LSIS TREC VIDEO 2008 High Level Feature Shot Segmentation using Compact Profil Entropy and Affinity Propagation Clustering

Herve GLOTIN and Zhongqiu ZHAO

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

Abstract. In this task, we build fast video indexing systems using a kind of efficient features based on the entropy of pixel projections. These features of 45 dimensions, called Profil Entropy Features (PEF), are derived using the projection in the horizontal orientation. These features are then fed to SVMs to produce the keyframe ranks, from which we can get the shot ranks. In the runs, we divided the training set into several subsets using randomly method or affinity propagation clustering in order to simplify learning, and then we combined the outputs of the SVMs on the subsets into the final output. Finally we also made some fusions of different runs using arithmetic and harmonic means. We got the inferred MAP of 0.05245 and the 19th rank among all the best by team of 37 automatic submitted runs, the average of which is 0.063 for a STD of 0.0458. Further, our system needs only 11 hours time consumption for training and testing on the whole TREC video sets (on a Linux Xeon 2.66GHZ).

Key words: TRECVID08, High Level Feature Extraction, video retrieval, entropy, affinity propagation

1 TRECVid 2008 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] 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.

2 The LSIS new Profil Entropy Features (PEF)

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], where 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. This feature, associated to mean and color STD, reached the second best rank in the official ImagEval 2006 campaign (see www.imageval.org and [5]).

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 1, or the harmonic mean of the pixels on each column or each row of I (noted 'Ha', then we have Π_X^{Ha}).

Then, we estimate the probability distribution function (pdf) of each profil according to [6,15]. Considering that the sources are ergodic, we finally calculate each PEF equal to the normalized entropy ($H(pdf)/log(\#bins(pdf))$). We detail below each step of the PEF extraction.

Let op be the selected projection,
for each color of I of $L(I)$ rows 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 finally complete the PEF features by the usual mean and standard deviation of each normalized color of I.

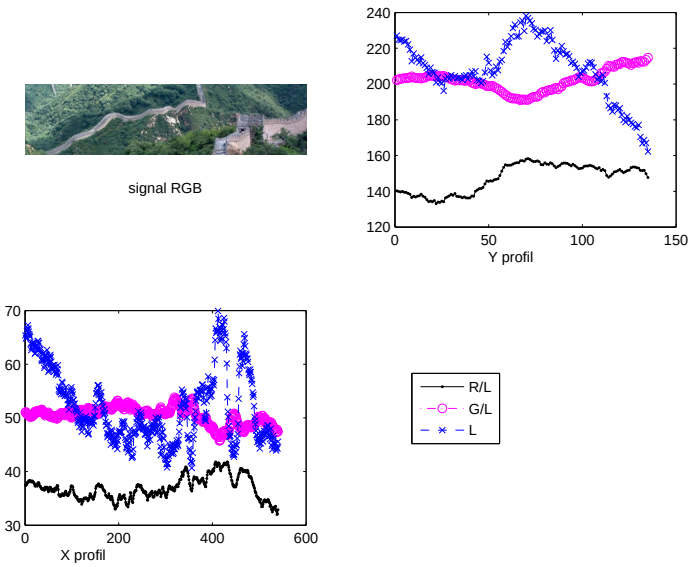

Fig. 1. Illustration of the horizontal and vertical profils using simple arithmetic projection (or sum) of each normalized color $r = R/L, g = G/L, L = R + G + B$.

Then we can calculate the PEF on three horizontal subimages as illustrated in Figure 2. We note such PEF using '='. For each image, 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 '=' features. It is interesting to note that only 1/20 second to compute the feature of one image on an usual linux Xeon.

3 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 [7,8] 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 [9,10]. 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

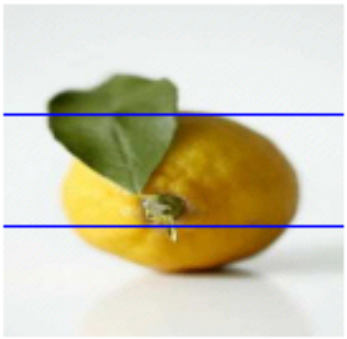

Fig. 2. Illustration of the 3 subimages of type '=' (horizontal).

cost function, a quadratic cost function can be used. The least squares support vector machines (LS-SVM) [11] 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.

4 Overview of the Video Retrieval System

The process we adopted in this task is overall shown in Figure 3. We used the labeled 2007 keyframes for the training. Then we split this labeled dataset into several subsets using random method or affinity propagation clustering [12]. Finally we got a set of SVMs on learning these subsets. We used these SVMs to evaluate the 2008 keyframes and combined the outputs of the SVMs using

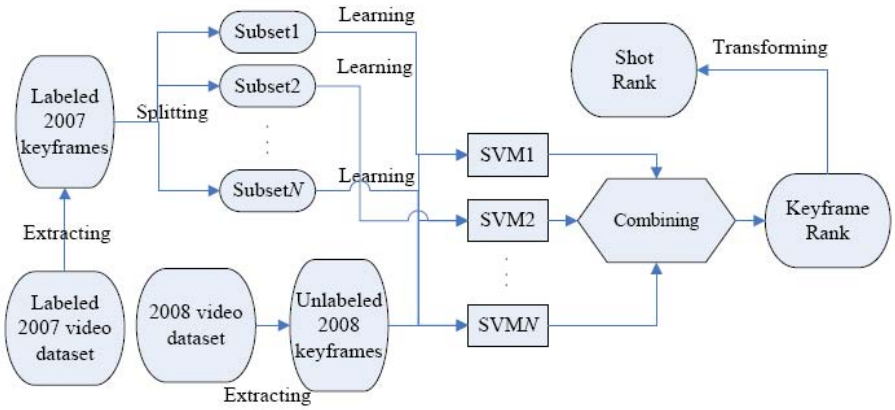

Fig. 3. The framework for TrecVid High Level Feature Extraction System. The SVM combination methods vary in the runs.

max or mean function to attain the keyframe ranks which was then transformed into the shot rank. The transformation is according to a max function of the keyframe scores of the shot, as defined in our IRIM group [16].

5 Submitted Runs

Shown in Table 1 and Figure 4 is the submitted 6 runs and their performances. The detail descriptions of these runs are as follows:

A_LSIS-1_1: Harmonic mean fusion of A_LSIS-5_5, A_LSIS-6_6, and other two runs (not submitted) using the similar method to that of A_LSIS-5_5, A_LSIS-6_6 but using max function for SVM combination instead of mean function.

A_LSIS-2_2: Harmonic mean fusion of ten runs: A_LSIS-5_5, A_LSIS-6_6, and the other 8 runs (not submitted) using the similar method but using randomly division, or affinity propagations with different parameters, or different combining methods for SVMs.

A_LSIS-3_3: Arithmetic mean fusion of A_LSIS-5_5, A_LSIS-6_6, and other two runs (not submitted) using the similar method to that of A_LSIS-5_5, A_LSIS-6_6 but using max function for SVM combination instead of mean function.

A_LSIS-4_4: Arithmetic mean fusion of ten runs: A_LSIS-5_5, A_LSIS-6_6, and the other 8 runs (not submitted) using the similar method but using randomly division, or affinity propagations with different parameters, or different SVM combination methods such as max and mean functions.

A_LSIS-5_5: Profil Entropy Features, divisions of training set using affinity propagation (the parameter in affinity propagation is set as 0), multiple SVMs on different subsets, mean function for SVM combination.

A_LSIS-6_6: Profil Entropy Features, divisions of training set using affinity propagation (the parameter in affinity propagation is set as 0.5), multiple SVMs on different subsets, mean function for SVM combination.

Table 1. The submitted 6 runs (on linux 4Gram, Intel®Xeon®CPU E5430 2.66GHZ)

Run Tag	Inferred MAP	Annotation Resources	Keyframe Selection Method	Low Level Features	Combination Methods	Fusion Method	Time for PEF extraction and training perHLF	Time for PEF extraction and test perHLF
A_LSIS-1_1	0.0334	LIG LSCOM [13]	Quenot [14]	PEF	max, mean	Harmonic mean	36m+4.5h	35m+25m
A_LSIS-2_2	0.0211	LIG LSCOM [13]	Quenot [14]	PEF	max, mean	Harmonic mean	36m+9h	35m+58m
A_LSIS-3_3	0.0479	LIG LSCOM [13]	Quenot [14]	PEF	max, mean	Arithmetic mean	36m+4.5h	35m+25m
A_LSIS-4_4	0.0525	LIG LSCOM [13]	Quenot [14]	PEF	max, mean	Arithmetic mean	36m+9h	35m+58m
A_LSIS-5_5	0.0504	LIG LSCOM [13]	Quenot [14]	PEF	mean	None	36m+1h	35m+4m45s
A_LSIS-6_6	0.0514	LIG LSCOM [13]	Quenot [14]	PEF	mean	None	36m+1.3h	35m+10m31s

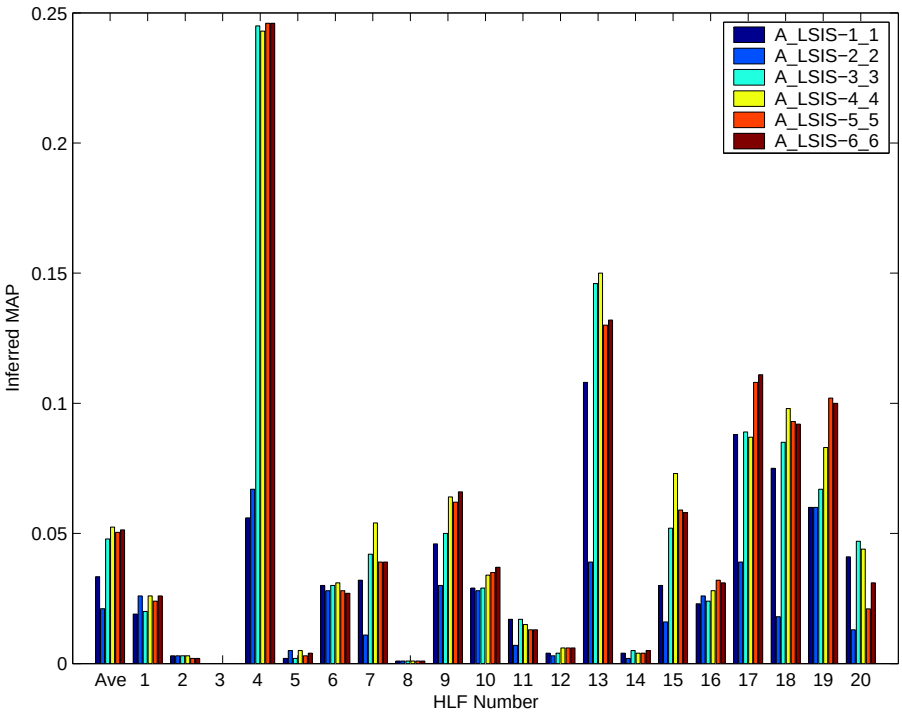

Fig. 4. The Inferred MAP performances of each run for each HLF

From Figure 4, we can see that the fusion does not always do well for all topics. The arithmetic mean fusion always does better than the harmonic mean

Task / Rank	TrecVid2008 HLF
1	dog
2	street
3	boat-ship
4	nighttime
5	flower
6	hands
7	driver
8	two people
9	singing
10	cityscape
11	airplane-flying
12	mountain
13	classroom
14	harbor
15	telephone
16	kitchen
17	demonstration-or-protest
18	bridge
19	bus
20	emergency-vehicle

Fig. 5. The descending rank of the topics by the inferred MAP of A_LSIS-4_4

fusion, and the former does better on the concepts of 'Two people', 'Street', 'Hand', 'Boat-ship', and 'Singing' than other concepts. The fusion does badly on 'Nighttime' and 'Flower', which may be caused by have not using the weighted fusion. And it needs our further research to optimize the arithmetic weights.

We also rank the topics by the descending MAP scores of the A_LSIS-4_4 run, as shown in Figure 5. The proposed system perform the best on the 'dog' concept (MAP = 0.243). This result is close to the best of all 200 runs (MAP = 0.271). It do badly on the concepts of 'bridge', 'bus', and 'emergency-vehicle'. The reason is that our PEF features are essentially mixture of color and texture of images. And these features are confusing for some similar concepts, for example, between 'bus' and 'emergency-vehicle'.

6 Conclusions

Considering the results of inferred MAP, we can see that A_LSIS-4_4 does the best but improves little. Further, from the results we can also concluded that:
1) the max function does worse than mean function for combination of SVMs, since we have some other runs not submitted, using max function, whose performances are worse than A_LSIS-5_5 and A_LSIS-6_6;
2) divisions using affinity propagation does much better than random divisions, since we have some other runs not submitted, using max function, whose performances are worse than A_LSIS-5_5 and A_LSIS-6_6;
3) arithmetic weighted mean fusion does better than Harmonic mean fusion, since A_LSIS-3_3 does better than A_LSIS-1_1 and A_LSIS-4_4 does better than A_LSIS-2_2.

In further work, we will extend the PEF using also vertical subimages, or using harmonic pixel projection as complement to the simple mean.

Also we will optimize the weights of the arithmetic mean fusion, which is expected to make the fusion more proper for some concepts.

Acknowledgment

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

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. Glotin, H.: Information retrieval and robust perception for a scaled multi-structuration, Thesis for habilitation of research direction, University Sud Toulon-Var, Toulon (2007)
5. 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)
6. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions, Signal Processing, 16(3), 233–246 (1989)
7. Vapnik, V.: The nature of statistical learning theory. Springer-Verlag, New York (1995)

8. Vapnik, V.: Statistical learning theory. John Wiley, New York (1998)
9. 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)
10. 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)
11. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers Neural Processing Letters, 9, 293–300 (1999)
12. Frey, B. J., Dueck D., Clustering by Passing Messages Between Data Points. Science, 315, 972–976 (2007)
13. Ayache S., Quenot G.: Video Corpus Annotation Using Active Learning, 30h European Conference on Information Retrieval (ECIR'08), pp 187–198 (2008)
14. 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)
15. Glotin, H., Zhao, Z., Profil Entropic visual Features for Visual Concept Detection in CLEF 2008 campaign, In Working Notes of ImageCLEF2008, Danmark, in conjuction with ECDL 2008.
16. http://mrim.imag.fr/irim/