Neural Network Combining Classifier Based on Dempster-Shafer Theory for Semantic Indexing in Video Content

Rachid Benmokhtar and Benoit Huet

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

Abstract. Classification is a major task in many applications and in particular for automatic semantic-based video content indexing and retrieval. In this paper, we focus on the challenging task of classifier output fusion[1]. It is a necessary step to efficiently estimate the semantic content of video shots from multiple cues. We propose to fuse the numeric information provided by multiple classifiers in the framework of evidence logic. For this purpose, an improved version of RBF network based on Evidence Theory (NN-ET) is proposed. Experiments are conducted in the framework of TrecVid high level feature extraction task that consists of ordering shots with respect to their relevance to a given semantic class.

1 Introduction

Classifier fusion is a promising way for improving the performance of pattern recognition algorithms. Many authors proposed different ways of fusing classifiers [1,2,3]. In [4], a state of the art is presented, along with a dichotomy and an evaluation of differents classifeurs fusion methods used in the literature. Neural Network approaches seem to be able to give the better performances than GMM and Decision Template [3,4]. In the aim of the neural network study, this paper gives a novel fusion method inspired by RBF neural network and evidence theory, called Neural Network based on Evidence Theory (NN-ET). These methods are implemented for this purpose and evaluated in the context of content-based retrieval of video data.

This paper presents a novel fusion scheme based on neural network which is build within our semantic video content indexing and retrieval system. First, an overview of the architecture is given. A description of RBF neural network is then provided along with an explanation of how evidence theory can be used for classification and fusion. The experimental results presented in this paper are performed 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

[1] The work presented here is funded by France Télécom R&D under CRE 46134752.

T.-J. Cham et al. (Eds.): MMM 2007, LNCS 4352, Part II, pp. 196–205, 2007.

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 [5]. The algorithm is fast and provides visually acceptable segmentation. Its low computational requirement is an important criterion when one needs to process a huge amount of data like the TrecVid'05 database. An illustration of the segmentation result is provided on figure 1. Secondly, color and texture are extracted for each segmented region. Thirdly, vectors obtained over the complete database are clustered using K-Means to find the N most representative elements.

Fig. 1. Example of segmentation outputs

Representative elements are then used as visual keywords to describe video shot content. To do so, features computed from a single video shot are matched to their closest visual keyword with respect to Euclidean distance (or another distance measures). The occurrence vector of the visual keywords in the shot, called Image Vector Space Model (IVSM) is then build. 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 initial classification which will then be used by the fusion mechanism [6]. The overall chain is presented in figure 2.

For the study presented in this paper we distinguish two types of modalities : visual and motion features. The two visual features are selected for this purpose: Hue-Saturation-Value color histograms and energies of Gabor's filters [7]. 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 [8]. For some concepts like people walking/running, sport, it is useful to have an information about the motion activity present in the shot. Two features are selected for this purpose: the camera motion and the motion histogram of the shot.

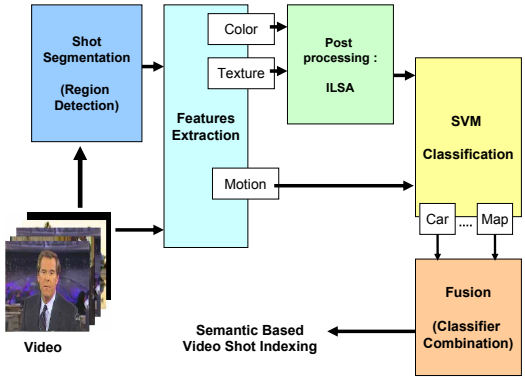

Fig. 2. General framework of the application

3 Classifier Fusion

3.1 Radial Neural Network (RBF)

RBF is a popular supervised neural network learning algorithm, which consists in a spacialization of the MLP network [9]. The RBF network is constitued by only the following three layer, as shown in (figure 3).

- *Input Layer :* Broadcast the inputs without distortion to hidden layer;
- *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 the basis function and σ_i is a width parameter that controls how is spread the curve. Generally, these centers are selected by using some

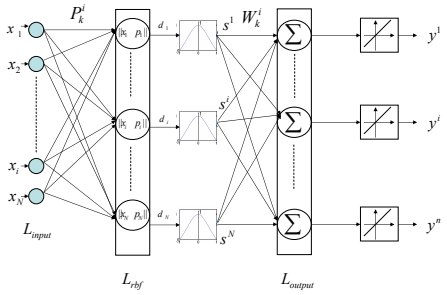

Fig. 3. RBF Classifier Structure

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 parameters (ΔW, $\Delta \mu$, $\Delta \sigma$) are given by (more detailed explanation can be found in [9]) :

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

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 in [4], 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 [10].

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 :

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

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 [11].

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

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

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

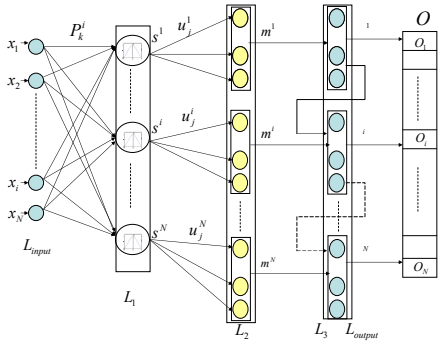

Fig. 4. Neural Network implementation of the evidence theoritic Classifier Structure

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 [12].

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

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

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 $m^i = (m_1^i, m_2^i, ..., m_{M+1}^i)$ 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,

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

where u_q^i represents the degree membership to each class w_q, by introducing a new parameter β [12] as $u_j^i = \frac{(\beta_j^i)^2}{\sum_{k=1}^{M} (\beta_k^i)^2}$.

3. The Dempster Shafer combination rule combine n different mass function in one single mass. It's given by :

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

This mass 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 $\vec{\mu^i} = (\mu^i(\{w_1\}), ..., \mu^i(\{w_M\}), \mu^i(\Omega))$. where μ^i is the conjunctive combination of the BBA's $m^1, ..., m^i$

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

The activation vectors for $i = \{2, ..., M\}$ 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:

$$\begin{cases} O = \frac{\mu}{K} \\ K = \sum_{k=1}^{M+1} m_k \end{cases} \tag{18}$$

The different parameters ($\Delta\beta$, Δu, $\Delta\gamma$, $\Delta\alpha$, ΔP, Δs) can be determined by gradient descent of output error for a given ν and input pattern x.

$$E_v(x) = \frac{1}{2}||P_v - t||^2 = \frac{1}{2}\sum_{q=1}^{M}(P_{v,q} - t_q)^2 \tag{19}$$

where $P_{v,q} = O_q + \nu O_{M+1}$ is the output vector with $q = 1, ..., M$ and $0 \leq \nu \leq 1$.

$P_{0,q}, P_{1,q}, P_{\frac{1}{M},q}$ represent the credibility, the plausibility and the pignistique probability rescpectively of each class w_q.

The derivate of $E_v(x)$ *w.r.t* β_j^i id given by :

$$\frac{\partial E_v(x)}{\partial \beta_j^i} = \sum_{k=1}^{M} \frac{\partial E_v(x)}{\partial u_j^k} \frac{\partial u_k^i(x)}{\partial \beta_j^i} \tag{20}$$

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

$$\frac{\partial E_\nu(x)}{\partial u_j^i} = \frac{\partial E_\nu(x)}{\partial m_k} \frac{\partial m_k}{\partial u_j^i} = (P_{\nu,j} - t_j)\frac{\partial m_k}{\partial u_j^i} \tag{21}$$

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.

$$m = m^i \bigcap \bar{m}^i \text{with } \bar{m}^i = \bigcap_{k \neq i} \bar{m}^k \tag{22}$$

The vector can be computed by [13]:

$$\begin{cases} \bar{m}_j^i = \frac{m_j - \frac{m_{M+1} m_j^i}{m_{M+1}^i}}{m_j^i + m_{M+1}^i} \\ \bar{m}_{M+1}^i = \frac{m_{M+1}}{m_{M+1}^i} \end{cases} \tag{23}$$

so,

$$\frac{\partial m_k}{\partial u_j^i} = s^i(\bar{m}_j^i + \bar{m}_{M+1}^i) \tag{24}$$

and,

$$\frac{\partial E_\nu(x)}{\partial u_j^i} = (P_{\nu,j} - t_j)s^i(\bar{m}_j^i + \bar{m}_{M+1}^i) \tag{25}$$

$$\frac{\partial E_\nu(x)}{\partial \eta^i} = \frac{\partial E_\nu(x)}{\partial s^i}\frac{\partial s^i}{\partial \epsilon^j} = \frac{\partial E_\nu(x)}{\partial s^i}(-2\eta^i(d^i)^2 s^i) \tag{26}$$

$$\frac{\partial E_\nu(x)}{\partial \epsilon^i} = \frac{\partial E_\nu(x)}{\partial s^i}\exp\left(-(\eta^i d^i)^2\right)(1-\alpha^i)\alpha^i \tag{27}$$

$$\frac{\partial E_\nu(x)}{\partial p^i_j} = \frac{\partial E_\nu(x)}{\partial s^i}\frac{\partial s^i}{\partial p^j_j} = \frac{\partial E_\nu(x)}{\partial s^i}(2(\eta^i)^2 s^i(x_j - p^i_j)) \tag{28}$$

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

$$\begin{aligned}\frac{\partial E_\nu(x)}{\partial s^i} &= \sum_{k=1}^{M}\frac{\partial E_\nu(x)}{\partial P_{v,k}}\frac{\partial P_{\nu,k}}{\partial s^i} = \sum_{j=1}^{M}(P_{\nu,j}-t_j)(\frac{\partial m_j}{\partial s^i} + \nu\frac{\partial m_{M+1}}{\partial s^i}) \\ &= \sum_{j=1}^{M}(P_{\nu,j}-t_j)(u^i_j(\bar{m}^i_j + \bar{m}^i_{M+1}) - \bar{m}^i_j - \nu\bar{m}^i_{M+1})\end{aligned}$$

4 Experiments

Experiments are conducted on the TrecVid'05 databases [14]. 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:Mean Average Precision (MAP).*

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 proposed approach 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 5 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 Dempster Shafer rule combination. The fusion output 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 approach has been shown to allow decision making with reject options, and to have good classifier fusion performance as compared to other methods.

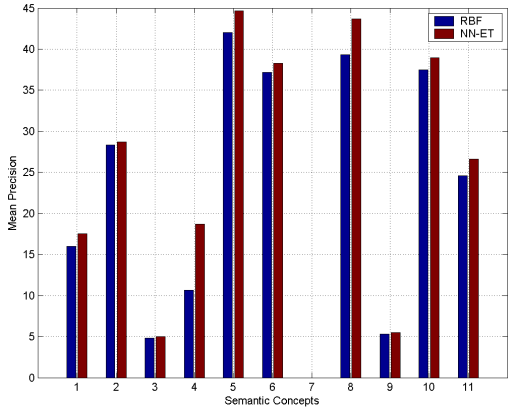

Fig. 5. Comparaison of RBF neural network and Neural Network based on Evidence Theory (NN-ET) fusion method

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 only the *posterior* probability, Evidence theory resolve this issue, 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 test data.

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.

References

1. L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their application to hardwriting recognition," *IEEE Trans. Systems Man Cybernet*, vol. 22, pp. 418–435, 1992.
2. R. Duin and D. Tax, "Experiements with classifier combining rules," *Proc. First Int. Workshop MCS 2000*, vol. 1857, pp. 16–29, 2000.
3. L. Kuncheva, J.C.Bezdek, and R. Duin, "Decision templates for multiple classifier fusion : an experiemental comparaison," *Pattern Recognition*, vol. 34, pp. 299–314, 2001.
4. R. Benmokhtar and B. Huet, "Classifier fusion : Combination methods for semantic indexing in video content," *Proceedings of ICANN*, vol. 2, pp. 65–74, 2006.
5. P. Felzenszwalb and D. Huttenlocher, "Efficiently computing a good segmentation," *Proceedings of IEEE CVPR*, pp. 98–104, 1998.
6. F. Souvannavong, "Indexation et recherche de plans video par contenu semantique," Ph.D. dissertation, Phd thesis of Eurecom Institute, France, 2005.
7. 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.
8. 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.
9. C. Bishop, "Neural networks for pattern recognition," *Oxford University Press, ch. Radial Basis Functions*, 1995.
10. G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.
11. P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–243, 1994.
12. T. Denoeux, "An evidence theoretic neural network classifier," *IEEE. International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 712–717, 1995.
13. ——, "A neural network classifier based on dempster-shafer theory," *IEEE transactions on Systems, Man and Cybernetics*, vol. 2, pp. 131–150, 2000.
14. TRECVID, "Digital video retrieval at NIST," *http://www-nlpir.nist.gov/projects/trecvid/*.