Text-to-Image Cross-Modal Retrieval of Magazine Articles Based on Higher-order Pattern Recall by Hypernetworks

Jung-Woo Ha, Byoung-Hee Kim, Hyun-Woo Kim, Woongchang Yoon, Jae-Hong Eom,
and Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
Seoul, Korea
Email: { jwha, bhkim, hwkim, wcyoon, jheom, btzhang } @bi.snu.ac.kr

***Abstract*—As the amount of multimedia data grows larger and multi-modal information is widespread, requirements for methods are increasing which can be used to analyze composite information and retrieve related items of one modality based on another modality. For contents-based retrieval, models that incorporate multiple modalities concurrently are recognized as a mandatory approach. In this study, we propose a method to reconstruct and retrieve images based on text-to-image cross-modal recall by hypernetworks. In our method, a probabilistic graphical model called hypernetwork learns the pattern of relations between text keywords and images and images related to given keywords are reconstructed based on the learned relation patterns. Then, original images are retrieved based on similarities which are evaluated between reconstructed images and original ones. Experimental results on Korean magazine articles show that when text keywords are given as a query, the original images related to the keywords are retrieved. In addition, the results show when both image patches and text keywords are given, the images are reconstructed more precisely.**

***Index Terms*— Image reconstruction, pattern matching, pattern recognition, text processing**

I. Introduction

CROSS-MODAL learning means a methodology based on transition from one modality to another one. That is, we can refer it to cross-modal that given auditory information, the data is converted to visual information with identical or similar contents. Cross-modal data retrieval and generation method is important with respect to both application and cognitive science. First, cross-modal techniques are applied to multimedia data mining [1] [2]. For example, 'text-to-image' can be applied to contents based image search and 'image-to-text' can be used to auto-tagging. Also, with respect to cognitive science, cross-modal retrieval is a trial to imitate the perception and cognition related to multi-modality in the brain [3] [4].

The contents in articles have their own subjects and the subjects are usually represented with more than one modality such as sentences and images. Text words and images related to the subject are used in the articles so as the content represent its subject with consistency. Therefore we can assume that there exist relations between words and images used in an article. Moreover, we can try to find a mapping method to convert one modality to the other based on the relation information. That is, when linguistic keywords are given as a query, we can acquire images related to the given keywords.

Since keywords and images are represented with a large number of features, the cross-modal retrieval requires a generative model which can represent relations among high-dimensional features. In this study, we use the hypernetwork model [5] as a generative model. The hypernetwork is a weighted hypergraph where evolutionary methods are embedded as learning strategies.

In this study, we make use of Korean magazine articles as experimental data and we generate images based on text-to-image cross-modal reconstruction using hypernetworks. In addition, we introduce a similarity measure to evaluate the difference between reconstructed images and original images to retrieve the most similar images. Experimental results show when the text keywords are given, the images are reconstructed and we can retrieve the original images related to keywords. Also, when partial images are given as query together, the reconstructed images are more recognizable.

The rest of this paper is organized as follows. In Section II, backgrounds for this study are summarized and we present the method of cross-modal text-to-image reconstruction with hypernetworks in Section III. In Section IV, experimental results are presented. Finally, we conclude with summary and future works in Section V.

II. Backgrounds

A. *Cross-modal Learning*

The term of cross-modal was from cross-modal perception, cross-modal integration, or cross-modal plasticity in brain and

This work was supported in part by IITA through the IT R&D program (IITA-2009-A1100-0901-1639, MARS), in part by KRF grant funded by Korean Government (MOEHRD) (KRF-2008-314-D00377), and in part by the BK21-IT program funded by Korean Government (MEST).

cognitive science. Since sensory information is commonly encoded in more than one sense modality, notably sight and sound, cross-modal perceptions and integrations occurs in brain [4]. Beyond brain science, cross-modal methodology has been adopted in analyzing multimodal information processing. D. Li *et al.* suggested cross-modal association based factor analysis method as alternatives to Latent Semantic Indexing (LSI) and Canonical Correlation Analysis (CCA) [1]. Ferecatu *et al.* showed that the joint use of visual features and concept-based features with relevance feedback scheme improves the quality of the cross-modal image retrieval [2].

B. *Hypernetworks*

Hypernetwork is a weighted hypergraphs and it represents combinatorial spaces of features with various orders. Hyperedges in hypernetwork models are the combinations of more than one vertex which is a pair of feature and its value. Formal definition of hypernetworks is presented in [1]. The hypernetwork consists of hyperedges which can connect to more than two vertices. Since a hyperedge is a combination of features, the hypernetwork represents the combinatorial spaces of features. Moreover, hyperedges can represent various amounts of information according to their order, the number of features in a hyperedge. Lower-order hyperedges have general information and higher-order ones have specific representation. Since the hypernetwork is a set of partial information, it can conduct recall of relevant information for given queries. Since the hypernetwork are proposed by Zhang *et al.* [6], the model has been applied to various domains such as pattern recognition [6] [7], bioinformatics [8] [9], and multimedia data mining [5].

III. Text-to-image Retrieval with Hypernetworks

The goal of this study is to retrieve images related to text keywords by image completion based on text-to-image cross-modal inferences. Therefore, the hypernetwork is built with two modalities concurrently. The suggested method for text-to-image retrieval consists of three steps such as building of hypernetwork, reconstruction of images with query, and retrieval of original images with generated images based on the measured similarity. Fig. 1 shows the entire process on text-to-image completion with a hypernetwork trained with a set of pairs of text and image.

A. *Building Hypernetworks*

The detailed process of building hypernetworks is explained in [5] and [10]. In this study, hyperedges are composed of two parts such as text part and image one. Text part is generated from a data sample by sampling the keywords with non-zero occurrence frequency in the sample. Sampling process of image part is identical to the one of previous hypernetworks. Unlike with the previous hypernetwork model, there is no explicit learning procedure in building a hypernetwork. The model is generated based on sampling only.

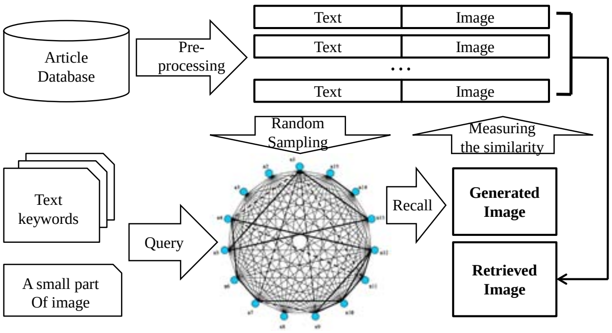

Fig 1. Flow of generation and retrieval of images with text keyword queries based on text-to-image cross modal recall by hyperenetworks. Hypernetworks are generated by sampling from preprocessed articles which are represented as a joint of text and image attributes. When text keywords and a partial image are given as queries, an image related to given queries is reconstructed by recall based inference of hypernetwork. For evaluation, we can measure the similarity between the generated image and the original image. For retrieval, we measure similarities between the generated image and images in a database and choose those with highest similarity values.

B. *Image Reconstruction*

Image reconstruction by hypernetworks is conducted by iterative matching and comparison of input queries with hyperedges. If text and image query are consistent with a hyperedge, the image part of the hyperedge participate in determining the value of pixels. Fig. 2 explains the way to reconstruct images with query. In addition, the algorithm for image reconstruction is presented in Fig. 3.

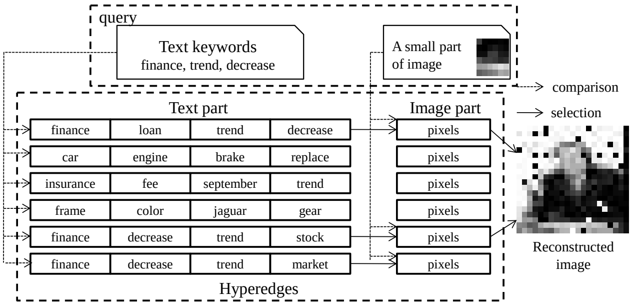
query
Text keywords
finance, trend, decrease
A small part of image
comparison
selection
Text part
Image part
finance | loan | trend | decrease | pixels
car | engine | brake | replace | pixels
insurance | fee | september | trend | pixels
frame | color | jaguar | gear | pixels
finance | decrease | trend | stock | pixels
finance | decrease | trend | market | pixels
Hyperedges
Reconstructed image

Fig. 2. Image reconstruction by comparison of query with hyperedges. When query with text keywords and partial images are given, the values of each modality are compared with the ones in hyperedges. Pixel values of a reconstructed image are determined with image part of matched hyperedges.

C. *Similarity Measure for Gray Images*

To retrieve images related to query, the similarity should be evaluated between an original image and a generated image. In this study, we introduce the similarity *s* between images as following:

$$s = \sum_{i=1}^{M} (g_{ri} - g_{oi})^2 . \qquad (1)$$

, where g_{ri} and g_{oi} is the gray value of the *i*-th pixel in a reconstructed image and an original image with *M* pixels respectively.

```
$e_{ti}$ : text order of the $i$-th hyperedges
$e_{vi}$ : image order of the $i$-th hyperedges
$e_{vil}$ : the $l$-th vertex of $e_{vi}$
$q_t$ : keywords in query, $q_v$ : image pixels in query
$g\,[M]$: reconstructed image with $M$ pixels
$cnt\,[M][G]$: array of gray value of image pixels with gray scale $G$
$E$ : the set of hyperedges
For $i \leftarrow 1$ to $|E|$
    $e_i \leftarrow$ the $i$-th element of $E$
    $flag \leftarrow$ true
    For $j \leftarrow 1$ to $|\,q_t\,|$
        $q_{tj} \leftarrow$ the $j$-th element of $q_t$
        If $q_{tj}$ does not exist in $e_{ti}$ Then $flag \leftarrow$ false
        End If
    End For
    If $flag$ is true Then
        For $j \leftarrow 1$ to $|\,q_v\,|$
            comparison $q_v$ with $e_{vi}$
        End For
        If $q_v$ matches $e_{vi}$ then
            For $l \leftarrow 1$ to $|\,e_{vi}\,|$
                $cnt[e_{vil}]$[gray value of $e_{vil}$]++
            End For
        End If
    End If
End For
For $i \leftarrow 1$ to $M$
    g[i] $\leftarrow \arg\max_{c} cnt[i][c]$
End For
```

Fig. 3. Algorithm of reconstructing image with a query

IV. Experimental Results

A. Data and Experimental Setup

In this study, we make use a set of Korean magazine articles whose categories are car-tuning and business respectively. The categories are selected because it may be probable that there exist the difference in frequently used keywords. Articles consist of text part and image part. Text part is made up of some sentences while image part contains several images in an article. To build hypernetworks, we preprocess articles in the following steps: sentences are converted to keyword frequencies matrix after stemming and images are renormalized to 20 by 20 pixels with 8bit gray scale as shown in Fig 4. As a result, from magazine articles, we build a set of bimodal records with keyword frequencies and pixel values. Since an article contains several images, there exist several pairs that share a single keyword configuration. Table 1 shows the summary of our experimental data.

For the hypernetwork, sampling rate, the number of hyperedge sampling for one record, is fixed as 20.

Table 1. The information of experimental data

Data size	Keywords	Pixels
1420	4612	400

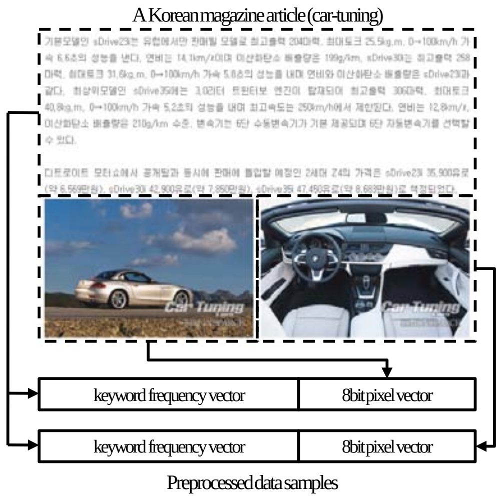

Fig 4. Data representation by preprocessing the articles. Since an article usually has several images, identical keyword frequency vector can appear in several data.

B. Experimental Results

Table 2 shows the accuracies of image retrieval with reconstructed images when text keywords are given only as queries. In Table 2, successful retrievals mean the number that an image in an article with keywords given as query is in candidates of retrieved images. Therefore, successful retrievals can be regarded as a measure of successful keywords-related image retrievals. The probability of successful retrieval increases from 50.1% to 75.2% as we increase the size of candidates from 3 to 20. Fig 5 shows reconstructed images and retrieved images for two types of text queries including business and car. Although reconstructed images are not vivid, retrieved images show misty outlines and shapes that are related to input text queries.

Table 2. Reconstruction accuracy of generated images with text keywords query

	The size of retrieved candidates			
	3	5	10	20
Successful retrievals	712	813.4	916	1067.6
Percentage (%)	50.1	59.2	64.5	75.2
Standard deviation	27.0	19.3	11.9	25.6

The number of percentage is the ratio of successful retrievals to 1420 data. The numbers are averaged for 10 times experiments.

Fig. 6 shows reconstructed images and retrieved images with text and additional image patch as a query. Compared to Fig 5, it is much easier to recognize the reconstructed images than ones without partial image query. Moreover, there are few differences between reconstructed images and retrieved images since the given partial images are from the best retrieved images in both cases of subjects. Therefore we can conjecture that partial images as well as text keywords queries play important roles on reconstructing the image. These results explain the need of multi-modal query to improve the performance of image search and retrieval.

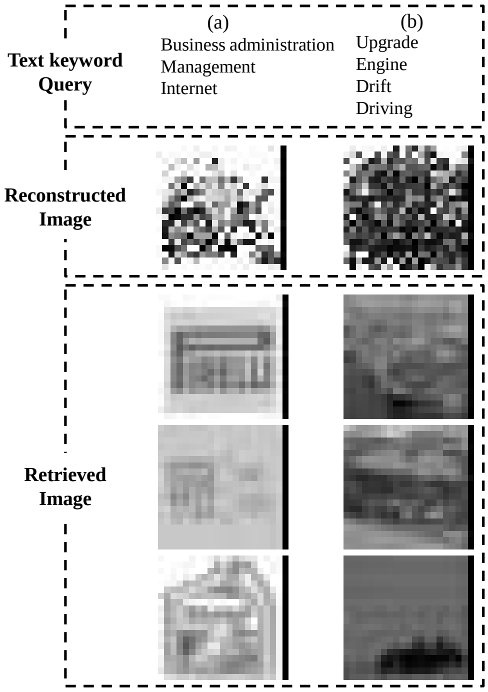

Fig. 5. Reconstructed and retrieved images with text query of two subjects. Text order and Image order of hyperedges are 100 and 40 respectively. Retrieved images are three images which are the most similar to the reconstructed image. Null pixels are represented with black pixels in reconstructed images.

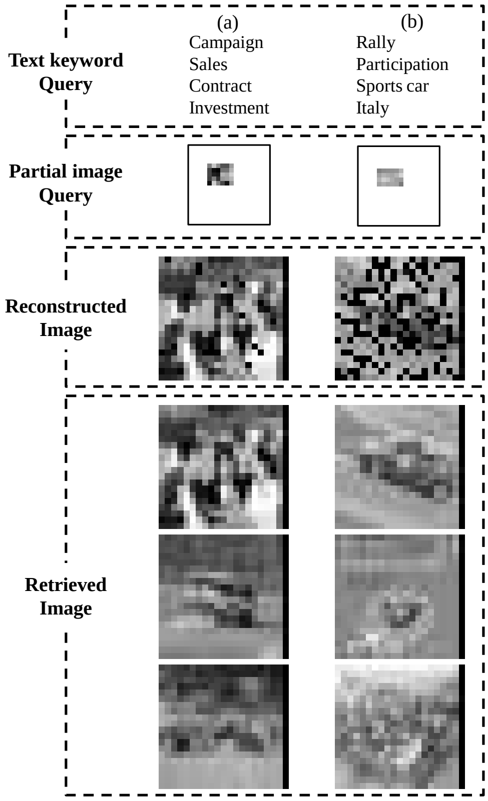

Fig. 6. Reconstructed and retrieved images with text and image query of two subjects. Text order and Image order of hyperedges are 100 and 40 respectively. Partial images are 5 by 5 pixels in specific location. Null pixels are represented with black pixels in reconstructed images.

V. Conclusions and Future works

We suggest a method of text-to-image reconstruction and retrieval based on cross-modal recall by hypernetworks. Experimental results show when a text query is given, the images are reconstructed and keywords-related images are retrieved based on the proposed similarity measure. In addition, when the partial image is given as query, the accuracy of reconstruction is improved dramatically.

In this study, there is no explicit learning process in hypernetworks. To improve the performance of reconstruction, we will adopt the proper learning mechanism to present hypernetworks. In addition, it is required to expand the scale of pixels and the number of used keywords. Considering aspects of applications, these results can be applied to multi-modal queries to improve the performance and accuracy of contents-based image searching.

Acknowledgment

Jung-Woo Ha thanks Min-Oh Heo, Je-Keun Rhee and Young-Kil Ko for discussion and data preparation.

References

[1] D. Li, N. Dimitrova, M. Li, and K. Sethi, "Multimedia content processing through cross-modal association", Proceedings of *the 11th ACM International Conference on Multimedia,* pp. 604~611, 2003.

[2] M. Ferecatu, N. Boujemaa, and M. Crucianu, "Semantic interactive image retrieval combining visual and conceptual content description," *Multimedia Systems*, Vol.13, pp. 309-322, 2008.

[3] M. M. Cohen and D. Massaro, "Synthesis of visible speech", *Behaviour Research Methods, Instruments and Computers*, Vol. 22, No. 2, pp. 260-263, 1990.

[4] J. M. Fuster, M. Bodner, and J. K. Kroger, "Cross-modal and cross-temporal association in neurons of frontal cortex," *Nature,* Vol. 405, pp. 347~351, 2000.

[5] B. -T. Zhang, "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory," *IEEE Computational Intelligence Magazine*, Vol 3, No. 3, pp. 49-63, 2008.

[6] B. -T. Zhang and H. -Y. Jang, "Molecular programming: evolving genetic programs in a test tube," *The Genetic and Evolutionary Computation Conference* (*GECCO* 2005), vol. 2, pp. 1761-1768, 2005.

[7] J. -K. Kim and B. -T. Zhang, "Evolving hypernetworks for pattern classification," *IEEE Congress on Evolutionary Computation* (*CEC 2007*), pp.1856~1862, 2007.

[8] J. -W. Ha, J. -H. Eom, S. -C. Kim and B. -T. Zhang, "Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis," *Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation* (2007), pp. 2709-2716, 2007.

[9] C. -H. Park, S. -J. Kim, S. Kim, D. -Y. Cho and B. -T. Zhang, "Finding cancer-related gene combinations using a molecular evolutionary algorithm," *IEEE 7th international conference on Bioinformatics & BioEngineering* (*BIBE 2007*), pp. 158-163, 2007.

[10] J.-W. Ha, J. H. Jang, D. -H. Kang, W. H. Jung, J. S. Kwon, and B. -T. Zhang, "Gender classification with cortical thickness measurement", IEEE International Conference on Fuzzy Systems, 2009 (accepted).