall precision averages are shown in Figure 4, and the summary statistics for the runs sorted by MAP are shown in Table 2, from which we see that TXT method did the best.

8 Conclusion

Our visual IMG NOFB run is the second best model in this campaign. So it can be concluded that our PEF VCDT visual concept partly fit this task. We show that even a simple text analysis

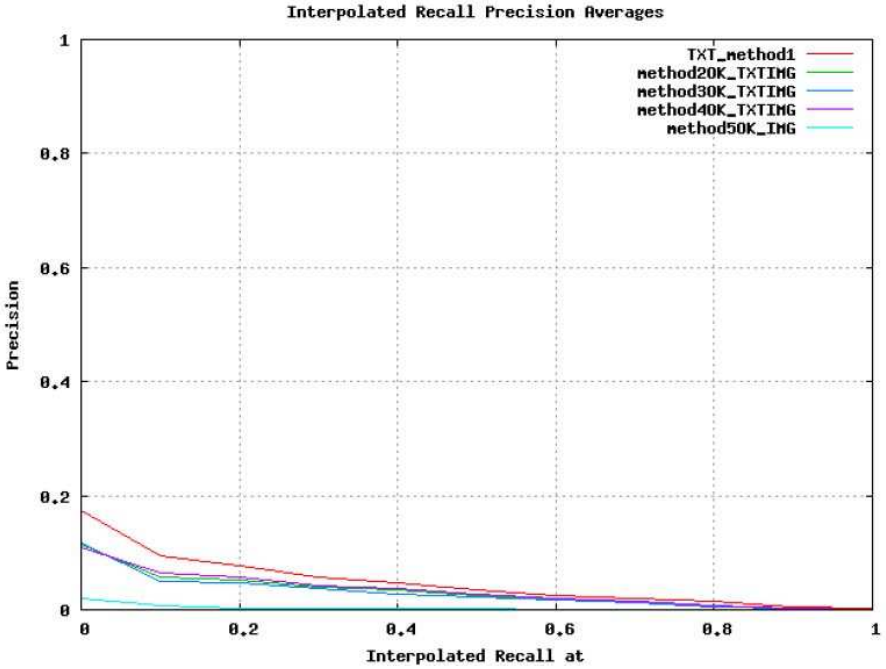


Figure 4. Interpolated Recall Precision Averages

Team	**Run**	**Modality**	**MAP**	**P@5**	**P@10**	**R-prec**	**Bpref**
imperial	visual only run AUTO NOFB	IMG	0.0037	0.0187	0.0147	0.0108	0.0086
LSIS utoulon	LSIS-IMG-AUTO-NOFB	IMG	0.0020	0.0027	0.0027	0.0049	0.0242
upmc-lip6	visual only run AUTO NOFB	IMG	0.0010	0.0107	0.0080	0.0053	0.0038
LSIS utoulon	LSIS-TXT-method1	TXT	0.0399	0.0507	0.0467	0.0583	0.0662
LSIS utoulon	LSIS4-TXTIMG-AUTO-NOFB	TXTIMG	0.0296	0.0347	0.0307	0.0421	0.0578
LSIS utoulon	LSIS-IMGTXT-AUTO-NOFB	TXTIMG	0.0260	0.0267	0.0253	0.0349	0.0547
LSIS utoulon	LSIS-TXTIMG-AUTO-NOFB	TXTIMG	0.0233	0.0533	0.0480	0.0300	0.0542

Table 2. The run results, with the 2 other runs of the campaign with visual NOFB runs showing that our visual concept approach is good for this run type. The basic text retrieval and fusion are also given for comparison.

overcomes the best IMG NO FEEDBACK from Imperial College which has 0.0037 MAP, against 0.399 for our TXT NOFB text run. This emphases the fact that visual retrieval for Wiki task is very difficult. This can explain why our 3 basic fusions methods TXTIMG did not improve the TXT run. One can then conclude that the Feedback seems necessary for such task.

Acknowledgment

This work was partially supported by the French National Agency of Research (ANR-06-MDCA-002).

References

[1] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In MULTIMEDIA 06: Proceedings of the 14th annual ACM international conference on Multimedia, pages 421-430, New York, NY, USA, 2006. ACM Press.

[2] T. Westerveld and R. van Zwol. The INEX 2006 Multimedia Track. In N. Fuhr, M. Lalmas, and A. Trotman, editors, Advances in XML Information Retrieval:Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI). Springer-Verlag, 2007.

[3] M. Jagersand, Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach, in Proc. of 5th International Conference on Computer Vision, 1995.

[4] Iyengar J. Zachary, S.S and Barhen J., Content based image retrieval and information theory: A generalized approach, in Special Topic Is- sue on Visual Based Retrieval Systems and Web Mining, Journal of the American Society for Information Science and Technology, 2001, pp. 841-853.

[5] H. Glotin, Z. Zhao, Profil Entropic visual Features for Visual Concept Detection in CLEF 2008 campaign, In Working Notes of ImageCLEF2008, Danmark, in conjuction with ECDL 2008.

[6] H. Glotin, "Robust Information Retrieval and perception for a scaled Lego-Audio-Video multi-structuration", Thesis of habilitation for research direction, University Sud Toulon-Var, 2007.

[7] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, Signal Processing, vol.16, no.3, pp. 233-246, 1989.

[8] Vapnik, V. 1998 Statistical learning theory. John Wiley, New York.

[9] Suykens, J.A.K. and Vandewalle, J. 1999. Least Squares Support Vector Machine Classifiers Neural Processing Letters, 9 (1999), 293-300.