

LSIS Scaled Photo Annotations: Discriminant Features SVM versus Visual Dictionary based on Image Frequency

Zhong-Qiu ZHAO^{1,2}, Herve GLOTIN¹, Emilie DUMONT¹

¹ University of Sud Toulon-Var, USTV, France

Lab. Sciences de l'Information et des Systemes LSIS, UMR CNRS 6168, La Garde-France

² School of Computer & Information, Hefei University of Technology, China

glotin@univ-tln.fr; zhongqiuzhao@gmail.com; emilie.r.dumont@gmail.com

Abstract

In this paper, we used only visual information to implement ImageCLEF2009 Photo Annotation Task. Firstly, we extract various visual features: HSV and EDGE histograms, Gabor, and recent Descriptor of Fourier and Profile Entropy Features. Then for each concept and features, we compute Linear Discriminant Analysis (LDA) to decrease the high dimension impact. Finally, we train support vector machines (SVMs), for which the outputs are considered as the confidences with which the samples belong to the concept.

Also we propose a second model, an improved version of a Visual Dictionary (VD), which is built by visual words extracted for frequency templates in the training set.

We describe the results of these 2 models, topics by topics, and we give perspectives for our VD method, that is more faster than SVM, and better than SVM for some topics. We also show that among the 19 teams, our SVM(LDA) run attains the AUC score of 0.721, and then occupies the 8th AUC rank among the 19 teams involved in this campaign, while our VD models would occupy the 10th rank.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Cross-Language Retrieval in Image Collections (ImageCLEF)—*ImageCLEFphoto*

Keywords

LDA, Visual Dictionary, Generalized Descriptor of Fourier, Profile Entropy Feature, SVM.

1 Introduction

ImageCLEF (<http://www.imageclef.org/ImageCLEF2009>) is the cross-language image retrieval track run as part of the Cross Language Evaluation Forum (CLEF) campaign. This track evaluates retrieval of images described by text captions based on queries in a different language; both text and image matching techniques are potentially exploitable.

Therefore, the task provides a training database of 5000 images which are labeled by 53 concepts which has a hierarchy. Only these data are used to train retrieval models. The test database consists of 13000 images, for each of which participating groups are required to determine the presence/absence of the concepts. More details on the dataset can be found in [13].

We use various features described in nexte section: PEF[2, 3], HSV and EDGE histograms, new Descriptor of Fourier [7], and Gabor. The we use an LDA to reduce their dimension. Finally, we use the Least Square support vector machine (LS-SVM) to produce concept similarity. Another original method called Visual Dictionary is proposed and implemented in section 5.

2 Visual Features

We use a new feature, the pixel 'profile' entropy (PEF) [2], giving the entropy of a pixel profiles in horizontal and vertical directions. The advantage of PEF is to combine raw shape and texture representations, with a low CPU cost feature, and already gave good performances (second best rank in the official ImagEval 2006 campaign (see www.imageval.org)).

Here we use extended PEF [3] using the harmonic mean of the pixel of each row or column. The idea is that the object or pixel region distribution, which is lost in arithmetic mean projection, could be partly represented by the harmonic mean. These two projections are then expected to give complementary and/or concept dependant information. PEF are computed into three equal horizontal and vertical image slices, yieding to a total of 150 dimensions.

We also use classical features : HSV and EDGE histograms, and Gabor, and recent Descriptor of Fourier robust to rotation [7]. We train our two models on these features that represent a total of 400 dimensions. We use LDA to reduce the feature dimensions as depicted in next section.

3 Linear Discriminant Analysis

In general, the LDA [11] is used to find an optimal subspace for classification in which the ratio of the between-class scatter and the within-class scatter is maximized. Let the between-class scatter matrix be defined as

$$S_B = \sum_{i=1}^c n^i (\bar{X}^i - \bar{X})(\bar{X}^i - \bar{X})^T$$

and the within-class scatter matrix be defined as

$$S_W = \sum_{i=1}^c \sum_{X_k \in C_i} (X_k - \bar{X}^i)(X_k - \bar{X}^i)^T$$

where $\bar{X} = (\sum_{j=1}^n X_j)/n$ is the mean image of the ensemble, and $\bar{X}^i = (\sum_{j=1}^{n^i} X_j^i)/n^i$ is the mean image of the i th class, n_i is the number of samples in the i th class, c the number of classes, and C_i the i th class. As a result, the optimal subspace, $E_{optimal}$ by the LDA can be determined as follows:

$$E_{optimal} = \arg \max_E \frac{|E^T S_B E|}{|E^T S_W E|} = [c_1, c_2, \dots, c_{c-1}]$$

where $[c_1, c_2, \dots, c_{c-1}]$ is the set of generalized eigenvectors of S_B and S_W corresponding to the largest generalized eigenvalues λ_i ; $i = 1, 2, \dots, c-1$, i.e.,

$$S_B E_i = \lambda_i S_W E_i, i = 1, 2, \dots, c-1$$

Thus, the feature vector, P , for any query face images, X , in the most discriminant sense can be calculated as follows:

$$P = E_{optimal}^T U^T X$$

In our image retrieval task, LDA output only 1 dimension since the classification problem for each concept is 2-class.

4 Fast classification using Least Squares Support Vector

In order to design fast image retrieval systems, we use the Least Squares Support Vector Machine (LS-SVM). The SVM [1] first maps the data into a higher dimensional input space by some kernel functions, and then learns 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 [4, 5]. The SVM is typically based on an ε -insensitive cost function, meaning that approximation errors smaller than ε 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) [6] are reformulations to the standard SVMs which lead to solving linear KKT systems instead, which is quite computationally attractive. Thus, in all our experiments, we will use the LS-SVMlab1.5 (<http://www.esat.kuleuven.ac.be/sista/lssvmlab/>).

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 stress minimizing of the complexity of the model and to stress good fitting of the training data points.

We set these two parameters as

$$\sigma^2 = [4 25 100 400 600 800 1000 2000]$$

and

$$\gamma = [4 8 16 32 64 128 256 512]$$

respectively. So a total of hundred SVMs were constructed for each SVM model, and then we selected the best SVM using the validation set.

5 Visual Dictionary Method

The visual dictionary is an original method to annotated images which is an improvement of the method proposed in [8]. We construct a Concept Visual Dictionary composed by visual words intended to represent semantic concept which consists of five steps :

- Visual elements. Images are decomposed into visual elements where a visual element is an image area, i.e. images are split into a regular grid.
- Representation of visual elements. We use the most classical and intuitive approach consisting in representing a visual word by usual features HSV, GABOR, EDGE, and also PEF [3], and DF [7].
- Global Visual Dictionary. For each feature, we cluster visual elements using the K-Means algorithm with a predefined number of clusters and using the Euclidean distance in order to group visual elements and to smooth some visual artifacts. And then, for each cluster, we select the medoid to be a visual word and to compose the visual dictionary of a feature.
- Image transcription. Based on the Global Visual Dictionary, we replace visual elements by the nearest visual word in the visual dictionary. And then, the image representation is based on the frequency of the visual words within the image for each feature.
- Concept Visual Dictionary. We select the most discriminative visual words for a concept given to compose a Concept Visual Dictionary. To filter the words, we use a entropy-based reduction, which is developed from work carried out in [10].

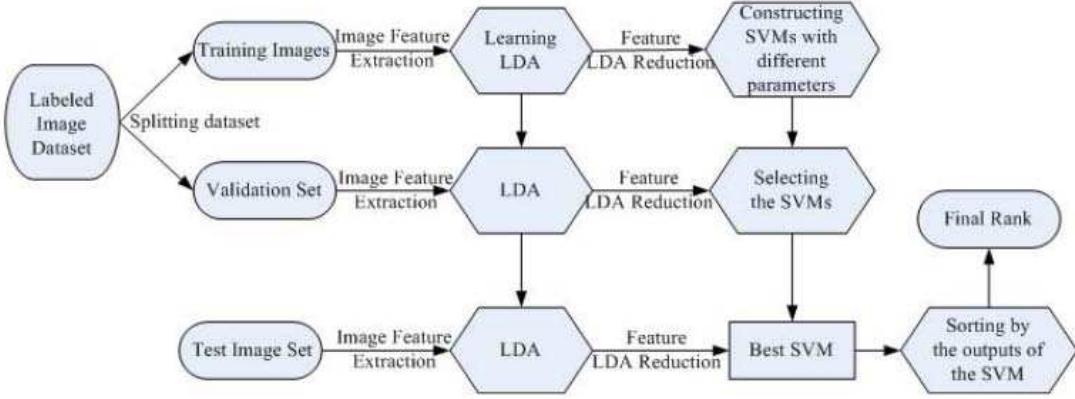


Figure 1: The framework for image retrieval of each topic

In a second step, we propose an adaptation of the common text-based paradigm to annotated images. We used the tf-idf weighting scheme [9] in the vector space model together with cosine similarity to determine the similarity between a visual document and a concept. To use this scheme, we represent an image by the frequency of the visual words within the image for different features : HSV, GABOR, PEF, EDGE and DF.

6 Experimental Results

The models based on SVM to implement the image retrieval in the task is shown in Figure 1 and contains the following steps:

Step 1) Split the VCDT labeled image dataset into 2 sets, namely training image dataset and validation set.

Step 2) Extract the visual features from the training image data using our extraction method; Learn and perform LDA reduction on these features; train and generate lots of SVM with different parameters.

Step 3) Use the validation set to select the best model

Step 4) Extract the visual features from the VCDT test image database using our extraction method; perform LDA reduction on these features; and then use the best model to find the best discriminant feature.

Step 5) Sort the test images by the distances from the positive training images and produce the final rank result.

The same train and development sets have been used for the VD and SVM training. We submitted five runs to the official evaluation, from which the two best are depicted here :

- Run SVM(LDA)

It consists in performing SVM on the LDA of [PEF150 + HSV + EDGE + DF] features to reduce the impact of the highdimension malediction. The test of 10K images and 50 topics costed 2 minutes on usual pentium IV PC.

- Run VD

Is a vector search system, using small icons from the images. The visual features are the HSV, edge, and Gabor. This model needs only 2 hours of training on a pentium IV 3Ghz, 4 GRam, and test is faster than SVM.

The Area Under the Curve (Receiving Operator Characteristic integral) for each topic and method are depicted in figure 2.

Table 1: Our two best submitted runs of ImageCLEF2009 Photo Annotation Task

RUN	LAB	Method	Average Equal Error Rate	AUC
SVM(LDA)	LSIS	SVM(LDA([PEF150+HSV+EDGE+DF]))	0.3308	0.7209
VD	LSIS	Visual Dictionary	0.3607	0.6855

Table 2: The team best runs of ImageCLEF2009 Photo Annotation Task (from the Official evaluations)

RANK	LAB	Average EER	AUC
1	ISIS	0.234476	0.838699
2	LEAR	0.249469	0.823105
3	FIR2	0.253566	0.817159
4	CVIUI2R	0.253296	0.813893
5	XRCE	0.267301	0.802704
6	bpacad	0.291718	0.773133
7	MMIS	0.312366	0.744231
* 8	LSIS (SVM(LDA))	0.330819	0.720931
9	IAM	0.330401	0.714825
10	LIP6	0.372169	0.673089
11	MRIM	0.383630	0.643450
12	AVEIR	0.440589	0.550866
-	Random	0.500280	0.499307
13	CEA	0.500495	0.469035
14	TELECOM ParisTech	0.526302	0.459922
15	Wroclaw	0.446024	0.220957
16	KameyamaLab	0.452374	0.164048
17	UAIC	0.479700	0.105589
18	INAOE	0.484685	0.099306
19	apexlab	0.482693	0.070400

We show in Table 2 that the SVM(LDA([PEF150+HSV+EDGE+DF])) is better than VD with AUC = 0.72. It is our best run, it occupies the 8th rank among the 19 participating teams. The same SVM(LDA) strategy has been applied on an another set of features (AVEIR group features described in [12]), but results to AUC = 0.50.

7 Conclusion

The SVM model has an higher average AUC than VD (0.722 against 0.682), but VD is lighter than SVM, and is, for some topics, better than it. The table 3 gives the list of worst and best topics for VD compared to SVM. The worst are for example "Snow, Winter, Sky, Desert, Beach ...", that are maybe topics with one clear visual representation, for example we can imagine a dominant color and texture for snow or sky ... On the contrary, VD is better than SVM for "Flower, Vehicle, Food, Autumn,..." that are maybe concepts with higher visual variations in color, texture... Thus it suggests that statistics of simpler visual concepts are maybe better modelized by SVM, while more complex visual concepts may be better represented by our Visual Dictionary model. The respective performances of these two models shall also be tied with the number of training samples. We currently investigate research on this promising improved VD, and we propose an optimal fusion with SVM in order to benefit of the properties of the both.

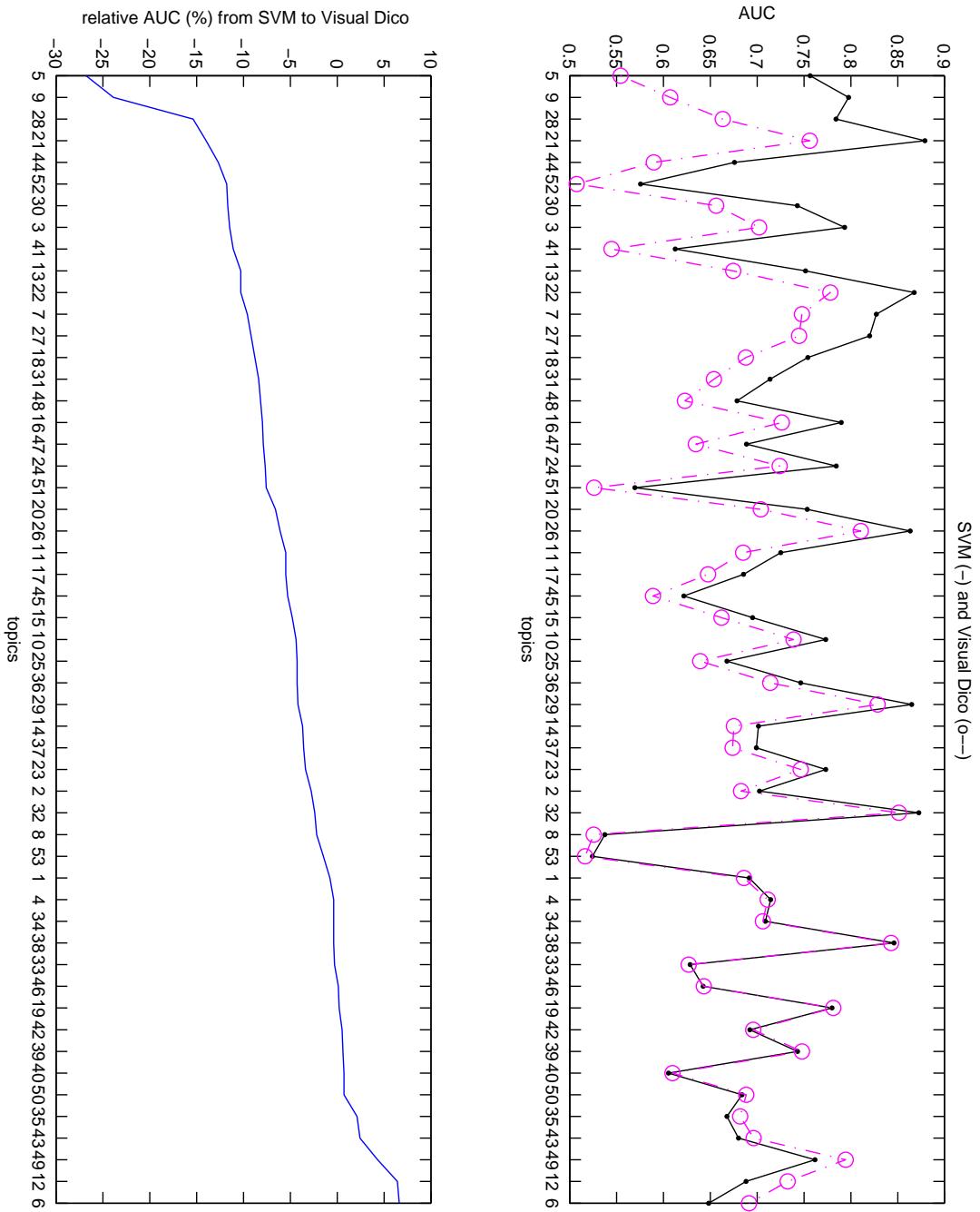


Figure 2: Area Under the Curve (AUC) evaluations for each topic and method. The number of each topic is the one given by the organizers. Here they are sorted according to the relative performance of VD against SVM. The absolute AUC for SVM and for VD (\circ) are given in the right figure, while their relative difference in percent is given in the left. The 10 worst and best topics are depicted in Table 3.

Table 3: Lists of the 10 worst and ten best topics for the VD method compared to SVM

Ten worst topics for VD	Ten best topics for VD
Snow	Flowers
Desert	Partly-Blurred
Day	Partly-Blurred
Sky	Motion-Blur
Single-Person	Vehicle
Overall-Quality	Macro
No-Visual-Time	No-Blur
Beach-Holidays	Food
Out-of-focus	Autumn
Winter	Citylife

Acknowledgment

This work was supported by French National Agency of Research (ANR-06-MDCA-002) and Research Fund for the Doctoral Program of Higher Education of China (200803591024).

References

- [1] Vapnik, V.: Statistical learning theory. John Wiley, New York (1998)
- [2] Glotin, H.: Information retrieval and robust perception for a scaled multi-structuration, Thesis for habilitation of research direction, University Sud Toulon-Var, Toulon (2007)
- [3] Glotin, H., Zhao, Z.Q., Ayache, S.: Efficient Image Concept Indexing by Harmonic & Arithmetic Profiles Entropy, IEEE International Conference on Image Processing, Cairo, Egypt, November 7-11, (2009)
- [4] Waring, C.A., Liu, X.: Face detection using spectral histograms and SVMs. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 35(3), 467–476 (2005)
- [5] Tong S., Edward, Chang : Support Vector Machine active learning for image retrieval, In Proceedings of the ninth ACM international conference on Multimedia, Canada, pp. 107–118 (2001)
- [6] Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers, In Neural Processing Letters, 9, 293–300 (1999)
- [7] Smach, F., Lemaitre, C., Gauthier, J.P., Miteran, J., Atri, M.: Generalized Fourier Descriptors with Applications to Objects Recognition in SVM Context, In 30, J. Math Imaging Vis 43–71 (2008)
- [8] Dumont E, Merialdo B. : Video search using a visual dictionary. In CBMI 2007, 5th International Workshop on Content-Based Multimedia Indexing, June 25-27, 2007, Bordeaux, France (2007).
- [9] Salton, G. and McGill, M. J. : Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA (1986)
- [10] Jensen R. and Shen Q.: Fuzzy-Rough Data Reduction with Ant Colony Optimization, In Fuzzy Sets and Systems (2004)
- [11] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces versus fisher faces. IEEE Trans. Pattern Anal. Machine Intell. 19, 711–720 (1997)

- [12] Glotin H. and al. : Comparison of Various AVEIR Visual Concept Detectors with an Index of Carefulness, In ImageClef09 proceedings (2009)
- [13] Nowak S., Dunker P. : Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task, CLEF working notes 2009, Corfu, Greece, (2009).