Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

7-2008

A self-organizing neural model for multimedia information fusion

Luong-Dong NGUYEN

Kia-Yan WOON

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Citation

A Self-Organizing Neural Model for Multimedia Information Fusion

Luong-Dong Nguyen
School of Computer Engineering
Nanyang Technological University
Singapore
Email: dongnl@pmail.ntu.edu.sg

Kia-Yan Woon
School of Computer Engineering
Nanyang Technological University
Singapore
Email: kywoon@ntu.edu.sg

Ah-Hwee Tan
School of Computer Engineering
Nanyang Technological University
Singapore
Email: asahtan@ntu.edu.sg

Abstract—**This paper presents a self-organizing network model for the fusion of multimedia information. By synchronizing the encoding of information across multiple media channels, the neural model known as fusion Adaptive Resonance Theory (fusion ART) generates clusters that encode the associative mappings across multimedia information in a real-time and continuous manner. In addition, by incorporating a semantic category channel, fusion ART further enables multimedia information to be fused into predefined themes or semantic categories. We illustrate the fusion ART's functionalities through experiments on two multimedia data sets in the terrorist domain and show the viability of the proposed approach.**

I. Introduction

With the information on the World Wide Web (WWW) continues to grow at an exponential pace in terms of quantity as well as diversity, efficient retrieval and management of multimedia information has been a great challenge for information users. Although search engines are already available for searching multimedia information on the web, it remains a challenge to integrate information across different media content into a coherent form for further analysis.

The focus of this paper is on multimedia information fusion, namely association and integration of information across multimedia forms, including text, image, audio, and video. Our end goal is to automate the consolidation and organization of mixed media content at a semantic level and to provide the end users with a unified view of the available information across media.

For association of multimedia data, a key issue is how to measure the similarity between data of different types. For example, texts are discrete symbols while images are encoded analog signals (for example, colors). The simplest way to measure the similarity is to use the text annotations of the text and images. However, text annotations are not always available. In addition, images are different from text in nature after all. So simply converting the multimedia association problem into pure text clustering is not a sound solution. Therefore, on top of clustering based on textual features, it is of importance to exploit the underlying statistical regularities present in the visual features of images.

This paper presents an associative memory modeling approach to the problem of multimedia information fusion. Specifically, a self-organizing neural network model, known

as fusion Adaptive Resonance Theory (fusion ART) [1], is used to integrate "related" images and text segments based on their similarities in both the text and visual feature spaces. Fusion ART is a direct generalization of a family of neural architectures known as Adaptive Resonance Theory (ART) [2]–[5]. By extending ART of one pattern channel to fusion ART consisting of multiple channels, fusion ART generates clusters automatically for information presented simultaneously across the various media channels in a real-time and continuous manner. More importantly, through the use of an additional semantic category channel, fusion ART provides an integrated framework for fusing multimedia data into predefined categories when ones are available.

To illustrate the fusion ART's functionalities and performance, we conduct experiments based on two multimedia data sets in the terrorist domain, one with clusters generated automatically and the other with predefined semantic categories. The experimental results show that fusion ART is able to support both fusion paradigms with a reasonable level of performance.

The rest of this article is organized as follows. Section two reviews related work in multimedia information fusion. Section three formalizes the problem of multimedia information fusion. Section four presents the fusion ART algorithm. Section five shows how fusion ART is used for the information fusion process. Section six and seven report our experiments on the two multimedia data sets. Section eight concludes and discusses outstanding issues and future work.

II. Related Work

Multimedia information fusion can be considered as a sub-field of information fusion, a more general problem of synthesizing information across distinct media forms and modalities from disparate sources in spatial and temporal domains [6]. A well studied topic of information fusion is that of document summarization [7]–[10]. The general approach to document summarization is to perform clustering of keywords and summarizing them into appropriate templates. However, most work along this line of research focuses on summarization of text documents and web images separately, but not across different media types.

As an early work in cross media fusion, SRA [11] describes a Multimedia Fusion System (MMF) that combines an automated clustering algorithm with a summarization module to automatically group multimedia information by content and simultaneously determine concise keyword summaries of each cluster. As MMF generates clusters in an unsupervised fashion, i.e., no pre-defined user profile need be used, the system can adapt to new and changing world events with no extra effort. The main components of MMF include keyword selection, document clustering, cluster summarization, and cluster display. Unfortunately, it is not disclosed how keywords are extracted from images and no evaluation of the system has been reported.

Another approach based on the concept of active documents advertising on the WWW using adlet is described in [6]. However, there remain many issues for further research and development in this approach. [12] proposes an ontology based approach for unifying the indexing and retrieval of mixed media educational content. Due to the low tolerance for technological complexity of the target students, the experiments are only based on extremely simple query input format. More recently, a novel method for discovering image and text association [13] has been proposed. The methods of associating image and text segments could be used towards the fusion of multimedia information fusion.

III. MEDIA FUSION

We consider the problem of cross-media information fusion as the process of organizing an incoming stream of multimedia data according to their semantic relations into a set of information groupings. More formally, given a collection of multimedia data, say consisting of a set of text segments $T = \{t_1, t_2, \ldots, t_n\}$ and a set of images $I = \{i_1, i_2, \ldots, i_m\}$, the problem of multimedia fusion, as illustrated in Figure 1, can be formulated as the simultaneous learning of two mappings,

$$f^t : T \longrightarrow C \qquad (1)$$

and

$$f^i : I \longrightarrow C \qquad (2)$$

where C is the set of groupings $c_1, c_2, \ldots, c_p$. The information groupings can be generated automatically as *clusters* during the fusion process or as *categories* predefined by users as templates for the specific problem domain. Text segments and images assigned to the same information grouping are deemed to be associated and fused.

As discussed earlier, many issues and challenges exist in the problem of cross media information fusion. We highlight some of the key issues considered in our work as follows.

Firstly, traditional information fusion techniques organize information into clusters automatically generated during the process. However, a user would have no control on the groupings created to fuse the information. It remains a challenge of designing a model with the flexibility of fusing information into known themes or semantic categories when ones are
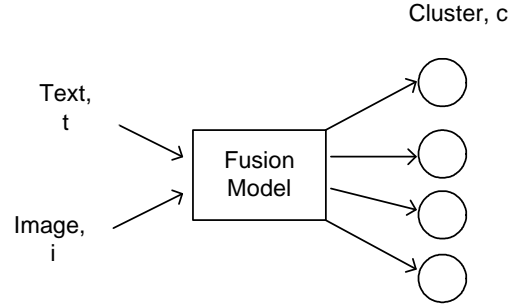

Fig. 1. The media fusion problem.

available and into automatically generated clusters at the same time.

Secondly, in a dynamic environment, various pieces of information may be available at different times in a continuous manner. For example, a pair of image and test is presented and then the image is presented with a semantic category. How do we design a model that is able to adapt according to incoming patterns and associate the information available at any given time remain a great challenge.

Our proposed solution to the multimedia fusion problem is that of a self-organizing network model, called fusion ART. We present the fusion ART algorithm and show how it can be used for media fusion in the subsequent sections.

IV. FUSION ART

Fusion ART is a natural extension of the Adaptive Resonance Theory (ART) network models from a single pattern field to multiple pattern channels. With well-founded computational principles, ART has been applied successfully to many pattern analysis, recognition, and prediction applications [14], [15]. These successful applications are of particular interest because the basic ART principles have been derived from an analysis of human and animal perceptual and cognitive information processing, and have led to behavioral and neurobiological predictions that have received significant experimental support during the last decade; see Grossberg 2003 and Raizada & Grossberg 2003 for reviews. Whereas the original ART models [18] perform unsupervised learning of recognition nodes in response to incoming input patterns, the extended neural architecture, known as fusion ART (fusion Adaptive Resonance Theory), learns multi-channel mappings simultaneously across multi-modal pattern channels in an online and incremental manner.

Fusion ART employs a multi-channel architecture (Figure 2), comprising a cluster field F_2 connected to a fixed number of (K) pattern channels or input fields through bidirectional conditional pathways. The model unifies a number of network designs developed over the past decades for a wide range of functions and applications. The generic network dynamics of fusion ART, based on fuzzy ART operations [19], is summarized as follows.

Input vectors: Let $\mathbf{I}^{ck} = (I_1^{ck}, I_2^{ck}, \ldots, I_n^{ck})$ denote the input vector, where $I_i^{ck} \in [0, 1]$ indicates the input i to channel ck.
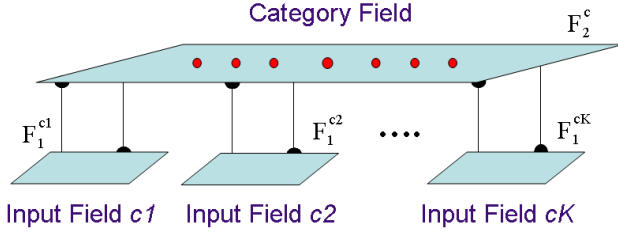
Fig. 2. The fusion ART architecture.

With complement coding, the input vector $\mathbf{I}^{ck}$ is augmented with a complement vector $\bar{\mathbf{I}}^{ck}$ such that $\bar{I}_i^{ck} = 1 - I_i^{ck}$. Complement coding is a normalization technique that has been found effective in ART systems in preventing the category proliferation problem.

Activity vectors: Let $\mathbf{x}^{ck}$ denote the F_1^{ck} activity vector for $k = 1, \ldots, K$. Let $\mathbf{y}$ denote the F_2 activity vector.

Weight vectors: Let $\mathbf{w}_j^{ck}$ denote the weight vector associated with the jth node in F_2 for learning the input patterns in F_1^{ck} for $k = 1, \ldots, K$. Initially, F_2 contains only one *uncommitted* node and its weight vectors contain all 1's.

Parameters: The fusion ART's dynamics is determined by choice parameters $\alpha^{ck} > 0$, learning rate parameters $\beta^{ck} \in [0, 1]$, contribution parameters $\gamma^{ck} \in [0, 1]$ and vigilance parameters $\rho^{ck} \in [0, 1]$ for $k = 1, \ldots, K$.

As a natural extension of ART, fusion ART responds to incoming patterns in a continuous manner. It is important to note that at any point in time, fusion ART does not require input to be present in all the pattern channels. For those channels not receiving input, the input vectors are initialized to all ones. The fusion ART pattern processing cycle comprises five key stages, namely code activation, code competition, activity readout, template matching, and template learning, as described below.

Code activation: Given the activity vectors $\mathbf{I}^{c1}, \ldots, \mathbf{I}^{cK}$, for each F_2 node j, the choice function T_j is computed as follows:

$$T_j = \sum_{k=1}^{K} \gamma^{ck} \frac{|\mathbf{I}^{ck} \wedge \mathbf{w}_j^{ck}|}{\alpha^{ck} + |\mathbf{w}_j^{ck}|}, \qquad (3)$$

where the fuzzy AND operation $\wedge$ is defined by $(\mathbf{p} \wedge \mathbf{q})_i \equiv min(p_i, q_i)$, and the norm $|.|$ is defined by $|\mathbf{p}| \equiv \sum_i p_i$ for vectors $\mathbf{p}$ and $\mathbf{q}$.

Code competition: A code competition process follows under which the F_2 node with the highest choice function value is identified. The winner is indexed at J where

$$T_J = \max\{T_j : \text{for all } F_2 \text{ node } j\}. \qquad (4)$$

When a category choice is made at node J, $y_J = 1$; and $y_j = 0$ for all $j \neq J$. This indicates a winner-take-all strategy.

Activity readout: The chosen F_2 node J performs a readout of its weight vectors to the input fields F_1^{ck} such that

$$\mathbf{x}^{ck} = \mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck}. \qquad (5)$$

Template matching: Before the activity readout is stabilized and node J can be used for learning, a template matching

process checks that the weight templates of node J are sufficiently close to their respective input patterns. Specifically, resonance occurs if for each channel k, the *match function* m_J^{ck} of the chosen node J meets its vigilance criterion:

$$m_J^{ck} = \frac{|\mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck}|}{|\mathbf{I}^{ck}|} \geq \rho^{ck}. \qquad (6)$$

If any of the vigilance constraints is violated, mismatch reset occurs in which the value of the choice function T_J is set to 0 for the duration of the input presentation. Using a *match tracking* process, at the beginning of each input presentation, the vigilance parameter ρ^{ck} in each channel ck equals a baseline vigilance $\bar{\rho}^{ck}$. When a mismatch reset occurs, the ρ^{ck} of all pattern channels are increased simultaneously until one of them is slightly larger than its corresponding match function m_J^{ck}, causing a reset. The search process then selects another F_2 node J under the revised vigilance criterion until a resonance is achieved.

Template learning: Once a resonance occurs, for each channel ck, the weight vector $\mathbf{w}_J^{ck}$ is modified by the following learning rule:

$$\mathbf{w}_J^{ck(\text{new})} = (1 - \beta^{ck})\mathbf{w}_J^{ck(\text{old})} + \beta^{ck}(\mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck(\text{old})}). \qquad (7)$$

When an uncommitted node is selected for learning, it becomes *committed* and a new uncommitted node is added to the F_2 field. Fusion ART thus expands its network architecture dynamically in response to the input patterns.

The network dynamics described above can be used to support a myriad of learning and predicting operations. We show how fusion ART can be used for media fusion in the subsequent section.

V. Image and Text Fusion

In this paper, we focus on the specific problem of fusing text and images. To this end, we adopt a fusion ART model, consisting of three pattern fields, namely a visual feature field F_1^{c1}, a textual feature field F_1^{c2} and a semantic category field F_1^{c3}, and a F_2 field encoding the association across the three pattern fields as illustrated in Figure 3.
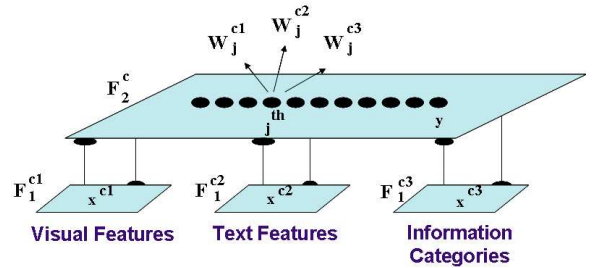

Fig. 3. A three-channel fusion ART model for fusion of text, images, and semantic categories.

For fusion ART to process the incoming information, features are extracted to describe the characteristics of each image and text segment using the available text and image analysis tools. The idea is to extract the unique image features (e.g.

color, texture) and the textual features (e.g. keywords) and to represent each text and image as a vector in the features space. The visual feature vector is encoded as $\mathbf{V}$, where V_i denotes the visual feature i and the text feature vector is encoded as $\mathbf{T}$, where T_i denotes the text feature i. The semantic categories are encoded as the input vector $\mathbf{L}$ where $L_i = 1$ if i is the semantic category; and $L_i = 0$ otherwise.

Upon input presentation, the fusion ART activity vectors are set accordingly by $\mathbf{x}^{c1} = \mathbf{V}$, $\mathbf{x}^{c2} = \mathbf{T}$, and $\mathbf{x}^{c3} = \mathbf{L}$. As mentioned, fusion ART does not require all pattern channels to be active at any given time. Only patterns presented in different channels at the same time will be encoded together. For example, when a pair of image and text is presented together, fusion ART searches for a cluster node to encode their association. Likewise, when a pair of image and semantic category is presented, a cluster will be selected for encoding their association. By synchronizing the encoding of incoming patterns across various media channels, fusion ART thus learns the association model based on the stream of incoming patterns in a continuous manner. Compared with prior systems, the multimedia fusion model described in this paper has a number of unique characteristics.

Firstly, as a generalized ART model, fusion ART learns to encode the association across multimedia information in an incremental manner. Given an incoming stream of input patterns across multiple channels, the model performs real-time search of suitable clusters for encoding the association and creating new ones when necessary. This is in contrast to prior fusion models that are typically trained in batch beforehand.

Secondly, the fusion model learns the association between multimedia feature spaces directly. Doing so allows us to discover and exploit the underlying statistical regularities in the original data directly. This is not possible in other models, that first extract keywords from images and then associate the visual keywords with text. Fusion through competitive learning is also a natural learning approach commonly found in human and animal learning paradigms.

Most importantly, fusion ART provides an integrated framework for fusion multimedia data into predefined categories (serving as templates) as well as automated generated clusters. In other words, besides that multimedia information can be associated through the clusters generated automatically, they can also be organized into a set of themes or semantic categories predefined by a user.

VI. Fusion without Themes

A. The Data Set

We build a data set consisting of images and their surrounding text retrieved from two popular news web sites www.bbc.co.uk and www.cnn.com. The images are related to the subject of "terrorist". The images' surrounding text are extracted from the web pages that contain the images. Here we define and extract five types of surrounding text, including image title, image caption, image alternative text, page title, and page metadata, which often hold images' semantics. After filtering

irrelevant images that have little relation to the subject of "terrorist" or have very short surrounding text, we obtain a data set with 159 images and the associated surrounding text.

For each pair of text and image, we extract a visual feature vector from the image and a text feature vector from the surrounding text. The visual feature vectors are based on global features, extracted from the entire image, that include the HSV domain color histogram and the Gabor texture feature. Specifically, Hue, Saturation, and Value are uniformly quantized into 18*4*4 bins, resulting in a global color feature vector of 288 dimensions. For other global texture features, Gabor filters with four scales and six orientations are applied on the entire image to generate a feature vector of 48 dimensions. The overall visual feature vectors consist of 336 attributes.

Text feature vectors are extracted from the images' surrounding text using the tf-idf scheme [20]. After extraction, each image's surrounding text is represented by a 1302-dimensional vector. The vectors are large in terms of dimensions but very sparse, which means most vectors' elements are zeros. In order to improve the efficiency of experimentation, the dimensionality of the text feature vectors is reduced using the Singular Value Decomposition (SVD) algorithm. SVD helps to reduce the text feature vectors' dimension without losing the relative distance and information between the data vectors. It means that the reduced vectors can be used for clustering or measuring similarity between the text vectors, as in the original tf-idf encoding. After dimensionality reduction, each text feature vector consists of 146 attributes.

B. Experiments

The extracted visual and text feature vectors are used as input to the fusion ART system. During the training phase, all the visual feature vectors and text feature vectors are presented in pairs to the image and text channels respectively. The semantic category channel that is not active here receives input vectors containing all ones.

For all the experiments, the choice parameters α^{ck} of fusion ART are fixed at 0.1 and the learning rates β^{ck} are set to 1 (fast learning). The contribution parameters γ^{ck} are set to 0.5 for both the image and text channels. Nonetheless, to see how the result varies with different parameter setting in fusion ART, we alter the vigilance parameter ρ^{c1} and ρ^{c2} gradually from 0.1 to 0.99 in different running of the experiments.

As shown in Table I, fusion ART creates a varying number of clusters in response to different settings of the vigilance parameter value. To evaluate the quality of the clusters created, we test the fusion model by presenting only the images' visual feature vectors. Upon identifying the cluster encoding an image, we extract the semantics of the cluster based on its cluster weight template vector in the text feature space. The semantics (cluster keywords) retrieved is then compared with the original text feature vector representing the surrounding text of the image for verifying its consistency. Specifically, a *precision* score is computed by counting the number of cluster keywords found in the original surrounding text over

the number of all cluster keywords while a *recall* score is calculated by that number of common keywords over the number of keywords in the surrounding text.

Table I summarizes the number of clusters the precision score, and the recall score, with respect to the setting of the vigilance parameter value. We see that the precision of cluster semantics for each images rises as the vigilance value increases. Along with that, however, is the climbing of the number of clusters created. Nevertheless, the recall rate does not improve when precision reaches a high value, which accords to the normal precision-recall curve.

TABLE I

THE NUMBER OF CLUSTERS, PRECISION AND RECALL OF FUSION ART WITH RESPECT TO VIGILANCE VALUE.

ρ^{ck}	0.1	0.5	0.75	0.9	0.95	0.99
Cluster	73	74	87	123	156	159
Precision	0.51	0.55	0.99	0.99	0.99	0.99
Recall	0.23	0.27	0.22	0.28	0.44	0.45

Image	Text	Cluster Semantics
	story.australia.emergency police .com - Australia agrees to amend emergency power laws - August 23, 2000 The Australian government has agreed to amend proposed legislation that would allow troops to respond to Olympic terrorist threats.	terrorist(0.82) polic(2.18) threat (2.50) troop(3.68) ha(2.50) stori (1.85) australia(6.25) emergen (6.92) agre(10.14) amend (10.14) power(3.97) law(3.68) august(4.38) australian(3.97) govern(3.46) propos(4.38) legisl (5.07) respond(5.07) olympic (2.99)
	story.91900.pool Ground Zero .com - 'Ground zero' tops 2001 word list - December 27, 2001 In the waning days of a year torn by terrorism and war, a list of 2001's top 10 words, compiled by yourDictionary.com, predominantly reflects the aftermath of September 11.	dai(3.28) terror(1.74) reflect (3.97) septemb(2.67) stori(1.85) pool(5.07) ground(10.14) top (6.25) word(10.14) list(6.25) decemb(4.38) wane(5.07) year (2.87) torn(5.07) war(3.12) compil(5.07) yourdictionary.com (5.07) aftermath(4.38)
	story.ressam.testifies Ressam .com - Terrorist: Y2K defendant wasn't told plot details - July 6, 2001 Convicted Algerian terrorist Ahmed Ressam testified Friday that Mokhtar Haouari, on trial for allegedly aiding Ressam's foiled plot to detonate a bomb at Los Angeles International Airport, was not told the details of the plan.	terrorist(1.64) juli(3.12) wa (5.01) bomb(1.93) intern(2.58) stori(1.85) convict(3.28) mokhtar(4.38) haouari(4.38) defend(4.38) plot(6.92) told (7.94) fridai(3.97) ressam (17.50) testifi(10.14) detail (10.14) algerian(5.07) ahm(4.38) trial(3.46) aid(5.07) foil(5.07) deton(4.38) lo(5.07) angel(5.07) airport(4.38) plan(3.46)

Fig. 4. Sample images and the associated semantics.

VII. FUSION INTO CATEGORIES

A. The Data Set

A terrorist domain web page collection is built in-house, containing 472 images related to terrorist attacks, downloaded from the CNN and BBC news web sites. For each image, we select a text paragraph, which is semantically related to the image, from the web page containing the image. In addition, we manually categorize the identified image-text pairs into eight predefined semantic categories, i.e. *Anti-Terror, Attack Detail, Ceremony, Government Response, Rescue, Terrorist Suspect, Victim*, and *Others*. The detailed data preprocessing methods are described as follows.

1) Textual Feature Extraction: We treat each text paragraph as a text segment. In our work, we use a text mining toolkit, known as Text2Knowleddge (T2K)

(http://alg.ncsa.uiuc.edu/do/tools/t2k), for preprocessing the text segments. The preprocessing steps include text tokenization, part-of-speech tagging, stop word filtering, stemming, removing unwanted terms (retaining only nouns, verbs and adjectives), and generating the textual feature vectors where each dimension corresponds to a remaining term after the preprocessing.

For calculating the term weights of the textual feature vectors, we use a model, named TF-ITSF (term frequency and inverted text segment frequency), similar to the traditional TF-IDF model. For a text segment ts in a web document d, we use the following equation to weight a term w in ts:

$$w^d(ts) = tf(ts, w) \cdot \log \frac{N^d}{tsf^d(w)}, \qquad (8)$$

where $tf(ts, w)$ denotes the frequency of w in the text segment ts, N^d is the total number of text segments in the web document d, and $tsf^d(w)$ is the text segment frequency of term w in d. Here, we use the text segment frequency for measuring the importance of a term for a web document.

After a term weight vector $(w_1^d(ts), w_2^d(ts), ..., w_n^d(ts))$ is extracted, L1-normalization is applied for normalizing the term weights into a range of $[0, 1]$:

$$v_{ts} = \frac{(w_1^d(ts), w_2^d(ts), ..., w_n^d(ts))}{max\{w_i^d(ts)\}_{i=1...n}}, \qquad (9)$$

where n is the number of textual features (i.e. terms).

2) Visual Feature Extraction: Each image is first segmented into 10×10 rectangular regions. For each region, we extract a visual feature vector, consisting of six color features and 60 Gabor texture features, which have been proven to be useful in many applications. The color features are the means and variances of the RGB color spaces. The texture features are extracted by calculating the means and variations of the Gabor filtered image regions on six orientations at five different scales. After that, all image regions are clustered using the k-means algorithm with k=500. The generated clusters, called *visterms* represented by $\{vt_1, vt_2, ..., vt_{500}\}$, are treated as a vocabulary for the images. An image is described by a visterm vt_j if it contains a region belonging to the jth cluster. For the terrorist domain data set, the visterm vocabulary is enriched with a high-level semantic feature, extracted by a face detection model of OpenCV. In total, a visterm vector of 501 dimensions is extracted for each image. The weight of each dimension is the corresponding visterm frequency normalized with the use of L1-normalization.

B. Experiments

During training, each of the images together with the associated text segments and semantic categories is presented to fusion ART for encoding. The system is trained until it is stabilized, such that there is no mismatch reset for each data sample. Then, testing is conducted through presenting the image feature vector or the text feature vector, without the semantic category.

In all the experiments, the choice parameters α^{ck} are fixed at 0.001. Fast learning is used with $\beta^{ck} = 1$. The contribution parameters γ^{ck} are dynamically set to be the average of the active channels except for the semantic category channel, which is fixed at 0.0. The baseline vigilance parameters $\bar{\rho}^{ck}$ of the active channels are set to 0.0, and the vigilance of the semantic category channel is set to 1.0 during training; they are all set to 0.0 during testing.

With the presence of semantic category information, the evaluation of the fusion performance is thus based on the semantic categories predicted for each image and text pattern. For each test pattern, the semantic category encoded by the cluster identified is compared with its associated semantic category. When an image or text segment is grouped into the cluster with the same semantic category, it is deemed to be correctly fused.

We first evaluate the performance of the system by using five-fold cross-validation in the experiments. In this approach, each record is used the same number of times for training and exactly once for testing. We partition the data into five equal-sized subsets. During each run, one of the partitions is chosen for testing, while the rest of them are used for training. This procedure repeated five times so that each partition is used for testing exactly once. The total accuracy is computed by averaging the accuracy across all five runs.

We also experiment with the leave-one-out paradigm, under which the experiments are repeated for N times, one for each data sample as the only sample in the test set. It has the advantage of utilizing almost all the available data for training and still using each of the data sample in turns as the test sample.

TABLE II
THE FUSION PERFORMANCE OF FUSION ART IN THE FIVE-FOLD CROSS
VALIDATION AND LEAVE-ONE-OUT EXPERIMENTS.

	Image	Text	Overall
five-fold	42.3	44.7	43.5
leave-one-out	41.3	44.1	42.7

The experimental results based on five-fold cross-validation and leave-one-out paradigms are summarized in Table II. The results indicate that fusion ART can only fuse around 40% of the unseen images and text segments into the correct semantic categories. Even leave-one-out, in this case, does not help to improve the performance. The small number of (472) data points in this data set, compared with the high dimensionality of the input vectors ($651 * 2 + 787 * 2 + 8 = 2097$), may have contributed to the less-than- satisfactory performance.

In view that poor generalization in the visual and text feature spaces may have affected the fusion performance, we further conduct experiments by generating synthetic patterns based on the original data patterns for testing the fusion model. The synthetic data are generated by adding random disturbance noise to the original data patterns with a probability from 10% to 50%.

Figure 5 shows the fusion ART's performance on the synthetic data set. we see that in this case the fusion model has shown a greater resilience to the variations in the image and text patterns and is able to provide a reasonably good performance up to a noise level of 40%.
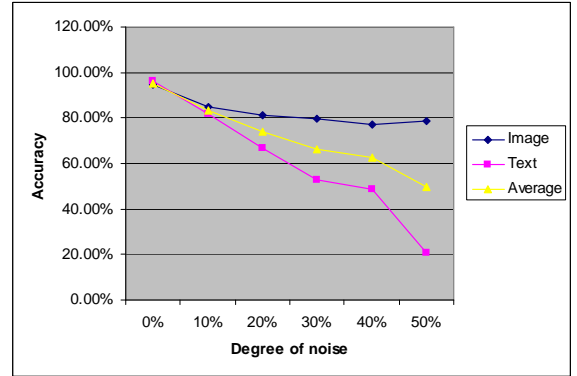

Fig. 5. The performance of fusion ART on the synthetic data set.

VIII. CONCLUSIONS

We have formulated multimedia information fusion as the problem of simultaneously learning the mappings across text, image, and semantic category spaces. To this end, we propose a self-organizing computational model, known as fusion ART, as a plausible solution towards cross-media information fusion. As fusion ART inherits the ART properties, including self-organizing, self-stabilizing, and fast incremental learning, the proposed model has the advantages and flexibility of learning the association model across multiple channels in real-time. More importantly, the fusion framework integrates the two distinct approaches of fusing into clusters generated automatically and fusing into *categories* predefined as templates. Our experimental results have shown that the proposed approach is viable and is promising for the purpose of information fusion.

Moving ahead, we plan to extend our experiments to larger data sets. Although fusion ART in principle can support a variety of fusion paradigms, this paper has only investigated two specific cases, one without the use of semantic categories and the other with the use of semantic categories. Our future study will include more complex scenarios, involving data patterns with and without semantic categories in the same experiments.

While we have no knowledge of a similar system with functions comparable to our fusion model, we wish to compare the performance of our model with some existing models, perhaps in selected aspects wherever possible. As the current results still show a relatively poor generalization performance in the text and visual feature spaces, we need to explore and incorporate better text and image analysis techniques, probably with the use of mid-level visual concepts, to narrow the semantic gap between text and images.

REFERENCES

[1] G. A. Carpenter, A. H. Tan, and S. Grossberg, "Intelligence through interaction: Towards a unified theory for learning," in *Proceedings of*

the International Symposium on Neural Networks (ISNN) 2007, LNCS 4491, vol. I.   Nanjing, China: D. Liu et al. (Eds.), pp. 1098–1107.

[2] G. A. Carpenter and S. Grossberg, Eds., *Pattern Recognition by Self-Organizing Neural Networks*.   Cambridge, MA: MIT Press, 1991.

[3] G. Carpenter and S. Grossberg, "Adaptive Rresonance Theory," in *The Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. Cambridge, MA: MIT Press, 2003, pp. 87–90.

[4] S. Grossberg, "Adaptive pattern recognition and universal recoding, I: Parallel development and coding of neural feature detectors," *Biological Cybernetics*, vol. 23, pp. 121–134, 1976.

[5] ——, "Adaptive pattern recognition and universal recoding, II: Feedback, expectation, olfaction, and illusion," *Biological Cybernetics*, vol. 23, pp. 187–202, 1976.

[6] S.-K. Chang and T. Znati, "Adlet: an active document abstraction for multimedia information fusion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 1, pp. 112–123, 2001.

[7] D. R. Radev, "A common theory of information fusion from multiple text sources step one: cross-document structure," in *Proceedings of the 1st SIGdial workshop on Discourse and dialogue*.   Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 74–83.

[8] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*.   New York, NY, USA: ACM Press, 2001, pp. 269–274.

[9] R. Barzilay., "Information fusion for multidocument summarization: paraphrasing and generation," Ph.D. dissertation, Columbia University, 2003.

[10] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*.   New York, NY, USA: ACM Press, 2005, pp. 112–121.

[11] S. International, "Multilingual nametag(tm) multilingual internet surveillance system multimedia fusion system," in *Proceedings of the Fifth Conference on Applied Natural Language Processing*.   New York, NY, USA: ACM Press, 1997, pp. 31–32.

[12] J. Ng, K. Rajaraman, and E. Altman, "Mining emergent structures from mixed media for content retrieval," in *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*.   New York, NY, USA: ACM Press, 2004, pp. 316–319.

[13] T. Jiang and A.-H. Tan, "Discovering image-text associations for cross-media web information fusion," in *PKDD '06: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*.   Springer Berlin / Heidelberg, 2006, pp. 561–568.

[14] R. O. Duda, P. Hart, and D. Stock, Eds., *Pattern Classification (2nd edition)*.   New York: John Wiley, Section 10.11.2, 2001.

[15] D. S. Levine, Ed., *Introduction to Neural and Cognitive Modeling*.   New Jersey: Lawrence Erlbaum Associates, Chapter 6, 2000.

[16] S. Grossberg, "How does the cerebral cortex work? development, learning, attention, and 3d vision by laminar circuits of visual cortex," *Behavioral and Cognitive Neuroscience Reviews*, vol. 2, pp. 47–76, 2003.

[17] R. Raizada and S. Grossberg, "Towards a theory of the laminar architecture of cerebral cortes: Computational clues from the visual system," *Cerebral Cortex*, vol. 13, pp. 200–213, 2003.

[18] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, pp. 54–115, June 1987.

[19] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759–771, 1991.

[20] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Ithaca, NY, USA, Tech. Rep., 1987.