Multi-Channel Bayesian ART for Robot Fusion Perception

Wei Hong, Chin*, Chu Kiong, Loo*, Naoki Masuyama*
*Faculty of Computer Science and Information Technology
University of Malaya, Kuala Lumpur, 50603 Malaysia
Email: weihong118118@gmail.com, ckloo.um@um.edu.my, naoki.masuyama17@gmail.com

***Abstract*—Multiple sensor data fusion is the technique of associate information from a number of different sensors to produce a robust and comprehensive description. Data fusion pose is using in various robotics application such as environment mapping, object recognition and robot localization. Their relation is generally hard coded and difficult to learn incrementally if new objects or events arise. In this paper, we propose a new learning architecture termed as Multi-Channel Bayesian ART which is very flexible can be adapted to new domains or different sensor configurations easily. The other advantages of the proposed method are: 1) it is capable of incremental on-line learning without forgetting previously-learned knowledge 2) It can process data real time and does not require any prior training to make it work in natural environment. The effectiveness of our proposed method is validated by real experimental results implemented on robot.**

***Keywords*—Hybrid map, Intelligent robotics, Mobile robot, Navigation, SLAM,**

I. Introduction

Multi-sensory information is a universal perception considering such information is of involvement in all robotic systems where information processing is essential. In such frameworks for the improvement of the precise activity data repetitive sensors are vital where the quantity of the sensors as well as the resolution of the sensors can change due to data with various sampling time from the sensors. Distinctive sensors can have distinctive benefits relying upon their individual working conditions and such assorted data can be a profitable addition for precise and additionally reliable autonomous robot control through its dynamics and kinematics. The challenge for this situation is the unification of the regular data from different sensors in a manner that the resultant data presents enhanced information for desired activity. Autonomous robotics constitutes a crucial chapter of robotics and the autonomous robotics research is extensively described in previous work, e.g. [1], [2], [3]. In this branch of robotics continuous data from the environment is collected by sensors and processed in real time. The precise and reliable data driving the robot is fundamental for a safe navigation the direction of which is in general not defined in advance. The accuracy of this information is to accomplish by method for both physical and analytical repetition of the sensors. The precision is obtained by integrating the sensory information from the multiple sensors in a multi-sensor system. This coordination is completed by joining data from various sensors for a final estimation result and this is widely part termed as sensor combination.

978-1-4799-7560-0/15/$31 ©2015 IEEE

Robotic systems that can perform tasks autonomously and capable of decide which behavior is the best to complete the task usually possess multiple sensors such as a laser scanner, a camera and other sensor system. If these sensory data can be associated in an appropriate approach, the response of the robot can be more reliable because of the higher quality of data for the classification. Moreover, the environment of the robot is not often static, for example, new persons, new obstacles or furniture relocation might happen. Therefore, the robot should be able to handle with new situation and adapt its reaction, that is, it should learn to distinguish these new objects.

The ultimate goal of data processing as fusion is to enable the framework to measure the state of the environment and specifically we can refer to the condition of a robot's environment in the present case. A similar research handling with this challenge, to be specific a multiresolutional filter application for spatial data fusion in robot navigation has been described previously where data fusion is accomplished using several datasets gathered from wavelet decomposition and not from individual sensors [4]. Besides, several machine learning methods are developed to perform data fusion [5]. For example, a time-delayed neural network (TDNN) is implemented in an automatic lipreading system to fuse audio and visual information. A major limitation of these network is the problem of catastrophic forgetting, i.e., learned associations from input data to output classes could be negatively affected if the network is trained on-line.

Many of the previous research work for classifying information from various sorts of sources are generally based on two concepts. For sensor fusion, the obtained information per channel can be combined in advance (data-level fusion) [6], [7] and the classification is done later with all information. On the other hand, a classifier is connected to every sensor channel separately and the decision of every classifier is consolidated subsequently (decision-level fusion) [8], [9]. However, both data- and decision-level approaches suffer from the predefined fixed architecture and therefore adapt inadequately to new configurations of sensor channels.

In this paper, we propose an algorithm for integrating data originating from multiple sensors to overcome classification tasks like those already stated. To be more specifically, we propose a robotic system that capable of learn and classify objects in its environment. The proposed method is based on the Bayesian Adaptive Resonance Theory (ART), and we have developed a new Simplified Bayesian ARTMAP network optimized for classification works based on muli-modal information. Our proposed framework overcomes problems

from processing sensor channels independently from each other, which can lead to loss of inter-dependencies of sensor channels. In addition, the proposed method prevents the early combination of sensor information without considering the different modalities of the sensor channels.

The rest of the paper is organized as follows. Section II introduces the theoretical framework of the proposed. The experimental results shown in Section III while experiment result are discussed in Section IV. Concluding remarks are finally presented in Section V.

II. Multi-Channel Bayesian ART

Our proposed method architecture consists of two layers, as shown in Figure . The input layer is formed by multiple Bayesian ARTs to learn and categorize multiple sensory information. The higher level of layer is Simplified Bayesian ARTMAP module. The proposed method learns a mapping from input vectors provided by multiple sensors. The M-dimensional sensory information is transmitted to Bayesian ART channels and the learning undergo 3 main process which is i) Neuron competition, ii) Neuron matching and iii) Neuron learning.

The definition of neurons in each channels can be interpreted as a region in the input space. Each neuron contains a multidimensional Gaussian distribution, with mean vector $\hat{\mu}_j$, covariance matrix $\hat{\Sigma}_j$, and a prior probability $\hat{P}(w_j)$. The learning process is initialized with 3 parameters: the maximal hypervolume S_{MAX}, the initial covariance matrix $\hat{\Sigma}_{\text{init}}$, and a prior probability $\hat{P}(w_j)_{init}$.

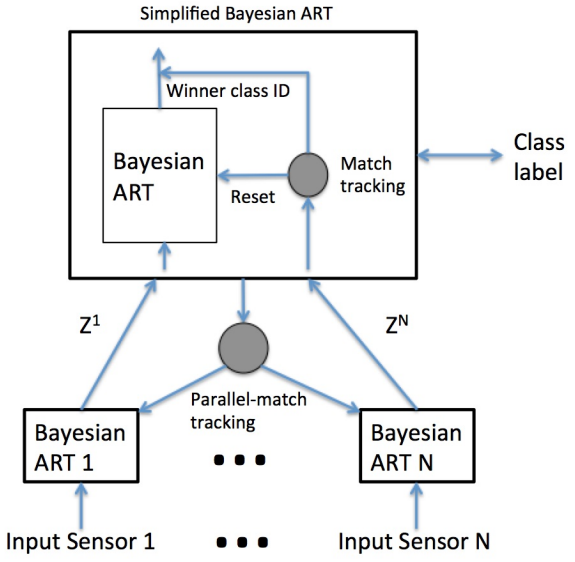


Fig. 1: Multi-channel Bayesian ART framework

A. Bayesian ART

In this section, we explain the Bayesian ART learning process. 1) *Neuron Competition:* In this stage, all existing neuron compete with each other to map the input data. The *a posteriori* probability of the *j*th neuron to represent the *M*-dimensional data $\boldsymbol{x}$ is calculated as follows:

$$M_j = \hat{P}(w_j|x) = \frac{\hat{p}(x|w_j)\hat{P}(w_j)}{\sum_{l=1}^{N_{cat}} \hat{p}(x|w_l)\hat{P}(w_l)} \tag{1}$$

where N_{cat} is the number of neurons and $\hat{P}(w_j)$ is the estimated prior probability of the *j*th neuron. The likelihood of w_j with respect to $\boldsymbol{x}$ is estimated using all data that have already been associated with the multivariate Gaussian neuron w_j:

$$\hat{p}(x|w_j) = \frac{1}{(2\pi)^{M/2}|\hat{\Sigma}_j|^{1/2}} \times \exp\{-0.5(x-\hat{\mu}_j)^T\hat{\Sigma}_j^{-1}(x-\hat{\mu}_j)\} \tag{2}$$

where $\hat{\mu}_j$ and $\hat{\Sigma}_j$ are the estimated mean and covariance matrix of the *j*th neuron.

The winning neuron $\boldsymbol{J}$ is the one with the maximum *a posteriori* probability (MAP):

$$J = \arg\max(M_j) \tag{3}$$

This proves that $\boldsymbol{J}$th neuron w_J is either more populated than other neurons (i.e. having high $\hat{P}(w_j)$) or more likely to be the correct neuron for $\boldsymbol{x}$ (i.e. having high $\hat{p}(x|w_j)$ since it is the closest neuron to $\boldsymbol{x}$) or both. Based on both probabilities and Bayes' theorem, the MAP criterion is expected to choose a winning neuron more accurately than if using only one of the probabilities. For example, the MAP criterion may prefer a neuron having *a priori* probability which is higher than that of another node although the normalized by the covariance distance of the former to the pattern is larger than that of the latter.

2) *Neuron Matching* (Vigilance Test): The neuron match is to ensure that the chosen neuron is able to represent the current data received from sensors. The test restricts the *J*th neuron hypervolume S_J to the maximal hypervolume allowed for a neuron S_{MAX}:

$$S_J \leq S_{\text{MAX}} \tag{4}$$

where the hypervolume is defined as the determinant of the Gaussian covariance matrix. For a diagonal covariance matrix, this hypervolume is reduced to the product of the variances each for a dimension:

$$S_J \triangleq \det(\Sigma_J) = \prod_{d=1}^{M} \sigma_{J_d}^2 \tag{5}$$

If the winning neuron fulfills the vigilance criterion in equation (4), learning is then performed. If the winning neuron fails the vigilance criterion, the neuron is removed from the competition for this sensory input and the Bayesian ART determines another neuron till one complies with equation (4).

3) *Neuron Learning*: When a chosen neuron fulfills the maximal hypervolume criteria in equation (4), the node elements are then updated as follows:

$$\hat{\mu}_{J(\text{new})} = \frac{N_J}{N_J+1}\hat{\mu}_{J(\text{old})} + \frac{1}{N_J+1}x \tag{6}$$

$$\hat{\Sigma}_{J(\text{new})} = \frac{N_J}{N_J+1}\hat{\Sigma}_{J(\text{old})} + \frac{1}{N_J+1}(x-\hat{\mu}_{J(\text{new})}) \times (x-\hat{\mu}_{J\text{new}}))^T * I \quad (7)$$

$$\hat{P}(w_J) = \frac{N_J}{\sum_{l=1}^{N_{node}} N_l} \quad (8)$$

$$N_J^{\text{new}} = N_J^{\text{old}} + 1 \quad (9)$$

where N_J is the number of times that ***J***th neuron have been chosen as winner for learning before receiving the current input sensory information and ***I*** is the identity matrix.

Another possibility is that the network not able to find any neuron that match the current input. In this case, an uncommitted neuron will be activated. Thus, the learning algorithm incrementally produces neurons to learn and remember feature patterns for representing the received input data.

B. Simplified Bayesian ARTMAP

In this paper, we have developed the Simplified Bayesian ARTMAP (SBAM) which is an extension of Bayesian ART suitable for supervised learning. However, compare to Bayesian ARTMAP, SBAM networks are classifiers, i.e., they learn a mapping from input vectors to a particular set of discrete class labels. This difference to Bayesian ARTMAP greatly decreases the complexity of the network. It only comprises a single Bayesian ART component whose $F2$ neurons were extended with an associated class label.

During learning, a SBAM network receives pairs of an input vector and an associated correct class label. The input is transmitted to the internal Bayesian ART, then the class label of the winner neuron will be compared to the given class label. Regardless of matching or not matching, the match-tracking algorithm temporary raises the vigilance to force the choosing of another neuron. A new neuron will be committed if all nodes are not match in the $F2$ layer. A trained SBAM framework can be utilized to predict a class label for a new input.

C. Overall Learning Process

The input of the Bayesian ART modules itself can be produced by different sensors or different features calculated on the sensory information from one physical sensor. Thus, the input channels do not compulsory have to represent a sensor but could also be a feature vector. In this framework, we implement the parallel match-tracking algorithm based on the information as far as available. However, it had to be adapted to our learning system [10], [11]. A detailed explanation of the functionality of the concrete implemented algorithm will follow within this section.

During learning, the SBAM network receives the correct class label in addition to the current input. Therefore, a training input consists of $i = 1, ..., N$ feature vectors $\vec{v}^i$ and the target class label. For the first iteration, each Bayesian ART module i tempt to assign its input to a known category. If that fails, it creates a new category with $\vec{v}^i$ as initial weight vector. When all Bayesian ART networks have categorized their input, a vector $\vec{z}$ termed *recognition code* is created from the weight vectors of the winner neurons. It is crucial to notice that the ART modules are not yet to perform a training step with their inputs. The input vector for the SBAM module is stated as follow: $\vec{z}_c = (\vec{z}^1, \ldots, \vec{z}^N)$ where the $\vec{z}^i$ are generated by

$$\vec{z}^i = \frac{N_J}{N_J+1}\hat{\mu}_J^i + \frac{1}{N_J+1}x^i \quad (10)$$

In Eq. 10, $\hat{\mu}_J^i$ represents the weight vector of the winner neuron of the i'th Bayesian ART module. If the SBAM categorizes the concatenated vector $\vec{z}_c$ into a category whose class label matches the expected class then all Bayesian ART modules and the SBAM are updated respectively. Otherwise, the parallel match-tracking algorithm is activated.

The parallel match-tracking algorithm requires the introduction of base hypervolume for each network module. The base hypervolume for the SBAM is denoted by S_{SBAM} and by S_{ARTi} (with $i = 1, \ldots, N$) for Bayesian ART modules. These base hypervolumes store the parameter values which were set at the beginning. The currently used hypervolume of the network is then indicated as the working hypervolume. Every time a new input is transferred to the network, all working hypervolumes are set back to their base hypervolume values. In fact, the base hypervolumes do not need to have the same value.

At the beginning, the parallel match-tracking process searches for the Bayesian ART module with the lowest confidence, that means the one with the highest hypervolume value. This module is denoted as ART_{lc}. If more than one Bayesian ART has the same matching value, the first one in the list is selected. Next, the hypervolume of all Bayesian ART modules and the SBAM is raised just enough so that the ART_{lc} resets the winner neuron. To fulfill this, a δ value is calculated by using the matching value of the winner neuron of ART_{lc}. The value of δ is calculated as follows:

$$\prod_{d=1}^{M} \sigma_{lc}^2 + \varepsilon - S_J, \ 0 < \varepsilon \leq 1 \quad (11)$$

This δ value is used to increase all working hypervolume values as follow:

$$S_i^{\text{new}} = S_i^{\text{old}} - \delta, \ \forall i \in \{\text{SBAM}, ART_1, \ldots, ART_N\} \quad (12)$$

The working hypervolume is set to the SBAM and the Bayesian ART networks by the parallel match-tracking algorithm. The parallel match-tracking here changes the hypervolume of the SBAM module externally which becomes the internal base hypervolume of the SBAM. In this way, the least confidential channel is blamed for the misclassification. Thus, not the entire network needs to change but rather just the part which is most likely the cause for the omission.

The ART_{lc} will select another category as the decrease of the hypervolume (12) will cause (4) to fail for this Bayesian ART network. Thus, another weight vector is chosen, which leads to a changed recognition code $\vec{z}_c$. If the uncommitted neuron is one Bayesian ART is chosen or its hypervolume reaches 0, this network has to commit a new neuron. This process will be repeated until the SBAM network classifies the input correctly. Only if all Bayesian ART modules are committed a new neuron, the SBAM also has to commit a new

neuron, and therefore, a new category as well. This category is labeled with new label.

If the proposed method is learned, the network is used to provide class predictions for new input data, the class label is provided as output.

III. Experimental Design & Result

Our proposed method was verified on data sets collected from H20 robot as shown in Figure 2. For analysis, two types of data are collected from a camera and a laser range finder (LRF). The visual data were recorded with a resolution of 640 x 320 pixels. The objects contained in this dataset are a person and a chair in front of the robot recorded in different distances and angles to the robot. Samples of the image are shown in Figure 3 and Figure 4 shows the image captured by the robot from different angle.

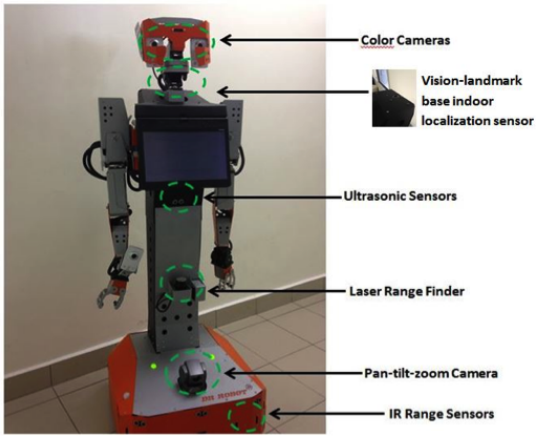
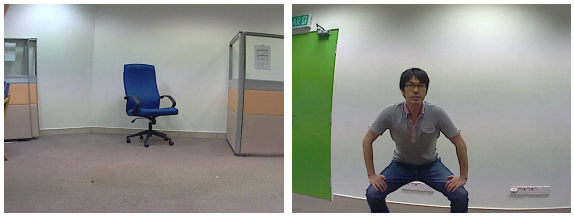


Fig. 2: H20 robot that gather experimental data.



(a) Chair subject (b) Human subject

(c) Sofa subject (d) Rack subject

Fig. 3: Sample pictures of test subjects

Fig. 4: Sample pictures of test subjects taken from different angle

Based on the raw data from the camera and laser range finder, first object detection is implemented on each sensor channel individually. For the camera image, the visual feature is a simple histogram computed on the value channel of the image converted into HSV color space. On the other hand, the feature extraction of laser range measurements is performed by dividing the overall measurement into segments of consecutive scan values with nearly the same length. This can be done by calculating the length difference of each pair of neighboring distance measurements and a threshold.

In this dataset, data are labeled manually and the decision which laser scan region belongs to which visual object is also determined by hand. In each picture of dataset only one object is included. The overall tested data contains 50 value sets for a chair, a human, a sofa and a rack that located in front of the robot.

In the first experiment, we validate the proposed method by clustering objects separately and determine the number of neurons are needed for representing these objects. Figure 5 illustrates the classification accuracy of the proposed method with different value of hypervolume. Next, Figure 6 shows the number of nodes added during the learning process. In addition, the error rates are used for evaluating the classification performance which is *false negative rate* (FNR). The FNR accumulates the errors where an object was not correctly detected. The experiment result of FNR value for our proposed method is plotted in Figure 7. In addition, Table I shows the confusion matrices for the proposed method to illustrate which errors occur and in which quantity. From the experiment result, the proposed method correctly classify the chair 47 times, the human 48 times, the sofa 45 times and the rack 46 times out of 50 trials. Next, it also capable of classify the human subject 48 times correctly out of 50 trials.

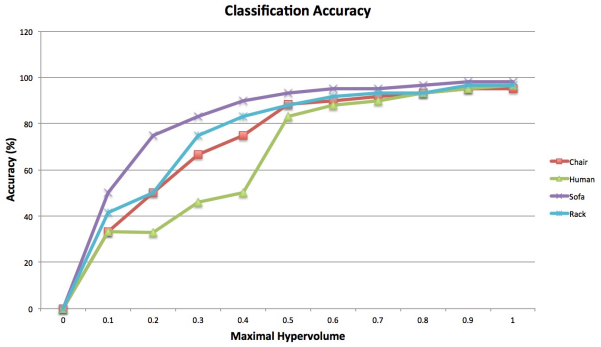


Fig. 5: Classification accuracy for different value of hypervolume.

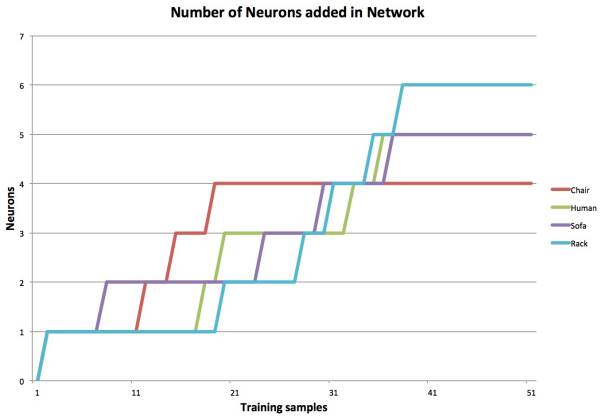


Fig. 6: Number of neuron added during learning process.

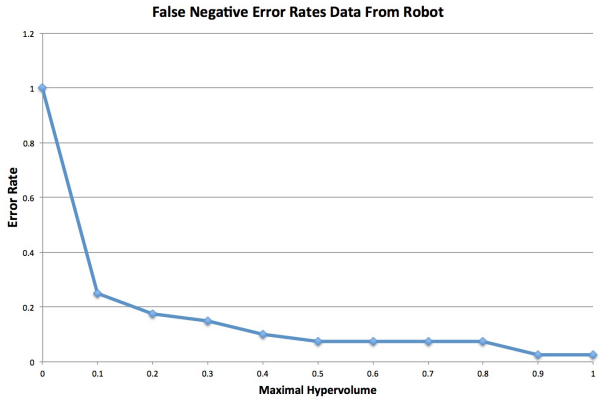


Fig. 7: Error rates for data gathered from the robot.

IV. Discussion

We have shown that the proposed framework is able to learn and classify objects and human without any prior knowledge. Based on the experimental plot, the optimum hypervolume value for learning is from 0.9 to 1. The performance of the framework is decreasing when maximal hypervolume is decreasing.

In the experiment, the proposed method classified incorrectly due to the noise of the image and laser scanner data collected from the robot. In addition, the graph shows that the optimum maximal hypervolume value for learning and classification is 1. All the experiment was run in real time and computed by the moderate specifications of laptop.

V. Conclusion

The experiments presented show the feasibility of the proposed approach. The proposed method capable of taking more than one sensory source for learning and classification purpose. In addition, it can also perform on-line learn and classify objects incrementally which overcome the plasticity-implicity dilemma. Lastly, it can process data in real time and does not prior training that suitable to implement in natural environment.

TABLE I: Confusion matrices of the proposed method using datasets gathered from robot. The percentage values are rounded.

Correct / prediction	Chair (%)	Human (%)	Sofa (%)	Rack (%)
Chair	47 (94)	0	3 (6)	0
Human	0	48 (96)	0	2 (4)
Sofa	5 (10)	45 (90)	0	0
Rack	0	4 (4)	0	46 (96)

Future work in this subject will include an analysis of effectiveness of the value of framework parameters. In addition, we will extend the proposed method to learn from more two sensors. Lastly, we will conduct more experiments using different kind of objects for further validation.

Acknowledgment

The authors would like to acknowledge a FRGS grant (Project No.: FP069-2015A) from Ministry of Higher Education in Malaysia.

References

[1] G. Oriolo, G. Ulivi and M. Vendittelli, "Real-time map building and navigation for autonomous robots in unknown environments", Systems, Man, and Cybernetics, IEEE Transactions on, Vol. 28, no. 3, pp. 316-333, (1998).

[2] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Pltz, G.A. Fink, G. Sagerer, "Multi-modal anchoring for humanrobot interaction", Robotics and Autonomous Systems, Vol 43, Issues 23, pp. 133-147, (2003).

[3] Henrik Jacobsson, Nick Hawes, Geert-Jan Kruijff, Jeremy Wyatt "Cross-modal Content Binding in Information-Processing Architectures", Proceedings of the International Conference on Human Robot Interaction, ACM, pp. 81-88 (2008).

[4] Ciftcioglu, O. "Multiresolutional Filter Application for Spatial Information Fusion in Robot Navigation", Advances in Robotics, Automation and Control, pp. 355-372, (2008)

[5] Ross Cutler , Larry Davis "Look Who's Talking: Speaker Detection Using Video And Audio Correlation", IEEE International Conference on Multimedia and Expo, pp. 1589-1592, (2000)

[6] X. Jin, S. Gupta, A. Ray and T. Damarla, "Multimodal sensor fusion for personnel detection", Information Fusion (FUSION), Proceedings of the 14th International Conference on, pp. 1-8 (2011)

[7] N. H. Nguyen, N. M. Nasrabadi and T. D. Tran, "Robust multi-sensor classification via joint sparse representation", Information Fusion (FUSION), Proceedings of the 14th International Conference on, pp. 1-8 (2011)

[8] Raju Damarla, David Ufford, "Personnel detection using ground sensros", Proceedings of SPIE, pp. 656205-656205-10 (2007)

[9] Huaifei Xing; Fang Li; Hao Xiao; Yongjie Wang; Yuliang Liu, "Ground target detection, classification, and sensor fusion in distributedfiber seismic sensor network", Proceedings of Adv. Sensor Syst. Appl. III pp. 683015-683015-10 (2007)

[10] Y. Asfour, G. Carpenter, S. Grossberg, G. Lesher, Fusion ARTMAP: An adaptive fuzzy network for multi-channel classification", Proceedings of the International Conference on Industrial Fuzzy Control and Intelligent Systems, pp. 155-160 (1993)

[11] A. Garg, V. Pavlovic, J. Rehg, "Boosted Learning in Dynamic Bayesian Networks for Multimodal Speaker Detection", pp. 1355-1369 (2003)