LSIS TREC VIDEO 2009 High Level Feature Retrieval using Compact Profile Entropy Descriptors

Herve GLOTIN, Zhongqiu ZHAO and Emilie DUMONT

Laboratoire des sciences de l'information et des systemes
UMR CNRS 6168 & Universite' Sud Toulon-Var, 83130, La Garde, France
glotin@univ-tln.fr, zhongqiuzhao@gmail.com

Abstract. We build a fast video shot retrieval system in the context of the NIST TREC Video 2009 evaluation campaign. We compare our efficient Profile Entropy Features (PEF) to usual features, using various classifiers. These PEF are derived using the projection in the horizontal and vertical orientations. These features are then fed to SVM or KNNG classifiers to produce the keyframe ranks, from which we can get the shot ranks. The experimental results show that our PEF features outperform other features such as EDGE, GABOR, HSV, and so on. Moreover, PEF are very compact and fast to compute, and thus may be improved in further video retrieval systems.

Key words: TRECVID09, High Level Feature Extraction, video retrieval, profile entropy, SVM

1 TRECVid 2009 High Level Feature Task

The High-Level semantic retrieval task concerns features or concepts such as "Indoor/Outdoor", "People", "Speech" etc., that occur frequently in video databases. The TRECVid HLF task [1] (http://www-nlpir.nist.gov/projects/tv2009/tv2009.html) contributes to work on a benchmark for evaluating the effectiveness of detection methods for semantic concepts. The task of high-level feature extraction is as follows: given the feature test collection composed of hundred of hours of videos, the common shot boundary reference for the feature extraction test collection, and the list of feature definitions, participants return for each feature the list of at most 2000 shots from the test collection, ranked according to the highest possibility of detecting the presence of the feature. Each feature is assumed to be binary, i.e., it is either present or absent in the given reference shot.

The 20 concepts of this task are shown as follows: 1 Classroom, 2 Chair, 3 Infant, 4 Traffic-intersection, 5 Doorway, 6 Airplane_flying, 7 Person-playing-a-musical-instrument, 8 Bus, 9 Person-playing-soccer, 10 Cityscape, 11 Person-riding-a-bicycle, 12 Telephone, 13 Person-eating, 14 Demonstration_Or_Protest, 15 Hand, 16 People-dancing, 17 Nighttime, 18 Boat_Ship, 19 Female-human-face-closeup, 20 Singing. The 100 hours used as test data for 2008 (tv8.sv.test)

are combined with 180 hours of new test data (tv9.sv.test) to create the 2009 test set for the search and feature tasks. This will allow participants to retest for progress in detection of some of the 2008 features against some of the 2008 test data.

These paper focuses on LSIS profile entropic descriptors for images, showing their performance in image and video indexing by evaluating on the high level feature extraction task of TRECVID 2009.

2 Profile Entropy Features definition

An important step in content-based image retrieval (CBIR) system is to quickly extract the discriminant visual features. Information theory and Cognitive sciences can provide some inspiration for developing such features.

Among the many visual features that have been studied, the distribution of color pixels of image is the most common one. The standard representation of color for content-based indexing in image databases is the color histogram. While a different color representation is based on the information theoretic concept of entropy. Such entropy feature can simply be equal to the entropy of the pixel distribution of the image, as proposed in [2]. 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 [3].

A new feature equal to the pixel 'profile' entropy has been proposed in [4][5], where a pixel profile 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. This feature, associated to mean and color STD, reached the second best rank in the official ImagEval 2006 campaign (see www.imageval.org).

Let I be an image (or a part of) of $L(I)$ rows, and $C(I)$ columns. The PEF are computed on these normalized RGB channels : $l = (R+G+B)/3$, $r = R/l$, and $g = G/l$. We consider the profiles of the orthogonal projections of the pixels to the horizontal X axis, noted Π_X^{op}, and to the vertical Y axis (Π_Y^{op}), where ^{op} is a projection operator. This one is either the arithmetic mean of the pixels (noted $\Pi_{\cdot}^{Ari}$), or their harmonic mean (noted $\Pi_{\cdot}^{Harm}$), as illustrated in Fig.1,2. Thus the length of a given profile is either $S = C(I)$ or $S = L(I)$.

Then, for each profile, we estimate its probability distribution function ($\hat{pdf}$) on N bins (where $N = round(\sqrt{S})$ as proposed in [7]).

For each channel, and each operator ^{op}, we compute :
$\Phi_X^{op}(I) = \hat{pdf}(\Pi_X^{op}(I))$. Considering that the sources are ergodic, we set PEF_X component to the normalised entropy of this distribution :
$PEF_X(I) = H(\Phi_X^{op}(I))/log(N)$,
where N the number of bins of the considered distribution, and H the usual entropy function. We compute similar PEF on Y axis :
$PEF_Y(I) = H(\Phi_Y^{op}(I))/log(N)$.

We set a third PEF component to the entropy of the direct distribution of all the pixels in I, $\hat{pdf}(I)$:
$PEF_B(I) = H(\hat{pdf}(I))/log(N)$,
where $N = round(\sqrt{L(I) * C(I)})$ bins.

The whole PEF features are the concatenation of PEF_X, PEF_Y and PEF_B[6], with the usual mean and standard deviation of each channel of I.

The PEF are computed on three horizontal (noted '=') or vertical ('‖‖') equal segmented subimages, and on the whole image. For exemple, for a given operator, we have the whole image plus the three '‖‖' subimages, and for each of the 3 channels we have $PEF_{X,Y,B}$, plus their mean and variance, thus we have $4 * 3 * (3 + 2) = 60$ dimensions. We note '#' the concatenation of '=' and '‖‖' PEF, without duplication.

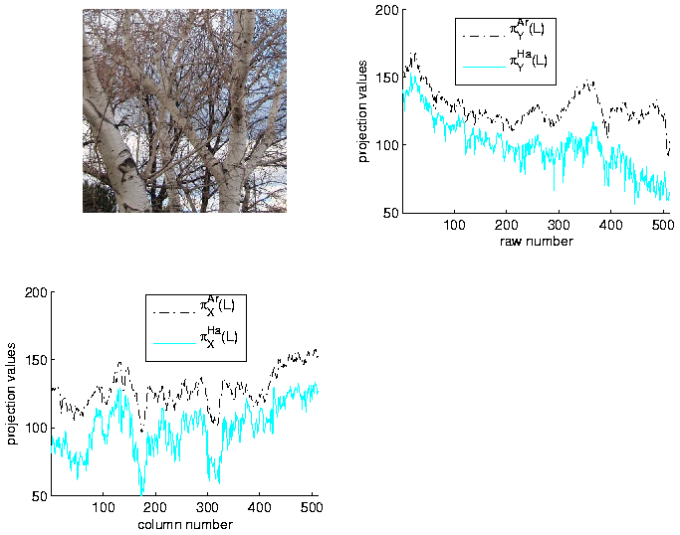

Fig. 1. Horizontal X (Bottom) and, vertical Y (Top Right) profiles using arithmetic (-.-) and harmonic (-) projections of the luminance of an image of a tree. It shows clearly the difference between the two projections for this structured pattern.

3 KNN comparison to hundreds of state of the art features

In this task, other features, such as HSV, EDGE, DF [16] are also used. We shared our PEF features with other teams from IRIM group. Table 1 shows the individual performance of each LSIS descriptor with up to 2 KNN classifiers from LIG [19]. These 2 classifiers were used for the evaluation of the descriptors (the

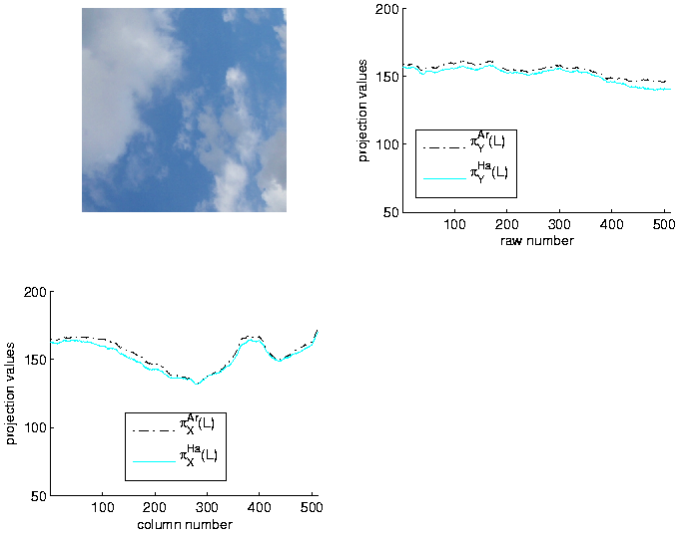

Fig. 2. Similar to Fig.1 but for an image of the concept sky : arithmetic and harmonic profiles are similar.

same classifiers were also used for producing the predictions of the TRECVID submission using the same descriptors):

– LIG_KNNC and LIG_KNNG: KNN-based classifier with hyper-parameters obtained by cross validation with an optimization respectively by concept or globally. The LIG_KNNC optimizes its parameters by cross-validation separately for each concept. The LIG_KNNG optimizes them globally.

The training and evaluation were done respectively on the development and test parts of the TRECVID 2007 collection. More details are given in [15]. For comparison, tests with randomly generated output indication a performance of 0.0022 ± 0.0005 for a random submission while a perfect submission would have a performance of 1.0000.

According to other features tested in IRIM [19], the typical MAP performance of a good "monomodal" descriptor is in the 0.0300-0.0500 range. This is about 20 times more than a random prediction but still 20 to 30 times less than a perfect prediction. So from the table, we can see that PEF performs better than other features such as EDGE, GABOR, and HSV, especially the PEF45 features with LIG KNNG classifier.

We also plotted the figure depicting the ratio between MAP and $\log 10(dim)$ (dim is the dimension of descriptors), as shown in Figure 3. It can be seen that LSIS PEF45 has good performance with relatively few dimensions.

Table 1. Performance of image descriptors from other teams. PEF45 denotes the PEF features exclude the harmonic ones, while PEF150 is the full concatenation (with certain redundancy) of harmonic and arithmetic operators

	Dims	LIG_KNNC	LIG_KNNG
LSIS_PEF45	45	0.0344	0.0382
LSIS_PEF150	150	0.0332	0.0330
LSIS_EDGE	72	0.0205	0.0213
LSIS_GABOR	60	0.0307	0.0328
LSIS_HSV	63	0.0239	0.0249

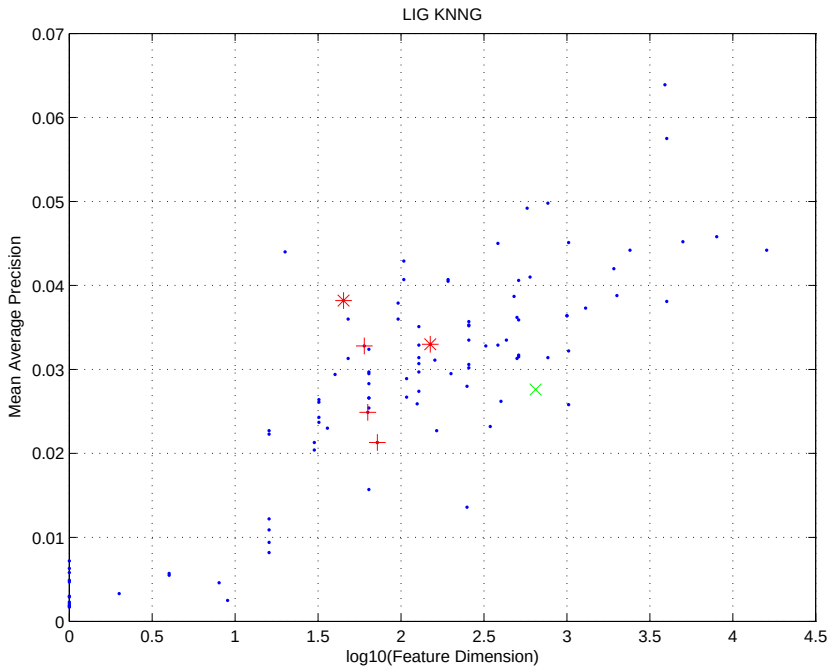

Fig. 3. Feature dimensions versus their MAP according to the LIG KNNG evaluations, for 200 features from TREC IRIM [19] consortium. The red '*' are (from left to right) the PEF45 and PEF150. The red '+' are (from left to right) the Gabor, HSV and EDGE histogram features. The average in dimension and MAP of all the IRIM features is the green 'X'. We see clearly that PEF45 is a particular feature in the top left, yielding to a good compactness property

4 Least Squares Support Vector Machines

In order to design fast video retrieval systems, we use the Least Squares Support Vector Machine (LS-SVM). The SVM [8][9] 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 [10][11]. 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) [12] 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 to 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 100 SVMs were constructed for each topic, and then we selected the best SVM using the validation set.

5 Submitted Run

We submitted 1 run, in which we trained a SVM model on the concatenation features [PEF150 HSV EDGE DF]. So we got the evaluation performance of mean inferred average precision equal to 0.028. The result details of this run are shown in Figure 4

From the results, it can be seen that the run performs well. In this run, the IAP for the concept "Person-eating" is the best among all concepts, which is much better than the average. And the IAPs for the concepts "Traffic-intersection" and "Person-riding-a-bicycle" are slightly better than the mean. The IAPs for the concepts "Classroom", "Infant", "Bus", "Telephone" and "Singing" are approximately equal to the mean. So the PEF features for the above concepts are more discriminant than the rest. The reason is that PEF features are the mixture of color and texture characters of images, and these concepts can be detected by the color or texture of the objects. For example, the texture of images of the concept "Person-eating" is much distinguished from other concepts.

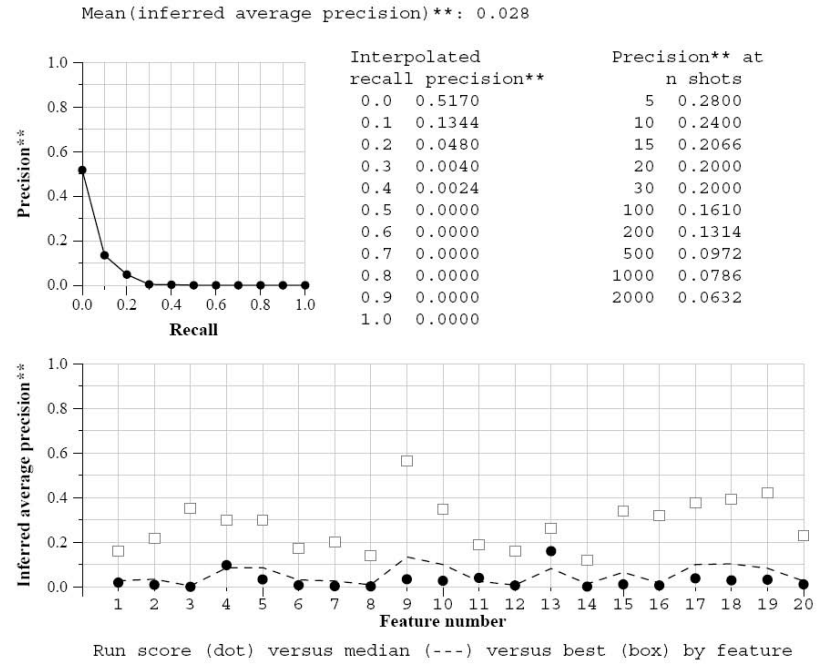

Fig. 4. The Inferred MAP performances of the A_LSIS_svm_PEF_4 run

6 Conclusions

Considering the results of inferred MAP, we can see that the PEF45 features that compiles texture and color informations in a compact vector, does the best, according to the KNNG classifier, compared to usal color or texture features, while PEF45 are of lowest dimensions. The TREC NIST official evaluation of the early fusion of these features (plus a Fourier descriptor), shows that our system is near the median of the TRECVID 2009, and significantly better for one topic. Further research will be conducted to enhance PEF.

Acknowledgment

This work was partially supported by the French National Agency of Research (ANR-06-MDCA-002). We also thank Stephane Ayache (LIF France) and Georges Quenot (LIG France) for their help on TREC files management.

References

1. Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722 (2006)
2. Jagersand, M.: Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach, in Proc. of 5th International Conference on Computer Vision (1995)
3. Iyengar, J., Zachary, S.S., Barhen J.: Content based image retrieval and information theory: A generalized approach, in Special Topic Issue on Visual Based Retrieval Systems and Web Mining, Journal of the American Society for Information Science and Technology, pp. 841–853 (2001)
4. Tollari, S., Glotin, H.: Web image retrieval on imageval: Evidences on visualness and textualness concept dependency in fusion model, in ACM Int Conf on Image Video Retrieval (2007)
5. S. Tollari, H. Glotin, "Learning optimal visual features from web sampling in online image retrieval", in: ICASSP, IEEE, vol. 4p, Las Vegas, march 2008
6. Glotin, H., Zhao, Z.Q., Ayache, S.: Efficient Image Concept Indexing by Harmonic & Arithmetic Profiles Entropy, 2009 IEEE International Conference on Image Processing, Cairo, Egypt, November 7-11, 2009 (2009)
7. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions, Signal Processing, 16(3), 233–246 (1989)
8. Vapnik, V.: The nature of statistical learning theory. Springer-Verlag, New York (1995)
9. Vapnik, V.: Statistical learning theory. John Wiley, New York (1998)
10. 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)
11. Tong S., Edward, Chang: Support vector machine active learning for image retrieval. In Proceedings of the ninth ACM international conference on Multimedia Ottawa, Canada, pp. 107–118 (2001)
12. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers Neural Processing Letters, 9, 293–300 (1999)
13. Frey, B. J., Dueck D., Clustering by Passing Messages Between Data Points. Science, 315, 972–976 (2007)
14. Ayache S., Quenot G.: Video Corpus Annotation Using Active Learning, 30h European Conference on Information Retrieval (ECIR'08), pp 187–198 (2008)
15. Quenot, G. M., Moraru, D., Besacier, L.: CLIPS at TRECVID: Shot boundary detection and feature detection, in 'Proceedings of the TRECVID 2003 Workshop', Gaithersburg, Maryland, USA, pp. 35–40 (2003)
16. Smach, F., Lemaitre, C., Gauthier, J.P., Miteran, J., Atri, M.: Generalized Fourier Descriptors with Applications to Objects Recognition in SVM Context, 30, J. Math Imaging Vis 43–71 (2008)
17. http://mrim.imag.fr/irim/
18. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces versus fisher faces. IEEE Trans. Pattern Anal. Machine Intell. 19, 711–720 (1997).
19. Georges Qunot, etal., IRIM keynotes of TRECVID 2009: High Level Feature Extraction, in this NIST TREC2009 proceedings (2009).