Perplexity-based Evidential Neural Network Classifier Fusion using MPEG-7 Low-level Visual Features

Rachid Benmokhtar
Institut Eurécom - Département Multimédia
2229, route des crêtes
06904 Sophia Antipolis - France
rachid.benmokhtar@eurecom.fr

Benoit Huet
Institut Eurécom - Département Multimédia
2229, route des crêtes
06904 Sophia Antipolis - France
benoit.huet@eurecom.fr

ABSTRACT

In this paper, an automatic content-based video shot indexing framework is proposed employing five types of MPEG-7 low-level visual features (color, texture, shape, motion and face). Once the set of features representing the video content is determined, the question of how to combine their individual classifier outputs according to each feature to form a final semantic decision of the shot must be addressed, in the goal of bridging the semantic gap between the low level visual feature and the high level semantic concepts. For this aim, a novel approach called "perplexity-based weighted descriptors" is proposed before applying our evidential combiner NNET [3], to obtain an adaptive classifier fusion PENN (Perplexity-based Evidential Neural Network). The experimental results conducted in the framework of the TRECVid'07 high level features extraction task report the efficiency and the improvement provided by the proposed scheme.

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content analysis and indexing—*Indexing methods*; I.5.2 [**Pattern recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Algorithms, Experimentation, Performance.

Keywords

Video semantic analysis, perplexity, entropy, visual descriptors, classifier fusion, neural network, evidence theory.

1. INTRODUCTION

With explosive spread of image and video data, video retrieval based on visual content is one of the challenging topic in the multimedia research, in particular to bridge the semantic gap between the low-level features and the high-level semantic concepts. Bridging the semantic gap via video classification requires to finely analyze the video shot content and to extract a set of features describing the content. The combination of these features toward an effective classification is however far from being trivial. Here, we focus in the case where the combination of cues from the various feature is realized post classification.

In this paper, we present our research conducted toward a semantic video content indexing and retrieval system. The general architecture of our system is depicted in Figure 1. The overall chain can be divided into 3 parts: (1) Feature extraction, (2) classification and (3) classifier fusion. The feature extraction step consists in extracting a set of low level features based on color, texture, shape, motion and face. Then, SVM classification is used to label the video shots. Finally, fusion of classifier outputs is performed thanks to a neural network based on evidence theory (NNET) [3]. The main objective is to show the importance and the role of fusion. Here, we propose a novel approach of weighting descriptors based on the entropy and perplexity measures to combine the individual classifier outputs according to each descriptor.

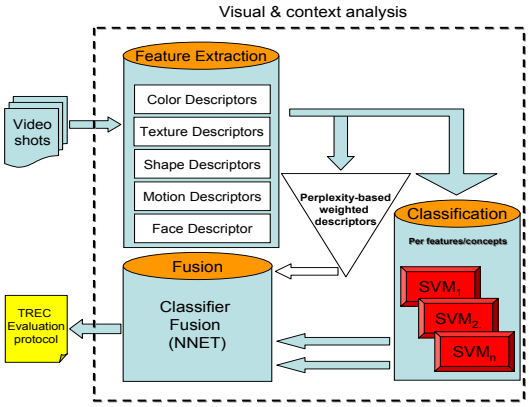

Figure 1: General indexing system architecture.

The rest of this paper is organized as follows. Section 2 presents the set of MPEG-7 visual descriptors employed by our system. Section 3 gives the proposed concept modeling, including the perplexity-based approach to weight the classifier outputs. Section 4 evaluates the experimentation results conducted on the TRECVid 2007 collection. Section 5 provides the conclusion of the paper.

MIR'08, October 30–31, 2008, Vancouver, Canada.

2. VISUAL DESCRIPTORS

The MPEG-7 standard defines a comprehensive, standardized set of audiovisual description tools for still images as well as movies. The aim of the standard is to facilitate quality access to content, which implies efficient storage, identification, filtering, searching and retrieval of media [10]. Our system employs five types of MPEG-7 visual descriptors: Color, texture, shape, motion and face descriptors. These descriptors are defined as follows [14]:

• **Scalable Color Descriptor (SCD)** is defined as the hue-saturation-value (HSV) color space with fixed color space quantization. The Haar transform encoding is used to reduce the number of bins of the original histogram with 256 bins to 16, 32, 64, or 128 bins [6].

• **Color Layout Descriptor (CLD)** is a compact representation of the spatial distribution of colors [7]. The color information of an image is divided into (8x8) block. The blocks are transformed into a series of coefficient values using dominant color descriptor or average color, to obtain $CLD = \{Y, Cr, Cb\}$ components. Then, the three components are transformed by 8x8 DCT (Discrete Cosine Transform) to three sets of DCT coefficients. Finally, a few low frequency coefficients are extracted using zigzag scanning and quantized to form the CLD for a still image.

• **Color Structure Descriptor (CSD)** encodes local color structure in an image using a structuring element of (8x8) dimension. CSD is computed by visiting all locations in the image, and then summarizing the frequency of color occurrences in each structuring element location on four HMMD color space quantization possibilities: 256, 128, 64 and 32 bins histogram [11].

• **Color Moment Descriptor (CMD)** provides some information about color in a way which is not explicitly available in other color descriptors. It is obtained by the mean and the variance on each layer of the LUV color space of an image or region.

• **Edge Histogram Descriptor (EHD)** expresses only local edge distribution in the image. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. The EHD basically represents the distribution of 5 types of edges in each local area called a sub-image. Specifically, dividing the image into (4x4) non-overlapping sub-images. Then, for each sub-image, we generate an edge histogram. Four directional edges (0°, 45°, 90°, 135°) are detected in addition to non-directional ones. Finally, it generates a 80 dimensional vector (16 sub-images, 5 types of edges). We make use of the improvement proposed by [13] for this descriptor, which consist in adding global and semi-global levels of localization of an image.

• **Homogeneous Texture Descriptor (HTD)** characterizes a region's texture using local spatial frequency statistics. HTD is extracted by Gabor filter banks (6 frequency times, 5 orientation channels), resulting in 30 channels in total. Then, computing the energy and energy deviation for each channel to obtain 62 dimensional vector [10, 18].

• **Statistical Texture Descriptor (STD)** is based on statistical methods of co-occurrence matrix such as: energy, maximum probability, contrast, entropy, etc [1], to model the relationships between pixels within a region of some grey-level configuration in the texture; this configuration varies rapidly with distance in fine textures, slowly in coarse textures.

• **Contour-based Shape Descriptor (C-SD)** presents a closed 2D object or region contour in an image. To create CSS description of contour shape, N equidistant points are selected on the contour, starting from an arbitrary point on the contour and following the contour clockwise. The contour is then gradually smoothed by repetitive low-pass filtering of the x and y coordinates of the selected contour points, until the contour becomes convex (no curvature zero-crossing points are found). The concave part of the contour is gradually flattered out as a result of smoothing. Points separating concave and convex parts of the contour and peaks (maxima of the CSS contour map) in between are then identified. Finally, eccentricity, circularity and number of CSS peaks of original and filtered contour are should be combined to form more practical descriptor [10].

• **Camera Motion Descriptor (CM)** details what kind of global motion parameters are present at what instance in time in a scene provided directly by the camera, supporting 7 camera operations: fixed, panning (horizontal rotation), tracking (horizontal transverse movement), tilting (vertical rotation), booming (vertical transverse movement), zooming (change of the focal length), dollying (translation along the optical axis), and rolling (rotation around the optical axis) [10].

• **Motion Activity Descriptor (MAD)** shows whether a scene is likely to be perceived by a viewer as being slow, fast paced, or action paced [15]. Our MAD is based on intensity of motion. The standard deviations are quantized into five activity values. A high value indicates high activity and the low value of intensity indicates low activity.

• **Face Descriptor (FD)** detects and localizes frontal faces within the keyframes of a shot and provides some face statistics (e.g, number of faces, biggest face size), using the face detection method implemented in OpenCV.

3. CONCEPT MODELING

Once the visual descriptors are extracted from the video image, the task of semantic concept modeling can be summarized as three steps: (1) classification, (2) perplexity-based weighted descriptors and (3) classifier fusion.

3.1 SVM-based Classification

SVMs have become widely used in the classification task due to their generalization ability in the high-dimensionality pattern recognitions [17]. The main idea is similar to the concept of a neuron: Separate classes with a hyperplane. However, samples are indirectly mapped into a high dimensional space thanks to a kernel function. In our paper, we use one SVM for each low-level feature, trained per concept under the *"one against all"* approach. We adopt a sigmoid function to compute the degree of confidence y_i^j (Eq. 1).

$$y_i^j = \frac{1}{1 + \exp(-\alpha d_i)} \quad (1)$$

where (i, j) represents the i^{th} concept and j^{th} low-level feature. d_i is the distance between the input vector and the hyperplane. α is the slope parameter obtained experimentally.

3.2 Perplexity-based Weighted Descriptors

Each LSCOM-lite (Large-Scale Concept Ontology for Multimedia) [12] semantic concept is best represented or described by its own set of descriptors. Intuitively, the *color descriptors* could be better to detect certain concepts such as "sky, snow, waterscape, and vegetation", and lower for "studio, meeting" for example.

For this aim, we propose to weight each low-level feature per concept, without any feature selection (fig. 2). The variance as a simple second order vector can be used to give the knowledge of the dispersion around the mean between descriptors and concepts. Conversely, the entropy depends on more parameters and measures the quantity of informations and uncertainty in a probabilistic distribution. We propose to maps the visual features onto a term weight vector via entropy and perplexity measures. This vector is then combined with the original classifier outputs [1] to produce the final classifier outputs. As presented in figure 2, we define now the four steps of the proposed approach.

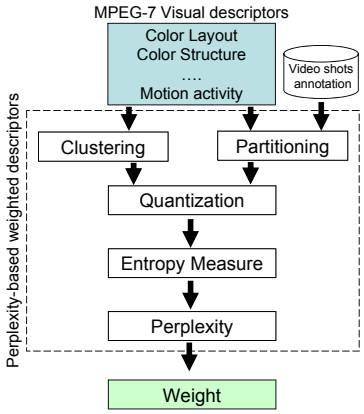

Figure 2: Perplexity-based weighted descriptors structure.

1. **K-means Clustering** computes the k center of cluster for each descriptor, in order to create a "visual dictionary" of the shots (the experimentations show that $k = 2000$ presents a compromise between efficiency and a low time-consuming computation).

2. **Partitioning** selects the positifs samples for each concept.

3. **Quantization** computes Euclidean distance between using each partitioning data set and dictionary.

4. **Entropy measure:** The entropy H (Eq. 2) of a certain feature vector distribution $P = (P_0, P_1, ..., P_{k-1})$ gives a measure of concepts distribution uniformity over the clusters k [9]. In [8], a good model is such that the distribution is heavily concentrated on only few clusters, resulting in low entropy value.

$$H = -\sum_{i=0}^{k-1} P_i \log(P_i) \quad (2)$$

where P_i is the probability of cluster i on the quantized vector.

5. **Perplexity measure:** In [5], perplexity PPL or normalized perplexity value $\overline{PPL}$ (Eq. 3) can be interpreted as the average number of clusters needed for an optimal coding of the data.

$$\overline{PPL} = \frac{PPL}{PPL_{max}} = \frac{2^H}{2^{H_{max}}} \quad (3)$$

If we assume that k clusters are equally probable, we obtain $H(P) = \log(k)$, and then $1 \leq PPL \leq k$.

6. **Weight:** In speech recognition, handwriting recognition, and spelling correction [5], it is generally assumed that lower perplexity/entropy correlates with better performance, or in our case, to a very concentrated distribution. So the relative weight of the corresponding feature should be increased. Many formula can be used to represent the weight such as Sigmoid, Softmax, Gaussian, etc. In our paper, we choose Verhulst evolution model (Eq. 4). This function is non exponential, it allows brake rate α_i, reception capacity (upper asymptote) K, and β_i defines the decreasing speed of weight function.

$$w_i = K \frac{1}{1 + \beta_i \exp(-\alpha_i(1/\overline{\mathrm{PPL}_i}))} \quad (4)$$

$$\beta_i = \begin{cases} K \exp(-\alpha_i^2) & \text{if } Nb_i^+ < 2 * k \\ 1 & \text{Otherwise} \end{cases} \quad (5)$$

β_i is introduced to decrease the negative effect of the training set limitation, due to the low number of positifs samples ($Nb_i^+ << k$) of certain concepts such as *weather,desert,mountain,*etc (see table 1). We observe a lower perplexity value, which could not be interpreted as a relevant relation between descriptor and concept. So, we increase β_i (Eq. 5) to obtain a rapid weight decrease for each concept presenting less than $2 * k$ positifs samples.

The relevance of the various descriptors at identifying high level concepts can be obtained through the perplexity distribution. So, the Boxplot provides an excellent visual summary of many important aspects of a distribution. The lower and upper lines express the data range, the lower and upper edges of the box indicate the 25^{th} and 75^{th} percentile. The line inside the box indicates the median value of the data. Figure 3 shows the normalized perplexity for each descriptor and its best concept presented by the minimum observation, such as: SCD is more effective to detect the concept

[1] We can also use the weight in the feature extraction step.

sky "13", EDH for *road* "12",etc. The first observation concerns the same value of median perplexity obtained for SCD, CLD, CMD, CSD, where color is more discriminant. Second, C-SD gives the smallest 25^{th} percentile of normalized perplexity for all data, followed by EDH and SCD. Third, it seems that EHD is very useful in the detection of the contour as in the *sport* and *road* concepts. Identical observation is given for C-SD. Conversely, MAD presents a large interval of perplexity but gives small value for the concepts *walking-running, people-marching* where the motion activity can be detected. Finally, FD is a relevant descriptor to detect *face* and *person* concepts which is not surprising.

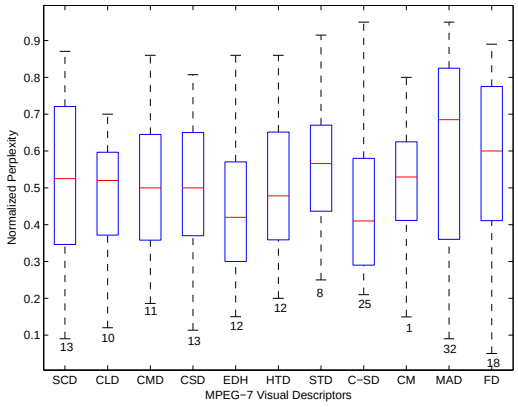

Figure 3: Normalized Perplexity Boxplot.

3.3 Classifier Fusion

In this part, we briefly describe our recently proposed neural network based on evidence theory (NNET) to address classifier fusion [2] (Figure 4) [3].

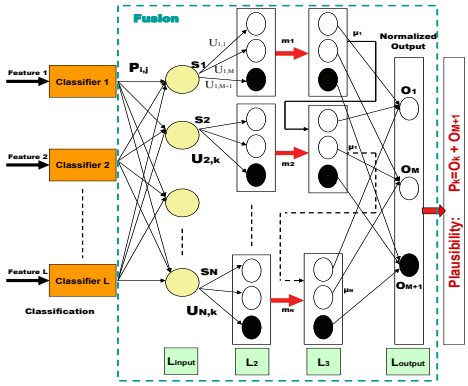

Figure 4: NNET classifier fusion structure.

1. **Layer** L_{input}**:** Contains N units. Identical to the RBF (Radial Basis Function) network input layer with an exponential activation function ϕ. d: distance computed using training data. $\alpha \in [0,1]$ is a weakening parameter associated to unit i.

$$\begin{cases} s^i = \alpha^i \phi(d^i) \\ \phi(d^i) = \exp\left(-\gamma^i (d^i)^2\right) \end{cases} \quad (6)$$

[2]The state of the art and the comparison study about the effectiveness of the classifier fusion methods are given in [2].

2. **Layer** L_2**:** Computes the belief masses m^i (Equ. 7) associated to each unit. The units of module i are connected to neuron i of the previous layer.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u_q^i \phi(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi(d^i) \end{cases} \quad (7)$$

where u_q^i is the membership degree to each class w_q, q class index $q = \{1, ..., M\}$.

3. **Layer** L_3**:** The Dempster-Shafer combination rule combines N different mass functions in one single mass. It is given by the conjunctive combination (Eq. 8):

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{B_1 \bigcap ... \bigcap B_N = A} \prod_{i=1}^{N} m^i(B_i) \quad (8)$$

The activation vector of modules i is defined as $\vec{\mu^i}$. The activation vectors can be recursively computed using:

$$\begin{cases} \mu^1 = m^1 \\ \mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{M+1}^i + \mu_{M+1}^{i-1} m_j^i \\ \mu_{M+1}^i = \mu_{M+1}^{i-1} m_{M+1}^i \end{cases} \quad (9)$$

4. **Layer** L_{output}**:** In [4], the output is directly obtained by $O_j = \mu_j^N$. The experiments show that this output is very sensitive to the number of prototype, where a small modification in the number can change the classifier fusion behavior. To resolve this problem, we use normalized output (Eq. 10). Here, the output is computed taking into account the activation vectors of all prototypes to decrease the effect of an eventual bad behavior of prototype in the mass computation.

$$O_j = \frac{\sum_{i=1}^{N} \mu_j^i}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu_j^i} \quad (10)$$

$$P_q = O_q + O_{M+1} \quad (11)$$

The different parameters (Δu, $\Delta\gamma$, $\Delta\alpha$, ΔP, Δs) can be determined by gradient descent of output error for an input pattern x. Finally, the maximum of plausibility P_q of each class w_q is computed.

4. EXPERIMENTAL RESULTS

The high level concept detection experiments presented in this paper are conducted on the TRECVid 2007 dataset [16] containing science news, news reports, documentaries, educational programs, and archival video. Of the 100 hours of video segmented into shots annotated with concepts from the 36 labels (Table 1), half is used to train the feature extraction system and the other half is used for evaluation purpose. The training set is divided into two subsets in order to train classifiers and subsequently the fusion parameters. For evaluation, we use the common measure from the information retrieval community: "Average precision, classification and the error rates".

Figure 5 presents the attributed normalized weight *vs* visual descriptors. We notice that each descriptor's importance varies depending on the concept under investigation. It has more or less importance in the concept detection. Color and texture descriptors have more weight value for "*vegetation,*

Id	Concepts	Neg.Train	Pos.Train	Pos.Test
1	Sports	11974	106	42
2	Weather	12029	51	34
3	Court	11967	113	5
4	Office	11159	921	453
5	Meeting	11532	548	270
6	Studio	11722	358	468
7	Outdoor	8643	3437	1812
8	Building	10964	1116	477
9	Desert	12019	61	15
10	Vegetation	10615	1465	499
11	Mountain	12004	76	17
12	Road	11420	660	297
13	Sky	10777	1303	853
14	Snow	12044	36	91
15	Urban	10746	1334	537
16	Waterscape	11725	355	414
17	Crowd	11159	921	552
18	Face	6596	5484	2325
19	Person	4981	7099	2972
20	Police Security	11824	256	63
21	Military	11848	232	74
22	Prisoner	12067	13	7
23	Animal	11675	405	271
24	Computer Tv	11617	463	202
25	US Flag	12070	10	0
26	Airplane	12052	28	7
27	Car	11663	417	187
28	Bus	12033	47	40
29	Truck	11985	95	19
30	Boat Ship	11979	101	151
31	Walking Running	11221	859	385
32	People Marching	11960	120	82
33	Explosion Fire	11068	12	19
34	Natural Disaster	12061	19	21
35	Maps	12030	50	31
36	Charts	11954	126	80

Table 1: Id of the TRECVid Concepts.

building, sky, etc". CM and MAD are more sensitive for the concepts "*walking-running* and *people-marching*". FD presents an important weight, essentially for the video shots presenting "*human-body* and *face*".

Now, we would like to study a number of ways in which the results previously detailed can be used to improve the retrieval's system performance. Figure 6 compares the performance of a simple system "No-weight" where all descriptors are taken as equal in terms if relevance to all semantic concepts, with four evolution models of weights (Softmax, Sigmoid, Gaussian, and Verhulst). Our proposed model based on Verhulst has the best average precision for all semantic concepts, in particular we observe a significant improvement for the concepts "4,5,6,16,17,18,19,23,31, and 32".

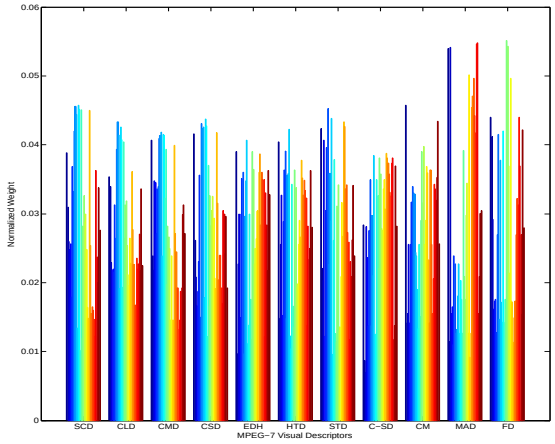

Figure 5: Normalized Perplexity-based weight descriptors.

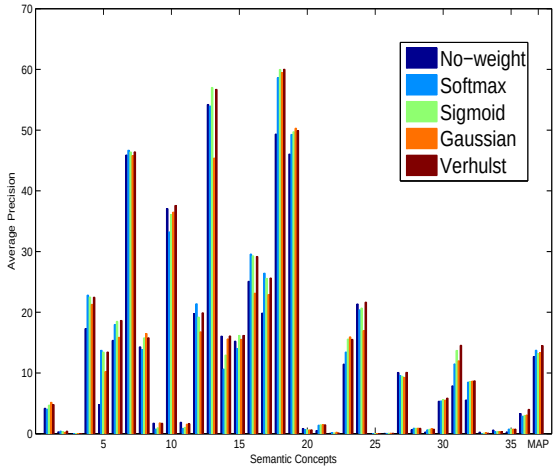

Figure 6: Performance comparison of the 5 approaches across the 36 benchmark concepts on the fusion set. PENN based on Verhulst model outperform all other approaches by a wide weighted combination.

As an example, to detect *face, person, meeting, or studio* concepts, PENN gives more importance to *FaceDetector, ContourShape, ColorLayout, ScalableColor, EdgeHistogram* than others descriptors. For the "Person" concept, the improvement was high as 11%, making it the best performing run.

Other models achieve respectable results in average with some decrease due to both numerous conflicting classification and limited training data. This also explains the extreme cases obtained for concepts 3,22,25,26 and 33.

In order to measure the overall performance for the content-based video shots classification, we calculate the Mean Average Precision (MAP), F-measure (F-meas), Positive Classification Rate (CR^+), and Balanced Error Rate (BER) [3] of all concepts, and for a subset of the 10 most frequent concepts in the dataset (Table 3). PENN "Verhulst" allows an overall improvement of the system and a significant increase of MAP, F-meas, CR^+, and decreases the global error "BER" comparing to the NNET "No-weight".

Table 2: Performances comparison.

Methods / Evaluation	NNET"No-weight" (%)	PENN"Verhulst" (%)
MAP	12.69	13.29
MAP@10	33.70	35.30
F-meas	11.84	14.10
F-meas@10	38.75	40.79
CR^+	11.93	13.43
CR^+@10	40.69	41.74
BER	45.02	44.13
BER@10	38.00	36.52

5. CONCLUSIONS

To bridge the semantic gap, an ideal video retrieval system should analyze finely the relationship between descriptors and concepts. To that end, we have developed a novel approach of descriptors weighting based on the entropy and perplexity measures using Verhulst model. The experiments show that our system is more effective to bridge this gap, and outperform all other approaches by a wide weighted combination.

The future works will concern the study of the ontology and the inter-concepts similarities between the classes, and the exploitation of this semantic informations on our classification or fusion system.

6. ACKNOWLEDGMENTS

The work presented here is supported by the European Commission under contract FP6-027026-K-SPACE. This work is the view of the authors but not necessarily the view of the community.

7. REFERENCES

[1] T. Adamek. Extension of mpeg-7 low-level visual descriptors for TRECVid07. Kspace Technical Report, FP6-027026, 2007.

[2] R. Benmokhtar and B. Huet. Classifier fusion: Combination methods for semantic indexing in video content. In *Proceedings of ICANN*, pages 65–74, 2006.

[3] R. Benmokhtar and B. Huet. Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In *Proceedings of MMM*, volume 4351, pages 196–205, 2007.

[4] T. Denoeux. An evidence-theoretic neural network classifer. In *International Conference on Systems, Man and Cybernetics*, volume 31, pages 712–717, 1995.

[3]BER is the average of the errors on each class.

[5] J. Gao, J. Goodman, M. Li, and K. Lee. Toward a unified approach to statistical language modeling for chinese. In *ACM Transactions on Asian Language Information Processing*, 2001.

[6] ISO/IEC/JTC1/SC29/WG11/N4062. Coding of moving pictures and associated audio. Information Technology, October 2001.

[7] E. Kasutani and A. Yamada. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/ video retrieval. In *Proceedings of ICIP*, volume 1, pages 674–677, 2001.

[8] M. Koskela and A. Smeaton. Clustering-based analysis of semantic concept models for video shots. In *Proceedings of ICME*, pages 45–48, 2006.

[9] J. Laaksonen, M. Moskela, and E. Oja. Class distributions on som surfaces for feature extraction and object retrieval. *Neural Network*, 17:1121–1133, 2004.

[10] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia content description interface*. John Wiley and Sons, New York, 2002.

[11] D. Messing, P. V. Beek, and J. Errico. The MPEG-7 color structure descriptor: image description using color and local spatial information. In *Proceedings of ICIP*, volume 1, pages 670–673, 2001.

[12] M. Naphade, L. Kennedy, J. Kender, S. Chang, J. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for trecvid 2005. IBM Research Technical Report, 2005.

[13] D. Park, Y. S. Jeon, and C. S. Won. Efficient use of local edge histogram descriptor. In *Proceedings of ACM workshop on Multimedia*, pages 51–54, 2000.

[14] A. Smeaton, Z. Obrenovic, B. Huet, F. Vallet, I. Kompatsiaris, Y. Avrithis, W. Bailer, E. Izquierdo, T. Sikora, and P. Praks. K-Space at TRECVid 2007. In *Processing of TRECVid*, 2007.

[15] X. Sun, B. Manjunath, and A. Divakaran. Representation of motion activity in hierarchical levels for video indexing and filtering. In *Proceedings of ICIP*, pages 149–152, 2002.

[16] TRECVID. Digital video retrieval at NIST. http://www-nlpir.nist.gov/projects/trecvid/.

[17] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[18] F. Xu and Y. Zhang. Evaluation and comparison of texture descriptors proposed in MPEG-7. *Journal of Visual Communication and Image Representation*, 17:701–716, 2006.