2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

Text-to-Image Retrieval Based on Incremental Association via Multimodal Hypernetworks

Jung-Woo Ha
School of Comp. Sci. & Eng.
Seoul National University
Seoul, Korea
jwha@bi.snu.ac.kr

Beom-Jin Lee
School of Comp. Sci. & Eng.
Seoul National University
Seoul, Korea
bjlee@bi.snu.ac.kr

Byoung-Tak Zhang
School of Comp. Sci. & Eng.
Seoul National University
Seoul, Korea
btzhang@bi.snu.ac.kr

***Abstract*—Text-to-image retrieval is to retrieve the images associated with the textual queries. A text-to-image retrieval model requires an incremental learning method for its practical use since the multimodal data grow up dramatically. Here we propose an incremental text-to-image retrieval method using a multimodal association model. The association model is based on a hypernetwork (HN) where a vertex corresponds to a textual word or a visual patch and a hyperedge represents a higher-order multimodal association. Using the HN incrementally learned by a sequential Bayesian sampling, in the multimodal hypernetwork-based text-to-image retrieval, a given text query is crossmodally expanded to the visual query and then similar images are retrieved to the expanded visual query. We evaluated the proposed method using 3,000 images with textual description from Flickr.com. The experimental results present that the proposed method achieves very competitive retrieval performances compared to a baseline method. Moreover, we demonstrate that our method provides robust text-to-image retrieval results for the increasing data.**

Keywords-text-to-image retrieval; incremetnal learning; hypernetworks; textual-visual association;

I. INTRODUCTION

Text-to-image (T2I) retrieval [1] involves getting images from text queries and it has been actively studied because of its diverse applications including content-based image retrieval [1-2] and article or video searching [3]. Various approaches have been applied to associate textual and visual modalities for T2I retrieval. Feng *et al*. used multiple Bernoulli model for image retrieval [3] and Zhang *et al*. applied Bayesian framework to learning latent semantic models for T2I retrieval [4]. Li *et al*. proposed multi-instance learning method using loosely labeled images for image retrieval [5]. Because the sizes of text and visual vocabularies increase continuously due to the growth of multimodal data, however, T2I retrieval models should facilitate to deal with these increasing data for their practical usages, thus requiring an incremental learning method. However, most of the models for T2I retrieval assume the vocabulary sizes are fixed like bag of words and they use batch approach-based learning methods [3-5]. Therefore, these models are not easy to be practically applied to data-increasing environments since this fixed vocabulary size has a limitation in representing new multimodal image data including unobserved textual words and new visual features.

Here we propose a T2I retrieval method based on a textual-visual association model which can efficiently treat increasing data. For the multimodal association, we use a hypernetwork (HN), which is a higher-order probabilistic graphical model using hypergraph structure [6]. In HNs, a vertex denotes a textual word or a visual feature and a hyperedge represents a multimodal subpattern of textual-visual data by connecting more than two vertices. Therefore, HNs can represent the higher-order associative relationships among textual and visual modalities. Learning HNs consists of generating hyperedges which reflect the relationships embodied in the given data and updating the weights of the hyperedges. This learning process is formulated by a sequential Bayesian framework. Whenever an image with a description is observed, new hyperedges are generated from the image by random selection-based evolutionary method and they are added into the HN. Then, the weights of the hyperedges of the model are updated by predicting and correcting the observed image, with comparing the subpattern of each hyperedge with the image and its description. Therefore, the weights are to reflect the associative strength of the hyperedge for predicting the images and descriptions. Especially, HNs can incrementally learn the increasing data by simply adding unobserved textual words and visual features involved in new data as new vertices into the model and generating hyperedges including them. When a text query is given, the query is expanded to a visual query consisting of visual patches associated with the textual query by the learned HNs. By measuring the similarity between the expanded visual query and stored images, images are retrieved semantically related to the given text query. Fig. 1 describes the proposed framework of T2I retrieval using HN models.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0005643, Videome) and in part by the BK21-IT program funded by MEST.

978-1-4673-1714-6/12/$31.00 ©2012 IEEE

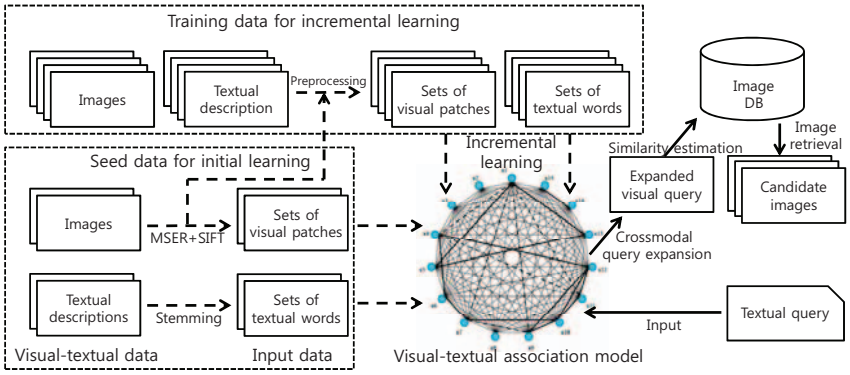

Figure 1. Overall flow of text-to-image retrieval using incrementally learned multimodal hypernetworks.

We apply the HN-based T2I retrieval method to retrieve about 3,000 images from Flickr.com for evaluation as shown in Fig. 2. In this study, several visual patches are extracted from an image by maximally stable external regions (MSER) [7] and the extracted patches are represented with 500 scale-invariant feature transform (SIFT) features [8]. Also, the image descriptions are represented with about 2,800 textual words. The experimental results present that our method shows good retrieval performances over a baseline method based on the co-occurrence of textual words and visual features. Moreover, we demonstrate that the proposed method provides robust T2I retrieval results with reflecting the increase of the data.

II. Multimodal Hypernetworks

A. Hypentworks as a Multimodal Association Model

A hypernetwork (HN) is a memory-based higher-order probabilistic graphical model using hypergraph structure as the model representation [6]. A hypergraph is an extended graph where an edge, i.e. hyperedge, can connect more than two vertices concurrently. In HNs, a vertex denotes a data variable and a hyperedge represents an arbitrary relationship among its vertices. Also, the weight of a hyperedge is defined to a function reflecting the strength of its association. Therefore, the HN can be considered as a model representing higher-order associative relationships among data variables. An HN is formally defined to H=(V, E, W) where V, E, and W denote a vertex set, a hyperedge set, and a weight set, respectively. HNs have been successfully applied in various problems including pattern recognition [9], bioinformatics [10], and multimodal data mining [11].

An HN can be used as a multimodal association model by defining vertices to textual words or visual features and hyperedges to associative relationships among textual and visual features [11]. The advantages of HNs as a multimodal association model are summarized as follows:

i) Representation of multimodal association based on higher-order relationships among textual and visual features,

ii) Robust and flexible model structure suitable for incremental learning,

iii) Crossmodal inference based on higher-order associative strength for text-to-image retrieval.

Fig. 3 illustrates hyperedges generated from a captioned image. As shown in the Fig. 3, each hyperedge represents a higher-order visual-textual association by consisting of the several visual and textual subpatterns. Moreover, the HN has the flexible model structure for incremental learning because new textual words and visual patches involved in unobserved data are added as new vertices and the relationships between the new vertices are included as new hyperedges into the model.

When a captioned image dataset $D = \{(\mathbf{x}_T, \mathbf{x}_I)^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}_T$ denotes the set of textual words in an image description and $\mathbf{x}_I$ is the set of visual patches comprising the image, is sequentially given, an HN can be formally considered as a mixture model of many hyperedges and the empirical distribution is represented with the model:

$$p(\mathbf{x}_T, \mathbf{x}_I \mid H) = \sum_{e \in E} w(e) f(\mathbf{x}_T, \mathbf{x}_I \mid e), \qquad (1)$$

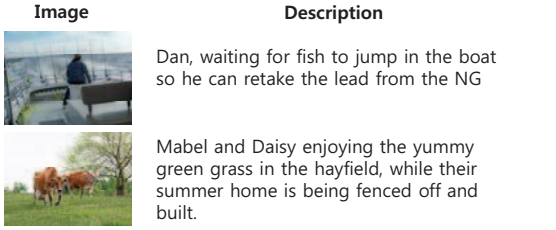

Figure 2. Examples of captioned images used in text-to-image retrieval

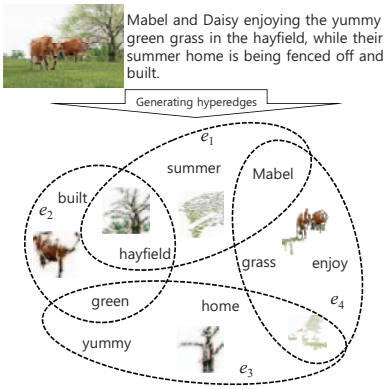

Figure 3. Hyperedges consisting of three texutual words and two visual pateches from a textual-visual data instance (a captioned image)

$$\text{s.t. } 0 < w(e) < 1 \quad \text{and} \quad \sum_{e \in E} w(e) = 1$$

where H denotes an HN model, $w(e)$ denotes the weight of a hyperedge e, and $f(\mathbf{x}_T, \mathbf{x}_I \mid e)$ is the density function. Then, the likelihood of the model is the probability of regenerating the observed data from the model and it is defined to this empirical distribution:

$$p(D \mid H) \approx p(\mathbf{x}_T^{(1)}, \mathbf{x}_I^{(1)}, \mathbf{x}_T^{(2)}, \mathbf{x}_I^{(2)}, ..., \mathbf{x}_T^{(N)}, \mathbf{x}_I^{(N)} \mid H)$$

$$= \prod_{n=1}^{N} p((\mathbf{x}_T, \mathbf{x}_I)^{(n)} \mid H) = \prod_{n=1}^{N} \sum_{e \in E} w(e) f(\mathbf{x}_T, \mathbf{x}_I \mid e) \quad (2)$$

where $\mathbf{x}_T^{(n)}$ and $\mathbf{x}_I^{(n)}$ denote the textual description and the image of the n-*th* captioned image, respectively.

B. Incremental Learning of Multimodal Hypernetworks

Whenever observing an unseen descripted image $(\mathbf{x}_T, \mathbf{x}_I)$, an HN is learned incrementally by predicting the image and updating the weight of the hyperedges. This learning procedure can be formulated by Bayes rule:

$$p(H_n \mid \mathbf{x}_T, \mathbf{x}_I) = \frac{p(\mathbf{x}_T, \mathbf{x}_I \mid H_n) p(H_n)}{p(\mathbf{x}_T, \mathbf{x}_I)}, \quad (3)$$

where H_n denotes the HN at time step n. By this rule, the prior $p(H_n)$ is updated to the posterior $p(H_n \mid \mathbf{x}_T, \mathbf{x}_I)$ by estimating the likelihood $p(\mathbf{x}_T, \mathbf{x}_I \mid H_n)$ and by normalized with $p(\mathbf{x}_T, \mathbf{x}_I)$. The posterior is then used as the prior $p(H_{n+1})$ at the next time step n+1. Reformulating this process recursively using all time steps on the sequence of n data, (3) is described:

$$p(H_n \mid \mathbf{x}_T^{(1:n)}, \mathbf{x}_I^{(1:n)}) = \frac{p(\mathbf{x}_T^{(n)}, \mathbf{x}_I^{(n)} \mid H_n) p(H_{n-1} \mid \mathbf{x}_T^{(1:n-1)}, \mathbf{x}_I^{(1:n-1)})}{P(\mathbf{x}_T^{(n)}, \mathbf{x}_I^{(n)} \mid \mathbf{x}_T^{(1:n-1)}, \mathbf{x}_I^{(1:n-1)})}, \quad (4)$$

$$p(H_n \mid \mathbf{x}_T^{(1:n)}, \mathbf{x}_I^{(1:n)}) \propto p(\mathbf{x}_T^{(n)}, \mathbf{x}_I^{(n)} \mid H_n) p(H_{n-1} \mid \mathbf{x}_T^{(1:n-1)}, \mathbf{x}_I^{(1:n-1)}), \quad (5)$$

where $\mathbf{x}_T^{(1:n)}$ and $\mathbf{x}_I^{(1:n)}$ denote the sequential stream of n textual-visual data. The posterior is estimated by predicting the new image with both hyperedges generated from the new observed image and hyperedges of H_{n-1} learning n-1 images. Each hyperedge is generated by randomly selecting visual patches of the given image and textual words of the description. This generation method assures that there always exists the subpattern involved in the hyperedge in the data. The details of generating hyperedges are explained in [6] and [11]. The weights of the hyperedges are updated by the prediction of the new observed image:

$$\begin{aligned} w_n(e) &= \eta g(e, (\mathbf{x}_T, \mathbf{x}_I)^{(n)}) + (1-\eta) w_{n-1}(e) \\ &= \eta g_T(e_T, \mathbf{x}_T^{(n)}) \cdot g_I(e_I, \mathbf{x}_I^{(n)}) + (1-\eta) w_{n-1}(e) \end{aligned}, \quad (6)$$

s. t.

$$g_T(e_T, \mathbf{x}_T) = | e_T \cap \mathbf{x}_T | \quad \text{and}$$

$$g_I(e_I, \mathbf{x}_I) = \alpha \sum_{\mathbf{u} \in e_I} \max_{\mathbf{v} \in \mathbf{x}_I} \frac{A(\mathbf{u}) A(\mathbf{v}^{\mathrm{T}})}{\left\| A(\mathbf{u}) \right\|_0^1} + (1-\alpha) \sum_{\mathbf{u} \in e_I} \sum_{\mathbf{v} \in \mathbf{x}_I} c(\mathbf{u}, \mathbf{v})$$

Algorithm 1: Incremental Learning of Hypernetworks

H: hypernetwork, E: hyperedge set
W: weight set, R: the iteration number for correction
$V_0 \leftarrow \{\}$, $E_0 \leftarrow \{\}$, $W_0 \leftarrow \{\}$,
For $n \leftarrow$ 1to N
  $V_n \leftarrow V_{n-1} \cup V_{new}^{(n)}$,
  For i $\leftarrow$ 1 to R
    $E_{new} \leftarrow$ Generate(E_{n-1}, $(\mathbf{x}_I, \mathbf{x}_T)^{(n)}$); // $p(H_n \mid \mathbf{x}_T^{(1:n-1)}, \mathbf{x}_I^{(1:n-1)})$
    $E' \leftarrow E_{n-1} \cup E$;
    $(\hat{\mathbf{x}}_T, \hat{\mathbf{x}}_I) \leftarrow$ PredictData(W_{n-1}, E'); // $p(\mathbf{x}_T^{(n)}, \mathbf{x}_I^{(n)} \mid H_n)$
    $W' \leftarrow$ Correct(W_{n-1}, E', $(\hat{\mathbf{x}}_T, \hat{\mathbf{x}}_I)$, $(\mathbf{x}_I, \mathbf{x}_T)^{(n)}$);
  End For
  $E_n \leftarrow E'$; $W_n \leftarrow W'$; $H_n \leftarrow (V_n, E_n, W_n)$; // $p(H_n \mid \mathbf{x}_T^{(1:n)}, \mathbf{x}_I^{(1:n)})$
End For

Figure 4. Algorithm for incremental learning of hypernetworks.

where η is the constant for the current image, and e_T and e_I denote the sets of textual words and visual patches included in hyperedge e, respectively. Also, $\mathbf{u}$ and $\mathbf{v}$ denote the visual patches of e_I and $\mathbf{x}_I$, respectively, $A(\mathbf{u})$ is the function which returns the occurrence vector of SIFT features of $\mathbf{u}$, and $\left\| A(\mathbf{u}) \right\|_0^1$ denotes L_0-norm of $A(\mathbf{u})$, the number of non-zero variables of $A(\mathbf{u})$. In addition, we add $c(\mathbf{u}, \mathbf{v})$ for reflecting the color similarity between two patches because SIFT does not consider a color property.

III. HYPERNETWORK-BASED TEXT-TO-IMAGE RETRIEVAL

An HN facilitates to translate text to image and vice versa by crossmodal inference because the model is the population of textual-visual associative subpatterns. In this paper, we focus on text-to-image translation for image retrieval from textual queries.

A. Preprocessing Textual-Visual Data

The description sentences are represented to the subsets of textual word set used in the training image descriptions by stemming and eliminating the stop words. An image is represented to the set of several visual patches that are extracted by combining two image processing methods: maximally stable external regions (MSER) [7] and scale-invariant feature transform (SIFT) [8]. MSER is a method for detecting an invariant stable subset of external regions of the images and SIFT is a method for extracting the distinctive invariant features from the images. The given images are separately represented to the several external regions by MSER and the set of salient features by SIFT. The regions are then represented with the vectors of the SIFT features using their locality information and we use the SIFT-based regions as the visual patches. For effectively representing the visual patches with SIFT features, k-means clustering method is used in this study because there are few features shared by the regions when images are represented with raw SIFT features. For incremental learning, in addition, the visual patches of a new image are represented with the clustered SIFT features of the observed images by using the clustered features as the centroids for k-means method and by clustering the raw features of the new image with the centroids. Fig. 5 illustrates the flow of converting an image into a visual patch set.

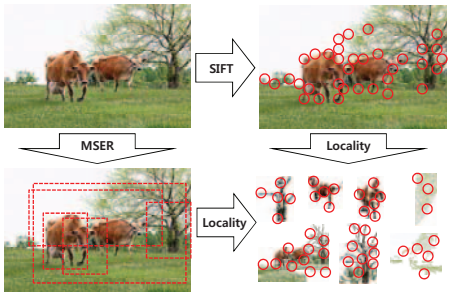

Figure 5. Flow of constructing visual patches from a given image with MSER and SIFT, two feature extraction methods. Boxes and circles denote the external regions extracted by MSER and the salient features by SIFT, respectively.

B. *Image Retrieval from Textual Query using Hypernetworks*

Text-to-image retrieval formally involves calculating the retrieved probability of an image $\mathbf{x}_I$ when a learned model H and a textual query $\mathbf{x}_T$ are given using (2):

$$p(\mathbf{x}_I \mid \mathbf{x}_T, H) = \frac{p(\mathbf{x}_T, \mathbf{x}_I \mid H)}{p(\mathbf{x}_T \mid H)} \propto \sum_{e \in E} w(e) f(\mathbf{x}_T, \mathbf{x}_I \mid e) , \quad (7)$$

Then, the images related to the given textual words are selected as follows:

$$\begin{aligned} I^* &= \arg\max_{\mathbf{x}_I} p(\mathbf{x}_I \mid \mathbf{x}_T, H) \\ &= \arg\max_{\mathbf{x}_I} \sum_{e \in E} w(e) f(\mathbf{x}_T, \mathbf{x}_I \mid e) . \end{aligned} \quad (8)$$

When textual words are given as a query, in order to find I^*, we use the textual-to-visual query expansion method and several images are selected as the candidates of I^* which are most similar to the visual query crossmodally expanded from the given textual query. Thus, (8) is reformulated by substituting $\mathbf{x}_T$ for the textual query Q as follows:

$$I^* = \arg\max_{\mathbf{x}_I} \sum_{e \in E} w(e) f(Q, \mathbf{x}_I \mid e) \approx \arg\max_{\mathbf{x}_I} \delta(\hat{I}, \mathbf{x}_I) , \quad (9)$$

where $\hat{I}$ denotes the visual query expanded from Q and $\delta(\hat{I}, \mathbf{x}_I)$ denotes a similarity function. Formally, a visual query $\hat{I}$ is defined to the set of visual patches involved in hyperedges including the elements of the textual query $Q = \{q_1, \ldots, q_{|Q|}\}$:

$$\hat{I} = \bigcup_{e \in E} \{\mathbf{u} \mid \mathbf{u} \in e, e \in E, Q \cap e \neq \varnothing\} , \quad (10)$$

where $\mathbf{u}$ denotes visual patches. Fig. 6 illustrates an example of the crossmodal query expansion from the textual query to the visual query. Then, the candidate images are retrieved by measuring the similarity between the expanded visual query and the stored images. The similarity is estimated by summing the similarity among the patches involved in $\hat{I}$ and $\mathbf{x}_I$:

$$\delta(\hat{I}, \mathbf{x}_I) = \frac{1}{|\mathbf{x}_I|} \sum_{\mathbf{v} \in \mathbf{x}_I} \sum_{\mathbf{u} \in \hat{I}} w(\mathbf{u}) s(\mathbf{u}, \mathbf{v}) , \quad (11)$$

s.t.

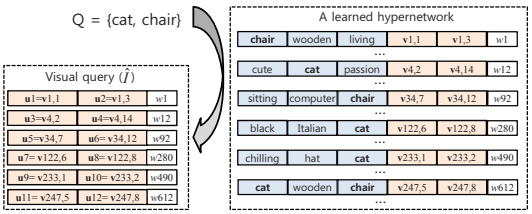

Figure 6. Flow of the crossmodal query expansion from the given textual query to the visual query. $\mathbf{u}_i$ denotes the i-th visual patch of the visual query and $\mathbf{v}_{j,k}$ is the k-th patch of the j-th training image.

$$s(\mathbf{u}, \mathbf{v}) = \begin{cases} g_I(\mathbf{u}, \mathbf{v}), & \text{if } g_I(\mathbf{u}, \mathbf{v}) > \theta \\ 0, & \text{otherwise} \end{cases} \quad \text{and}$$

$$g_I(\mathbf{u}, \mathbf{v}) = \alpha \frac{A(\mathbf{u}) A(\mathbf{v}^{\mathrm{T}})}{\|A(\mathbf{u})\|_0^1} + (1 - \alpha) c(\mathbf{u}, \mathbf{v})$$

where $w(\mathbf{u})$ is the weight of the hyperedge including a visual patch $\mathbf{u}$, $|\mathbf{x}_I|$ is the number of patches in $\mathbf{x}_I$, and θ denotes the threshold to prevent many low-valued patches from distorting the similarity. Therefore, the similarity becomes larger when the image involves the visual patches sharing more SIFT features with the patches of the visual query. Then, the images with large $\delta(\hat{I}, \mathbf{x}_I)$ are retrieved as candidate images related to the textual query.

IV. Experimental Results

A. *Data and Experimental Setup*

We evaluate the proposed T2I retrieval method with the dataset consisting of 3,000 photography images described by the sentences from Flickr.com. For evaluation, we divide the dataset into training set and test set consisting of 1,000 and 2,000 images, respectively. Each description is represented with the subset of 2,814 textual words. An image is converted into the set of the visual patches represented with the occurrence vector of 500 clustered SIFT features. Table I shows parameter setups for the method.

TABLE I. Parameter Setup for Model Learning

Parameters	Values
Number of visual patches in a hyperedge	2
Number of textual words in a hyperedge	3
Number of hyperedges generated from an image	5
Number of iterations for correction	5
α (constant for balancing SIFT and color)	0.99
θ (Patch similarity threshold)	0.9

B. *Text-to-Image Retrieval Performance*

We use three measures such as precision, recall, and successful retrieval (SR) for evaluating the performance of the HN-based T2I retrieval method. In order to define the measures, we call it correct retrieval (CR) that a retrieved image explicitly includes the object that is described by the given textual query, i.e., query-object. Then, each measure is defined as follows:

$$precision = \frac{\#\text{of CR}}{\#\text{of the retrieved images}} , \quad (12)$$

$$recall = \frac{\#\text{of CR}}{\#\text{of all images including the query-object}}, \quad (13)$$

$$SR = \begin{cases} 1, & precision > 0 \\ 0, & \text{otherwise} \end{cases}. \quad (14)$$

Moreover, we use two types of textual queries such as 10 queries and 30 queries for measuring the performance, and the textual queries are enumerated in Table II.

Table III presents the precision and recall of T2I retrieval of the HN-based method with models learning all the training images (1000-HNs) for 10 queries on the test set compared to a baseline method. The baseline method uses all visual patches of the training images including the textual query in their description as the visual query without any learning process. From Table III, the proposed T2I method outperforms the baseline method in terms of both precision and recall. The performances of the baseline method are lower than those of the proposed method since the expanded query of the baseline method is blurred by too many patches and the specificity is thus weakened. Meanwhile, large-weighted hyperedges with the strong associative relationships only survive in the HN by the incremental learning. Fig. 7 presents the average precision and recall of T2I retrieval on the training set and the test set for

TABLE II. CONTENTS OF TWO TYPES OF TEXTUAL QUERIES

Query	Textual words
10 queries	beach, boat, cat, flower, girl, grass, sand, sky, tree, water
30 queries	beach, bird, boat, bridge, building, castle, cat, chair, dog, dress, fish, floor, flower, girl, grass, house, kitchen, mountain, office, river, road, rock, room, sand, sky, table, tower, tree, wall, water

TALBE III. PRECISION AND RECALL OF TEXT-TO-IMAGE RETRIEVAL FOR 10 QUERIES

Methods	Precision	Recall
HN-based (1000-HN)	**0.24**	**0.055**
Baseline	0.155	0.035

TALBE IV. SUCCESSFUL RETRIEVAL FOR 30 QUERIES ON THE TRAING SET AND THE TEST SET

SR	Number of retrieved images				
	1	5	10	15	20
Training	0.6	0.933	1.0	1.0	1.0
Test	0.2	0.567	0.733	0.8	0.9

Figure 7. Precision and recall of text-to-image retrieval using 1000-hyperentworks for 30 queries on (a) the training dataset and (b) the test set. Values are averaged for 30 queries.

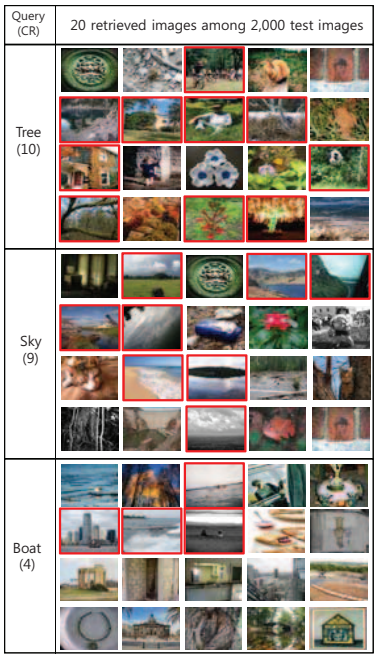

Figure 8. Retrieved images from the test dataset using 1000-hypernetworks for three queries. Red boxes are images including query-objects.

30 textual queries with 1000-HNs. As shown in Fig. 7, the results show the general pattern that the precision slightly decreases and the recall increases as the number of retrieved images grows up. For the training set, the precisions are mostly larger than 0.5 and it means that more than half of the retrieved images are related to the textual queries. From Fig. 7(b), we indicate that one or more images are associated with the given query when the number of the retrieved images is 10 from the test set. Table IV shows the SR of T2I retrieval for the same queries and model as Fig. 7. Therefore, our method can retrieve images related to the textual query even if the images have no textual description. Fig. 8 illustrates 20 retrieved images from the test dataset for each query, totally 60 images for three queries including 'tree', 'sky', and 'boat'. The precisions of 'sky' and 'tree' are higher than that of 'boat' because the patches of 'sky' and 'tree' are similar to each other due to sharing more SIFT features than the patches of boat. Fig. 9 shows the precision and the recall for 10 queries on the test set as the retrieval size increases. Although the performances are better than those of 30 queries because textual words in 10 queries are more selective, the result in Fig. 9 is also consistent to Fig. 7. In terms of SR, therefore, we indicate that images

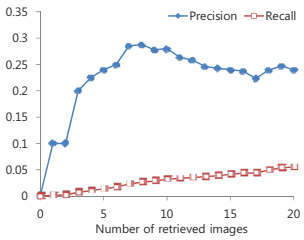

Figure 9. Precision and recall of text-to-image retrieval using 1000-hypernetworks for 10 queries on the test set as the retrieval size grows.

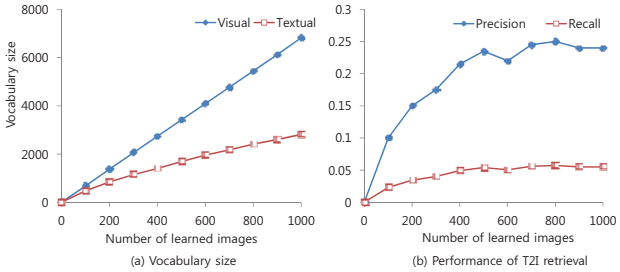

Figure 10. (a) Vocabulary size of hyperentworks and (b) average performance of text-to-image retrieval for 10 queries as learning proceeds incrementally. In (b), the number of retrieved images is 20.

associated with the text query can be retrieved with textual-visual crossmodal association by the HN-based T2I retrieval method from Table IV and Figs. 7, 8, and 9.

C. *Incremental Learning for Text-to-Image Retrieval*

Fig. 10 shows (a) the sizes of the textual and the visual vocabularies of the model and (b) the performances of T2I retrieval, as the learning incrementally proceeds. As the observed data grow up, from Fig. 10(a), visual information increases linearly due to the uniqueness of the patch while the textual vocabulary grows up slowly due to the frequently used words. In terms of performance, the precision and the recall are enhanced in early learning steps due to the increase of the number of hyperedges as well as the textual and the visual vocabulary sizes. Meanwhile, the performances are saturated after learning more than 500 training images. The reason is that the model contains redundant information despite the uniqueness of the visual patches because the patches including the same object share many SIFT features. In addition, we can indicate that the decay of the performances is caused by the patches sharing SIFT features but including the different objects and this issue can be solved by specifically representing the visual patches with more SIFT features. Fig. 11 illustrates the retrieved images for two textual queries including 'sky' and 'grass' as the increase of the observed images. Same as Fig. 10(b), the model observing more images shows the higher retrieval performance for both queries in Fig. 11.

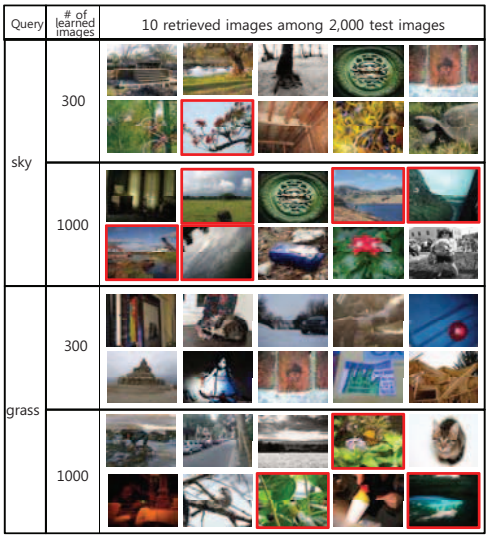

Figure 11. Retrieved images on the test dataset for two textual queries such as 'sky' and 'grass' with 300-hypernetworks and 1000-hypernetworks. Red boxed photos are images including the query-object.

V. Concluding Remarks

We have proposed a novel text-to-image (T2I) retrieval method based on a textual-visual association model and we use hypernetwork models for incrementally learning the associative relationships between two modalities. Moreover, the images related to the text queries are retrieved by crossmodal query expansion with the learned model. The proposed method was evaluated on 3,000 images with text descriptions from Flick.com to retrieve images associated with various text queries. Experimental results show that our method achieves the high retrieval performance in terms of precision, recall, and successful retrieval on the test dataset compared to a baseline method. Moreover, the results demonstrate that the hypernetworks can learn robustly increasing data with the proposed incremental learning method.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang. "Image retrieval: Ideas, influences, and trends of the new age." *ACM Compuler Surveys*, vol. 40(2), pp. 1-60, 2008.

[2] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* vol. 22, no. 12, pp.1349-1380, 2000.

[3] S. L. Feng, R. Manmatha, and V. Lavrenko. "Multiple bernoulli relevance models for image and video annotation." In *Proc of Computer Vision (ICCV), 2004 IEEE International Conference on*, pp. II-1002-1009, 2004.

[4] R. Zhang, Z. Zhang, M. Li, W.-T. Ma, and H.-J. Zhang, "A probabilistic semantic model for image annotation and multimodal image retrieval," In *Proc. of Computer Vision, (ICCV) 2005 IEEE International Conference on*, pp. 849-851, 2005.

[5] W. Li, L. Duan, D. Xu, and I.W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," In *Proc of Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2049-2055, 2011.

[6] 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.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp.761-767, 2004.

[8] D. G. Lowe, Distinctive image features from scale-invariant keypoints. *International Journal of Computer Visioin*, vol. 2, no. 60, pp. 91-110, 2004.

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

[10] S.-J. Kim, J.-W. Ha, B. Lee, and B.-T. Zhang, "Evolutionary layered hypernetworks for identifying microRNA-mRNA regulatory modules," *IEEE World Congress Computational Intelligence (WCCI-CEC 2010)*, pp. 2299-2306, 2010.

[11] J.-W. Ha, B.-H. Kim, B. Lee, and B.-T. Zhang, "Layered hypernetwork models for cross-modal associative text and image keyword generation in multimodal information retrieval," In *Proc. of the 11th Pacific Rim International Conference on AI (PRICAI2010), Trends in Artificial Intelligence*, vol. 6230, pp. 76-87, 2010.