Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 3–10
ISSN 1896-7094

Neural Network Combining Classifier Based on Dempster-Shafer Theory

Rachid Benmokhtar and Benoit Huet

Institut Eurécom - Département Multimédias
2229, route des crêtes
06904 Sophia-Antipolis - France
(Rachid.Benmokhtar, Benoit.Huet)@eurecom.fr

Abstract. In this paper, we propose an improved version of RBF network based on Evidence Theory (NN-ET) using one input layer and two hidden layers and one output layer, to improve classifier combination and recognition reliability in particular for automatic semantic-based video content indexing and retrieval. Many combination schemes have been proposed in the literature according to the type of information provided by each classifier as well as their training and adaptation abilities. Experiments are conducted in the framework of the TrecVid 2005 features extraction task that consists in ordering shots with respect to their relevance to a given class. Finally, we show the efficiency of NN-ET combination method.

1 Introduction

Multimedia digital documents are readily available, either through the Internet, private archives or digital video broadcast. Tools are required to efficiently index this huge amount of information and to allow effective retrieval operations. Unfortunately, most existing systems rely on the automatic description of the visual content through color, texture and shape features whereas users are more interested in the semantic multimedia content. To answer this need, the MPEG-7 [1] standard offers the possibility to describe the video content and in particular its semantic content. In practice an important gap remains between the visual descriptors and the semantic content. The most common solution to date is the use of manually or semi-automatically annotated content. However, in both cases the annotation task is very time consuming and error prone. New tools for automatic semantic video content indexing are highly awaited and an important effort is now conducted by the research community to automatically bridge the existing gap [2,3].

The retrieval of complex semantic concepts requires the analysis of many features per modalities. The task consisting of combining all these different parameters is far from trivial. The fusion mechanism can take place at different levels of the classification process. Generally, it is either applied on signatures (feature fusion) or on classifier outputs (classifier fusion). Unfortunately, complex signatures obtained from fusion of features are difficult to analyze and it results in classifiers that are not well trained despite of the recent advances in machine learning. Therefore, the fusion of classifier outputs remains an important step of the classification task.

In classifier fusion systems, information coming from the various classifiers are fused to obtain the final classification score. In this paper, RBF neural network and neural network based on evidence theory are implemented for this purpose and evaluated in the context of content-based retrieval.

This paper presents our research conducted toward a semantic video content indexing and retrieval system. It starts with the presentation of the video latent semantic analysis architecture. It is followed by a description of RBF neural network and how the evidence theory can be used in an effort to evaluate their classification and fusion ability. The experimental results presented in this paper are conducted in the framework of TrecVid'05. This study reports the efficiency of different combination methods and shows the improvement provided by our proposed scheme. Finally, we conclude with a summary of the most important results provided by this study along with some possible extension of work.

2 System Architecture

This section describes the workflow of the semantic feature extraction process that aims to detect the presence of semantic classes in video shots, such as building, car, U.S. flag, water, map, etc ...

First, key-frames of video shots, provided by TrecVid'05, are segmented into homogeneous regions thanks to the algorithm described in [4]. Secondly, color and texture are extracted for each region obtained from the segmentation. Thirdly, the obtained vectors over the complete database are clustered to find the N most representative elements. The clustering algorithm used in our experiments is the well-known k-means.

Representative elements are then used as visual keywords to describe video shot content. To do so, computed features on a single video shot are matched to their closest visual keyword with respect to the Euclidean distance (or other distance measures).

Then, the occurrence vector of the visual keywords in the shot is build and this vector is called the Image Vector Space Model (IVSM) signature of the shot. Image latent semantic analysis (ILSA) is applied on these features to obtain an efficient and compact representation of video shot content. Finally, support vector machines (SVM) are used to obtain the first level classification which output will then be used by the fusion mechanism [5]. The overall chain is presented in figure 1.

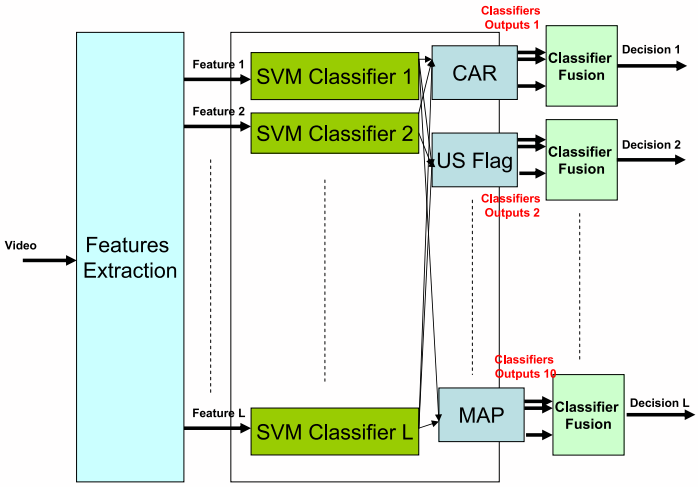

Fig. 1. General framework of the application.

2.1 Video features extraction

For the study presented in this paper we distinguish visual features. To describe the visual content of a shot, features are extracted from key-frames. Two visual features are selected for this purpose: Hue-Saturation-Value color histograms and energies of Gabor's filters [6]. In order to capture the local information in a way that reflects the human perception of the content, visual features are extracted on regions of segmented key-frames [7]. Then, to have reasonable computation complexity and storage requirements, region features are quantized and key-frames are represented by a count vector of quantization vectors. At this stage, we introduce latent semantic indexing to obtain an efficient region based signature of shots. Finally, we combine the signature of the key-frame with the signatures of two extra frames in the shot, as it is described in [5], to get a more robust signature.

2.2 Classification

The classification consists in assigning classes to video shots given some description of its content. It is an important step for video indexing systems since it allows completing the visual description of the content with class information. Unfortunately, many different cues are implied in the classification process. The visual content is extremely rich in semantic classes, but limited data is available to build classification models. We decided to conduct the classification on individual features in order to have enough training data with respect to input vector sizes.

Allwein and al [8] showed that it was possible to transform a multi-classes classification problem into several binary classification problems. They propose *one-against-all method*, which consists in building a system of binary classification by class. Every binary system classifies samples in a class or in the other (i.e. who understands all the remaining classes). In our work, this method is adopted using the SVM classification.

Support Vector Machines SVMs are one of the most popular machine learning techniques, since they have shown very good generalization performance on many pattern classification problems. They have the property to allow a non linear separation of classes with very good generalization capacities. They were first introduced by Vapnik [9] for the text recognition task.

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 that respects the Mercer's condition [10]. This allows leading the classification in a new space where samples are assumed to be linearly separable. The selected kernel denoted $\mathcal{K}(.)$ is a radial basis function which normalization parameter σ is chosen depending on the performance obtained on a validation set. The radial basis kernel is chosen for his good classification results comparing to Polynomial and Sigmoidal kernels [5].

3 Classifier Fusion

3.1 Radial Neural Network (RBF)

RBF is a popular supervised neural network learning algorithm, it's a spacialization of the MLP network [11]. The RBF network is constitued by only the following three layer, as shown in (figure 2).

- *Input Layer :* It broadcast the inputs without distortion;
- *RBF Layer :* Hidden layer that contain the RBF function;
- *Output Layer :* Simple layer that contain a lineaire function.

Basis functions normally take the form $\phi = \| \vec{x} - \vec{\mu_i} \|$. The function depends on the distance (usually taken to be Euclidean) between the input vector $\vec{x}$ and a vector $\vec{\mu_i}$. The most common form of basis function used is the Gaussian function $\phi = \exp \frac{\|\vec{x}-\vec{\mu_i}\|^2}{2\sigma_j^2}$.

where $\vec{\mu_i}$ determines the center of basis function and σ_i is a width parameter that controls how is spread the curve. Generally, these centers are selected by using some fuzzy or non-fuzzy clustering algorithms. In this work, we have used the k-means algorithm to select the initial cluster centers in the first stage and then these centers are further fine tuned by using point symmetry distance measure.

The number of neurons in the output layer is equal to the possible classes of the given problem. Each output layer neuron computes a linear weighted sum of the outputs of the hidden layer neurons as follows:

$$y_i(x) = \sum_{i=1}^{N} \phi_i(x) W_i \tag{1}$$

The weight vectors are determined by minimizing the mean squared differences between the classifier outputs $y_k = \sum_{j=0}^{M} w_{k,j} s_i$ and target values t_k as following :

$$E = \frac{1}{2} \sum_{k=1}^{M} (y_k - t_k)^2 \tag{2}$$

The parameter (ΔW, $\Delta\mu$, $\Delta\sigma$) are given by (for more explication, see [11]) :

$$\frac{\partial E}{\partial w_{k,i}} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial w_{k,i}} \tag{3}$$

or $\frac{\partial E}{\partial y_k} = -(t_k - y_k)$, thus,

$$\frac{\partial E}{\partial w_{k,i}} = -(t_k - y_k) s_i \tag{4}$$

after computation, we obtain :

$$\frac{\partial E}{\partial \mu_{j,i}} = \sum_k \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial s_j} \frac{\partial s_j}{\partial \mu_{j,i}} = \frac{s_j}{\sigma_j^2} (x_i - \mu_{j,i}) \sum_k (t_k - y_k) w_{k,j} \tag{5}$$

$$\frac{\partial E}{\partial \sigma_j} = \sum_k \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial s_j} \frac{\partial s_j}{\partial \sigma_{j,i}} = \frac{2 s_j}{\sigma_j} \log s_j \sum_k (t_k - y_k) w_{k,j} \tag{6}$$

3.2 Evidence Theory

As we have seen, solutions in combining multiple classifiers are numerous but each of them has weaknesses. Most treat imprecision, but uncertainty and reliability are ignored. Evidence theory allows to use uncertain data [12].

Let Ω be a finite set of mutually exclusive and exhausive hypotheses, called the *frame of dicernement*. A basic belief assignment (BBA) is a function m from 2^{Ω} to $[0, 1]$ verifying :

$$\begin{cases} m(\emptyset) = 0 \\ \sum_{A \subseteq \Omega} m(A) = 1 \end{cases} \tag{7}$$

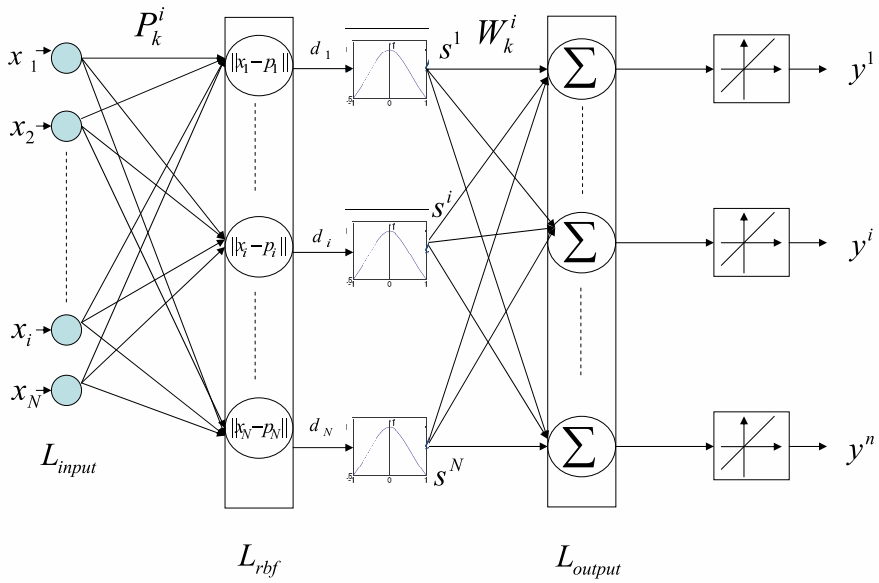

Fig. 2. RBF Classifier Structure

For any $A \subseteq \Omega$, $m(A)$ represents the belief that one is willing to commit exactly to A, given a certain piece of evidence. The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. Associated with m are a *belief* or *credibility* function *bel* and a *plausibility* function *pl*, defined, respectively, for all $A \in \Omega$ as :

$$bel(A) = \sum_{B \subseteq A} m(B) \tag{8}$$

$$pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{9}$$

The quantity $bel(A)$ can be interpreted as a global measure of one's belief that hypothesis is true, while $pl(A)$ may be viewed as the amount of belief that could potentially be placed in A, if further information became available [13].

The decision rule can be given by differents approches as following :

- Choose the maximum plausibilty hypothesis (*pl*);
- Choose the maximum pignistique probability hypothesis (*BetP*).

$$BetP(w) = \sum_{w \in A} \frac{m(A)}{|A|} \tag{10}$$

Application to Pattern Classification The response of hidden unit i to an input vector x is defined as a decreasing function of the distance between x and a weight vector p^i. The output signal y^j from the j^{th} output unit with weight vector w_i^j is obtained as a weighted sum of the activations in the n hidden layer:

$$y^j = \sum_{i=1}^{n} w_i^j s^i \tag{11}$$

The evidence-theoretic classifier introduced in this paper can also be represented in the connectionist formalism as a neural network with an input layer L_{input}, two hidden layers L_1 and L_2, and an output layer $L_3 = L_{output}$ (Fig. 3). Each layer L_1 to L_3 corresponds to one step of the procedure described in following:

1. Layer L_1 contains n units (prototypes). It is identical to the hidden layer of an RBF network with exponential activation function ϕ and d is a distance computed using data. $\alpha \in [0, 1]$ is a weakning parameter associated to prototype i, where $\epsilon = 0$ at the initialization [14].

$$\begin{cases} s^i = \alpha^i \phi(d^i) \\ \phi(d^i) = \exp(-\gamma^i (d^i)^2) \\ \alpha^i = \frac{1}{1+\exp(-\epsilon^i)} \end{cases} \tag{12}$$

where $(\gamma^i = (\eta^i)^2)$ is a positive parameter defining the receptive field size of prototype $i = \{1, \ldots, n\}$.

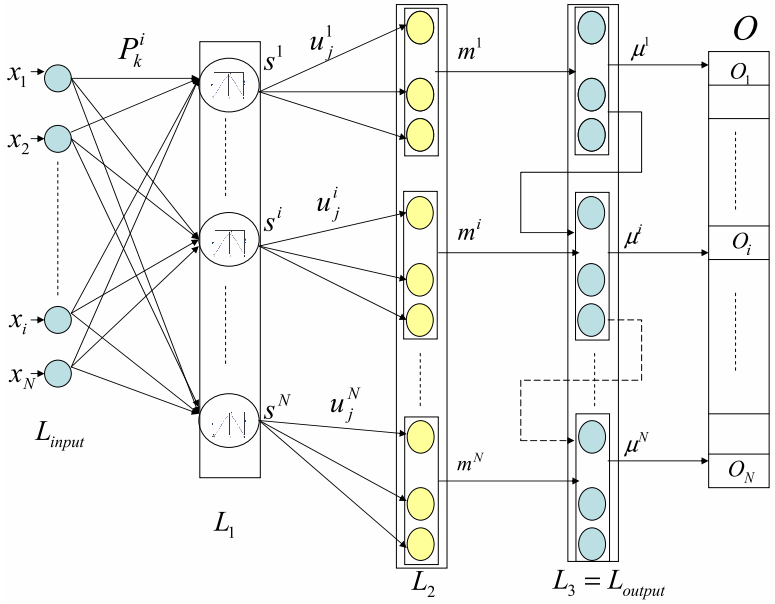

Fig. 3. Neural Network implimentation of the evidence theoritic Classifier Structure

2. Layer L_2 computes the BBA associated to each prototype. It is composed of n modules of $M+1$ units each. The units of module i are connected to neuron i of the previous layer. The vector of activations $m^i = (m_1^i, m_2^i, \ldots, m_{M+1}^i)$ of module corresponds to the belief masses assigned by m^i.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u_q^i \phi(d^i) \\ m^i(\{\Omega\}) = 1 - \alpha^i \phi(d^i) \end{cases} \tag{13}$$

so,

$$m^i = (m^i(\{w_1\}), m^i(\{w_2\}), \ldots, m^i(\{w_{M+1}\})) = (u_1^i s^i, \ldots, u_M^i s^i, 1 - s^i) \tag{14}$$

where u_q^i represents the degree membership to each class w_q, by introducing a new parameter β [14] as $u_j^i = \frac{(\beta_j^i)^2}{\sum_{k=1}^M (\beta_k^i)^2}$.

3. The D-Shefer combination rule combine n different masse function in one single masse. It's given by :

$$m(A) = (m_1 \oplus m_2 \oplus \cdots \oplus m_N) = \sum_{B_1 \cap \cdots \cap B_n = A} \prod_{j=1}^{n} m_j(B_j) \tag{15}$$

This masse function has a particular structure, indeed, the mass restarted only on singleton and γ hypothesis. This particular structure is going to play an important role during the implementation of decision rule.
The n BBA's m^i are combined in L_3, composed of n modules of $M+1$ units. The activations vector of modules i is defined $\vec{\mu^i} = (\mu^i(\{w_1\}), \ldots, \mu^i(\{w_M\}), \mu^i(\Omega))$. where μ^i is the conjunctive combination of the BBA's $m^1, \ldots, m^i$

$$\begin{cases} \mu^i = \bigcap_{k=1}^{i} m^k = \mu^{i-1} \bigcap m^i \\ \mu^1 = m^1 \end{cases} \tag{16}$$

The activation vectors for $i = \{2, \ldots, M\}$ can be recursively computed using the following formula :

$$\begin{cases} \mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{M+1}^i + \mu_{M+1}^{i-1} m_j^i \\ \mu_{M+1}^i = \mu_{M+1}^{i-1} m_{M+1}^i \end{cases} \tag{17}$$

4. Layer L_{output} gives vector O defined as $O = \frac{\mu}{K}$, where $K = \sum_{k=1}^{M+1} m_k$.

The different parameters ($\Delta\beta, \Delta u, \Delta\gamma, \Delta\alpha, \Delta P, \Delta s$) can be determined by gradient descent of output error for a given ν and input pattern x (stopping threshold = 10^{-4} or iteration number = 500).

$$E_\nu(x) = \frac{1}{2}\|P_\nu - t\|^2 = \frac{1}{2}\sum_{q=1}^{M}(P_{\nu,q} - t_q)^2 \tag{18}$$

where $P_{\nu,q} = O_q + \nu O_{M+1}$ is the output vector with $q = 1, \ldots, M$ and $0 \le \nu \le 1$.
$P_{0,q}, P_{1,q}, P_{\frac{1}{M},q}$ represent the Credibility, the plausibility and the pignistique probability rescpectively of each class w_q.
The derivate of $E_\nu(x)$ *w.r.t* β_j^i id given by :

$$\frac{\partial E_v(x)}{\partial \beta_j^i} = \sum_{k=1}^{M} \frac{\partial E_v(x)}{\partial u_j^k} \frac{\partial u_k^i(x)}{\partial \beta_j^i} \tag{19}$$

Let us now compute $\frac{\partial E_v(x)}{\partial u_j^i}$

$$\frac{\partial E_v(x)}{\partial u_j^i} = \frac{\partial E_v(x)}{\partial m_k} \frac{\partial m_k}{\partial u_j^i} = (P_{v,j} - t_j) \frac{\partial m_k}{\partial u_j^i} \tag{20}$$

In order to express $\frac{\partial m_k}{\partial u_j^i}$, we use the commutativity and associativity of the $\bigcap$ operator to rewrite the output BBA m as the conjunctive combination of two terms.

$$m = m^i \bigcap \bar{m}^i \text{with } \bar{m}^i = \bigcap_{k \neq i} \bar{m}^k \tag{21}$$

The vector can be computed by [15]:

$$\begin{cases} \bar{m}_j^i = \frac{m_j - \frac{m_{M+1} m_j^i}{m_{M+1}^i}}{m_j^i + m_{M+1}^i} \\ \bar{m}_{M+1}^i = \frac{m_{M+1}}{m_{M+1}^i} \end{cases} \tag{22}$$

so,

$$\frac{\partial m_k}{\partial u_j^i} = s^i(\bar{m}_j^i + \bar{m}_{M+1}^i) \tag{23}$$

and,

$$\frac{\partial E_v(x)}{\partial u_j^i} = (P_{v,j} - t_j) s^i (\bar{m}_j^i + \bar{m}_{M+1}^i) \tag{24}$$

$$\frac{\partial E_v(x)}{\partial \eta^i} = \frac{\partial E_v(x)}{\partial s^i} \frac{\partial s^i}{\partial \epsilon^j} = \frac{\partial E_v(x)}{\partial s^i} (-2\eta^i (d^i)^2 s^i) \tag{25}$$

$$\frac{\partial E_v(x)}{\partial \epsilon^i} = \frac{\partial E_v(x)}{\partial s^i} \exp(-(\eta^i d^i)^2)(1 - \alpha^i)\alpha^i \tag{26}$$

$$\frac{\partial E_v(x)}{\partial p_j^i} = \frac{\partial E_v(x)}{\partial s^i} \frac{\partial s^i}{\partial p_j^j} = \frac{\partial E_v(x)}{\partial s^i} (2(\eta^i)^2 s^i (x_j - p_j^i)) \tag{27}$$

we need to compute $\frac{\partial E_v(x)}{\partial s^i}$:

$$\begin{aligned} \frac{\partial E_v(x)}{\partial s^i} &= \sum_{k=1}^{M} \frac{\partial E_v(x)}{\partial P_{v,k}} \frac{\partial P_{v,k}}{\partial s^i} = \sum_{j=1}^{M} (P_{v,j} - t_j)\left(\frac{\partial m_j}{\partial s^i} + \nu \frac{\partial m_{M+1}}{\partial s^i}\right) \\ &= \sum_{j=1}^{M} (P_{v,j} - t_j)(u_j^i(\bar{m}_j^i + \bar{m}_{M+1}^i) - \bar{m}_j^i - \nu \bar{m}_{M+1}^i) \end{aligned}$$

4 Experiments

Experiments are conducted on the TrecVid'05 databases [3]. It represents a total of over 85 hours of broadcast news videos from US, Chinese, and Arabic sources. About 60 hours are used to train the feature extraction system and the remaining for the evaluation purpose. The training set is divided into two subsets in order to train classifiers and subsequently the fusion parameters. The evaluation is realized in the context of TrecVid'05 and we use the common evaluation measure from the information retrieval community: the Average Precision.

The feature extraction task consists in retrieving shots expressing one of the following semantic concepts: *1:Building, 2:Car, 3:Explosion or Fire, 4:US flag, 5:Map, 6:Mountain, 7:Prisoner, 8:Sports, 9:People walking/running, 10:Waterscape, 11:MAP (Mean Average Precision)*.

The RBF and NN-ET were trained with the same optimization algorithm (gradient descent). The number n of prototypes was varied between 2 and 10. For each value of n, the average training error rates are computed. Our method yields better results for small values of n and similar performance for higher values of n. The best number is $n = 5$, where we obtain the lower training error.

Figure 4 shows Mean Precision results of the two classifiers fusion methods compared in this work: the standard RBF and the evidence theory neural networks (NN-ET). The improvement in mean precision is clearly visible for all semantic concepts using NN-ET. It is a foreseen result since in the decision rule RBF takes just the *a posterior* probability. NN-ET, in contrast, convert this probability in the form of BBA's, which are then combined using D-shafer rule combination. The fusion output

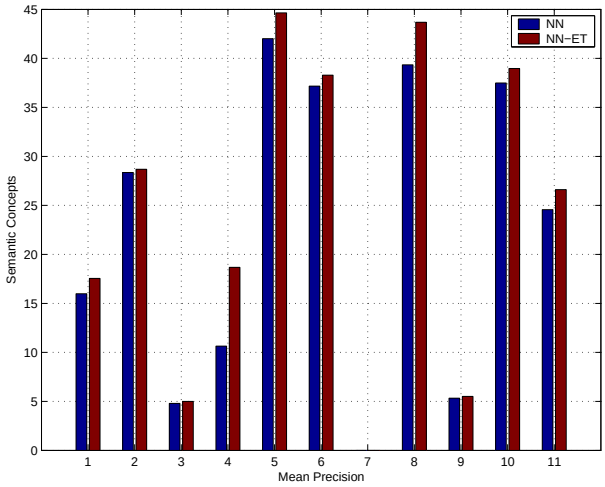

Fig. 4. Comparaison of RBF neural network and Neural Network based on Evidence Theory (NN-ET) fusion method.

can be presented as a belief function defining for each class a posterior probability interval. The width of this interval can be used as a mesure of the uncertainty attached to a fusion. This approch has been shown to allow decision making with reject options, and to have good classifier fusion performance as compared to other methods.

Besides, NN-ET presents more improvement for the concepts (4, 5, 8) that on the rest, it can be explained, by the high number of false decision in classification using just the *posterior* probability, Evidence theory resolve this inconvenient, introducing the degree of belief in our probability and the ignorance of our system.

We also notice, a precision equal to zero for the concept (7), it can be explained by the fact that there is no video shot that represents this concept in the Trecvid'05 data test.

5 Conclusion

In this paper, we have presented an automatic semantic video content indexing and retrieval system. The reported system first employs visual features (HSV Histogram, Gabor filters) in order to obtain a compact and effective representation, followed by SVM based classification to solve the challenging task of video shot content detection. two methods for combining classifiers are investigated in details. The RBF and Neural network based on Evidence Theory approach that it managed all the features most effectively and appears therefore to be particularly well suited for the task of classifier fusion.

This approach is based on a feeling of uncertainty to the classification model, considering complete or partial knowledge of the class. Inferior and superior expectations as well as of pignistique probability, propose several strategies of decision with arbitrary costs. We think that this methodology can be useful in the situations where the available informations are very incomplete and soiled by uncertainty.

We have started to investigate the effect of the addition of many other visual features (Dominant Color, RGB, Canny edges features,...) as well as audio features (MFCC, PLP, FFT), to see their influence on the final result. The addition of other modalities will allows us to evaluate how the different approaches are able to deal with potentially irrelevant data. In parallel, we have initiated a program of work about descriptor fusion. We believe such an approach, which may be seen as normalization and dimensionality reduction, will have considerable effect on the overall performance of multimedia content analysis algorithms.

Acknowledgement

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.

References

1. H. Eidenberger, *Statistical analysis of content based mpeg7 descriptors for image retrieval*, ACM Multimedia Systems journal, Springer, 2004.
2. M. Naphade, T. Kristjansson, B. Frey, and T. Huang, *Probabilistic multimedia objects (multijets): a novel approach to video indexing and retrieval*, IEEE Trans. Image Process., vol. **3**, pp. 536–540, 1998.
3. TRECVID, *Digital video retrieval at NIST*, `http://www-nlpir.nist.gov/projects/trecvid/`.

4. P. Felzenszwalb and D. Huttenlocher, *Efficiently computing a good segmentation*, Proceedings of IEEE CVPR, pp. 98–104, 1998.
5. F. Souvannavong, B. Merialdo, and B. Huet, *Multi modal classifier fusion for video shot content retrieval*, *Proceedings of WIAMIS*, 2005.
6. W. Ma and H. Zhang, *Benchmarking of image features for content-based image retrieval*, Thirtysecond Asilomar Conference on Signals, System and Computers, pp. 253–257, 1998.
7. C. Carson, M. Thomas, and S. Belongie, *Blobworld: A system for region-based image indexing and retrieval*, Third international conference on visual information systems, 1999.
8. E. Allwein, R. Schapire, and Y. Singer, *Reducing multiclass to binary : A unifying approach for margin classifiers.* Journal of Machine Learning Research, vol. 1, pp. 113–141, 2000.
9. V. Vapnik, *The nature of statistical learning theory*. Springer, 1995.
10. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, ch. Kernel-Induced Feature Spaces, 2000.
11. C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, ch. Radial Basis Functions, 1995.
12. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.
13. P. Smets and R. Kennes, *The Transferable Belief Model*, Artif. Intell, vol. **66,** pp. 191–243, 1994.
14. T. Denoeux, *An Evidence Theoretic Neural Network Classifier*, IEEE International Conference on Systems, Man and Cybernetics, vol. **3**, pp. 712–717, 1995.
15. T. Denoeux. *A Neural Network Classifier Based on Dempster-Shafer Theory,* IEEE transactions on Systems, Man and Cybernetics, vol. **2**, pp. 131-150, 2000.