Visual Query Expansion via Incremental Hypernetwork Models of Image and Text

Min-Oh Heo, Myunggu Kang, and Byoung-Tak Zhang

Biointelligence Lab, School of Computer Science and Engineering,
Seoul National University,
599 Gwanak-ro, Gwank-gu, Seoul 151-744, Korea
{moheo,mgkang,btzhang}@bi.snu.ac.kr

Abstract. Humans can associate vision and language modalities and thus generate mental imagery, i.e. visual images, from linguistic input in an environment of unlimited inflowing information. Inspired by human memory, we separate a text-to-image retrieval task into two steps: 1) text-to-image conversion (generating visual queries for the 2 step) and 2) image-to-image retrieval task. This separation is advantageous for inner representation visualization, learning incremental dataset, using the results of content-based image retrieval. Here, we propose a visual query expansion method that simulates the capability of human associative memory. We use a hyperenetwork model (HN) that combines visual words and linguistic words. HNs learn the higher-order cross-modal associative relationships incrementally on a set of image-text pairs in sequence. An incremental HN generates images by assembling visual words based on linguistic cues. And we retrieve similar images with the generated visual query. The method is evaluated on 26 video clips of 'Thomas and Friends'. Experiments show the performance of successive image retrieval rate up to 98.1% with a single text cue. It shows the additional potential to generate the visual query with several text cues simultaneously.

Keywords: hypernetwork, incremental data, visual query expansion, vision-language, text-to-image, multimodal information processing.

1 Introduction

Conventional text-to-image retrieval methods for image-text corpus have used the annotated tags on images that are used for searching for the target [1]. Recently, multi-modal data such as video, sound, images as well as web-pages including images are increasing explosively. Consequently, the underlying data distribution may change over time [3]. So, we need incremental models to learn the data of multi-modality.

Humans can associate vision and language modalities and thus generate mental imagery, i.e. visual images, from linguistic input in the environment of unlimited inflowing information. Considering human capability of multimodal memory [2,5,16], we separate a text-to-image retrieval task into two steps. In the first step, text-to-image conversion is used to generate the visual concept from the related

B.-T. Zhang and M.A. Orgun (Eds.): PRICAI 2010, LNAI 6230, pp. 88–99, 2010.

images associated with text cues. And the second step is to search for similar images with the expanded visual query from the first step. This approach gives some advantages. First, we can visualize the inner representation of the form of visual images. Secondly, we can deal with incremental data by updating visual queries incrementally in the first step. Thirdly, we can bring the result from content-based image retrieval (CBIR) for the second step. In addition, after generating visual queries with enough large data, we expect the visual queries to be the universal visual concepts when retrieving from all image databases.

Here, we propose a novel visual query expansion method that simulates the capability of human associative memory. Hypernetwork models (HN) have cognitive properties of continuity, glocality, and compositionality [5]. And HNs learn higher-order cross-modal association to solve the difference of granularity in image and text features. HNs can be appended and updated partially by adding new hyperedges from new observations as incremental learning. Especially, we built a visual word dictionary keeping the regional information from an image beforehand. This enables us to visualize the visual query and avoid the limitation of computational complexity for the image representation. As Fig. 1 shows, 1007 image-text pairs were captured from 26 video clips of Thomas and Friends. And we simply used the sum of absolute difference in RGB scale between images as the second step.

This paper is organized as follows. Section 2 summarizes related works. Then hypernetworks will be introduced briefly in Section 3 and a proposed method is explained in Section 4. Section 5 shows the experimental results. Finally, Section 6 concludes this paper with concluding remarks.

2 Related Work

Crossmodal data retrieval has been focused on the information retrieval field, as a result of readily available multimedia data. Approaches using multimodal data have been introduced using tagging based methods such as automatic tagging and annotation and statistical dependency based methods such as co-occurrence and canonical correlation analysis (CCA) [1-2]. And approaches using image annotation

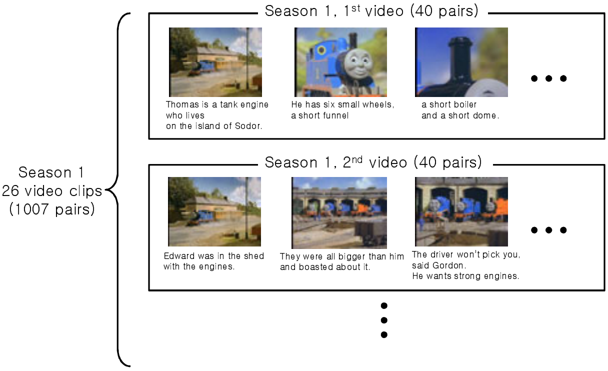
Season 1, 1st video (40 pairs)
Thomas is a tank engine who lives on the island of Sodor.
He has six small wheels, a short funnel
a short boiler and a short dome.
Season 1 26 video clips (1007 pairs)
Season 1, 2nd video (40 pairs)
Edward was in the shed with the engines.
They were all bigger than him and boasted about it.
The driver won't pick you, said Gordon. He wants strong engines.

Fig. 1. The training dataset used in this paper. The pairs from one clip are one unit of instances for sequential presentation on incremental learning.

information were studied. Jeon *et al.* proposed a cross-media relevance model (CMRM) [6] using annotated images and grouping small blobs of images manually. And Pan *et al.* studied graph-based methods for the correlated nodes discovery across other modalities [7]. And cross-modal association learning has been applied to video data. Yan *et al.* studied a text-image multimodal retrieval task on data of a broadcast new video [9] and Snoek *et al.* suggested a concept-based video retrieval method [8]. Additionally, D. Li *et al.* proposed a factor analysis method based on cross-modal association [10].

For the visual query expansion, it is mainly used to improve the performance of the retrieval task. Chum *et al.* introduced query expansion using images by analogy for the text retrieval. They used images as added queries giving spatial constraints and improved the retrieval performance for false negatives [12]. Joly *et al.* applied this concept to logo retrieval in large image collection [13] and Jiang et al. did this to bag-of-visual-words [14]. As visual representational aspects, a visual mental imagery is used as inner representation of cognitive processes of humans [16], AIs [17] and even robots [18].

In [4], Ha *et al.* studied the image-text cross-modal retrieval task with multimodal queries based on pixels of the gray scale on the fixed dataset. On the contrary, we deal with the relevant image retrieval task based on incremental HNs with color image patches on the increasing dataset.

3 Multi-modal Hypernetwork Models

3.1 Hypernetwork Models

A hypernetwork (HN) is a hypergraph which is represented with vertices and weighted hyperedges. Hypergraphs refer to generalized simple graphs by allowing for edges of higher cardinality. The edges in a hypergraph are called hyperedges. Fig. 2 shows an example of HN. In formal definition, a HN is defined as *H* = (*V*, *E*, *W*) where *V*, *E* and *W* are a set of vertices, hyperedges, and weights respectively. And the elements of *W* correspond to the elements of E. A HN is formulated on the basis of probabilistic theory. Given a data set $D = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ of *N* samples, the HN can be

$$P(D \mid W) = \prod_{n=1}^{N} P(\mathbf{x}^{(n)} \mid W) \tag{1}$$

$$P(\mathbf{x}^{(n)} \mid W) = \frac{1}{Z(W)} \exp\left(-\varepsilon(\mathbf{x}^{(n)}; W)\right) \tag{2}$$

where *Z(W)* denotes the partition function as the normalization term and $\mathbf{x}^{(n)}$ means the *n*-th instance of data. And ε is the energy function of HN and the partition function are defined as

$$\varepsilon(\mathbf{x}^{(n)}; W) = -\sum_{m=1}^{|E|} w_m \delta(\mathbf{x}^{(n)}, E_m) \tag{3}$$

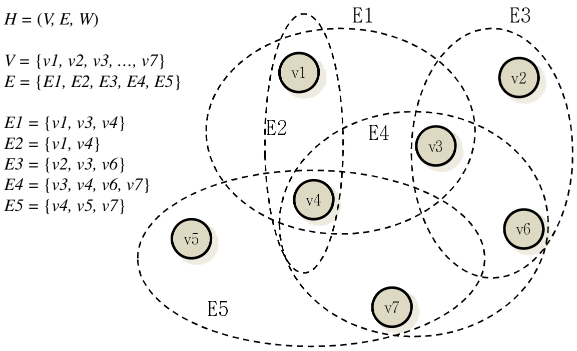

Fig. 2. An example of a hypernetwork. Hypernetwork *H* is composed of vertices set *V*, hyperedge set *E* and the corresponding weight *W*.

$$Z(W)=\sum_{n=1}^{|D|}\exp\left(-\sum_{m=1}^{|E|}-w_m\delta(\mathbf{x}^{(n)},E_m)\right) \tag{4}$$

where $w_i^{(k)}$ is a positive real-valued weight of *i*-th hyperedge E_i and $\delta(\mathbf{x}^{(n)}, E_i)$ denotes the identity function depending on input parameter elements of $\mathbf{x}^{(n)}$ and hyperedge E_i.

Taking the derivative of log-likelihood function of (2), we can derive the following

$$\ln P\left(D|W\right)=\ln\prod_{n=1}^{N}P(x^{(n)}\,|\,W) \tag{5}$$

$$\nabla_W \ln\prod\nolimits_{n=1}^{N}P(x^{(n)}\,|\,W)=\nabla_W\left\{\ln\prod\nolimits_{n=1}^{N}\frac{1}{Z(W)}\exp\left(-\varepsilon(x^{(n)};W)\right)\right\} \tag{6}$$

$$=N\left\{\left\langle\delta(\mathbf{x}^{(n)},E_m)\right\rangle_{Data}-\left\langle\delta(\mathbf{x}^{(n)},E_m)\right\rangle_{P(x|W)}\right\}$$

And minimizing the difference between two average frequencies is equivalent to maximizing the likelihood by making (6) be equal to zero [5].

Then, the term

$$\sum\nolimits_{n=1}^{N}\sum_{m=1}^{|E|}\delta(\mathbf{x}^{(n)},E_m)=N\left\langle\delta(\mathbf{x}^{(n)},E_m)\right\rangle_{Data} \tag{7}$$

can also be derived and it means that the total number of matching hyperedges with the given data set *D* follows the average frequencies of the hyperedges in the data set.

3.2 Cross-Modal Associative Learning on Incremental Hypernetwork Models

To learn cross-modal associative information, we create cross-modal hyperedges composed exclusively of the textual part and visual part, which are sampled from text and image respectively, as shown in Fig. 3. Formally, given an instance x= {x$_I$, x$_T$}, x$_I$ is the feature set for image representation and x$_T$ is that for text representation:

$$X_I = \{x_1^i, x_2^i, x_3^i, \ldots, x_P^i\} \tag{8}$$

$$X_T = \{x_1^t, x_2^t, x_3^t, \ldots, x_Q^t\} \tag{9}$$

where P and Q are the number of features for images and text respectively, which means the size of visual word dictionary and linguistic word dictionary. x^i_k and x^t_j are features denoting the k-th element of the visual word dictionary and the j-th one of the text word dictionary respectively. Then the joint distribution given arbitrary weights from (1) can be converted using the composition of hyperedges, and written into the formulation taken from (7) by changing the weight reflecting the number of matched instances among the size N of dataset.

$$P(D|W) = P(D_I, D_T|W) \propto \sum_{n=1}^{N}\sum_{m=1}^{|E|}\delta(\mathbf{x}^{(n)}, E_m) = \sum_{m=1}^{|E|} w_m \delta(\mathbf{x}, E_m) \tag{10}$$

where D_I is the dataset of image features and D_T is the one of text features. Then, the distribution is represented by weighted nonzero basis functions having a zero-one binary value. However, all of the possible hyperedges from order 1 to the order of the number of total features is almost impossible by virtue of combinatorial explosion which dictates that the number of cases will massively increase. So, we should approximate this with the relatively small number of hyperedges by using random sampling strategy. We can approximate the joint distribution using M hyperedges like this formula,

$$P(D_I, D_T|W) \propto \sum_{m=1}^{|E|} w_m \delta(\mathbf{x}, E_m) \simeq \sum_{m=1}^{M} w_m \delta(\mathbf{x}, E_m) \tag{11}$$

if M is large enough to express the distribution, the error between the estimation result and the distribution will be decreased. By this fact, we can estimate the distribution roughly by simply using a reasonably small number of hyperedges.

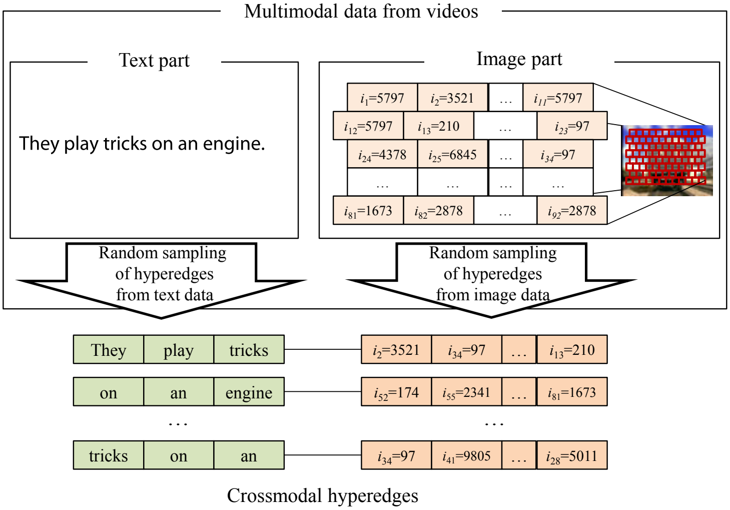
Multimodal data from videos
Text part
They play tricks on an engine.
Image part
i_1=5797 | i_2=3521 | … | i_{11}=5797
i_{12}=5797 | i_{13}=210 | … | i_{23}=97
i_{24}=4378 | i_{25}=6845 | … | i_{34}=97
… | … | … | …
i_{81}=1673 | i_{82}=2878 | … | i_{92}=2878
Random sampling of hyperedges from text data
Random sampling of hyperedges from image data
They | play | tricks — i_2=3521 | i_{34}=97 | … | i_{13}=210
on | an | engine — i_{52}=174 | i_{55}=2341 | … | i_{81}=1673
…
tricks | on | an — i_{34}=97 | i_{41}=9805 | … | i_{28}=5011
Crossmodal hyperedges

Fig. 3. An example of cross-modal hyperedges using the visual word dictionary. For the experiments, tri-gram is used for the sampling from text part and image patches random sampled among 92 regions on the grid for image part.

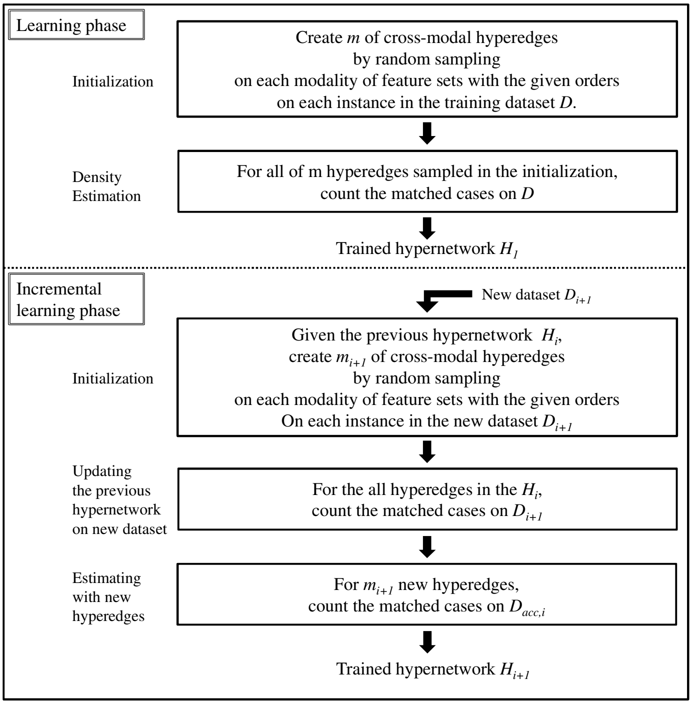
Learning phase
Initialization
Create m of cross-modal hyperedges by random sampling on each modality of feature sets with the given orders on each instance in the training dataset D.
Density Estimation
For all of m hyperedges sampled in the initialization, count the matched cases on D
Trained hypernetwork H_1
Incremental learning phase
New dataset D_{i+1}
Initialization
Given the previous hypernetwork H_i, create m_{i+1} of cross-modal hyperedges by random sampling on each modality of feature sets with the given orders On each instance in the new dataset D_{i+1}
Updating the previous hypernetwork on new dataset
For the all hyperedges in the H_i, count the matched cases on D_{i+1}
Estimating with new hyperedges
For m_{i+1} new hyperedges, count the matched cases on $D_{acc,i}$
Trained hypernetwork H_{i+1}

Fig. 4. The flow chart for cross-modal associative learning. The top shows the case for the fixed dataset and the bottom shows that for the incremental dataset.

For incremental HN learning, we can easily apply the same strategy with a small adjustment. Formally, we define the preliminary dataset as D_0 and the n-th new dataset as D_{n+1}. Then, the n-th accumulated training set $D^{(n+1)}$ can be written as follows:

$$D^{(n+1)} = D^{(n)} \cup D_{n+1} \tag{12}$$

Whenever there is an inflowing new dataset, adding new hyperedges from it by random sampling strategy can maintain the small error between the estimation and the distribution while keeping the condition that the number of hypedges is enough to follow. The process is summarized in Fig. 4.

4 A Visual Query Expansion Method

4.1 Building a Visual Word Dictionary for Image Patches

Visual query expansion needs image processing for using visual features. Avoiding the vast computational complexity on the image representation, we built a visual word dictionary including 10,000 visual words beforehand. This process is illustrated in Fig. 5. As image preprocessing, each image is firstly segmented into 15×15 square image patches on a regular grid shown in the second image in Fig. 3. Following the work of Feng *et al.* [15], using the rectangular regions could provide performance gains compared with using regions by automatic image segmentation methods. We were also able to avoid the problems associated with the computational cost. Secondly, we assigned all of the segmented patches into k groups by k-means clustering in the RGB color space using Koen's image processing package [11]. As a result, we made 10,000 visual words by choosing the closest visual word from the centroid of each cluster. This set of image patches worked as visual words in this paper.

4.2 Visual Query Expansion by Combining Image Patches

Expanded Visual query can be created by the following process. When given the linguistic cue which works as the condition on the (10), we can make inference with the trained HN by the following formula

$$P(D_I \mid D_{T_q}, W) = \frac{P(D_I, D_{T_q} \mid W)}{P(D_{T_q} \mid W)} \propto \sum_{m \in E_{T_q}} w_m \delta(\mathbf{x}, E_m) \tag{13}$$

where the set E_{Tq} of cross-modal hyperedges including the text T_q. Then, we choose the index x_p^i of visual word that makes conditional likelihood be the maximum at the *j*-th region on the grid as follows:

$$I_j^{vq} = \arg\max_p P(I_j = f(x_p^i) \mid D_{T_q}, W) = \arg\max_p \sum_{m \in E_{Tq}} w_m \delta(\mathbf{x}, E_m) \tag{14}$$

where *f* is the mapping function to the visual word. And combining them generates visual query. This process can be achieved on HNs by choosing the visual word that maximum weight of hyperedges which are relevant to the text T_q as in the following summarized procedure.

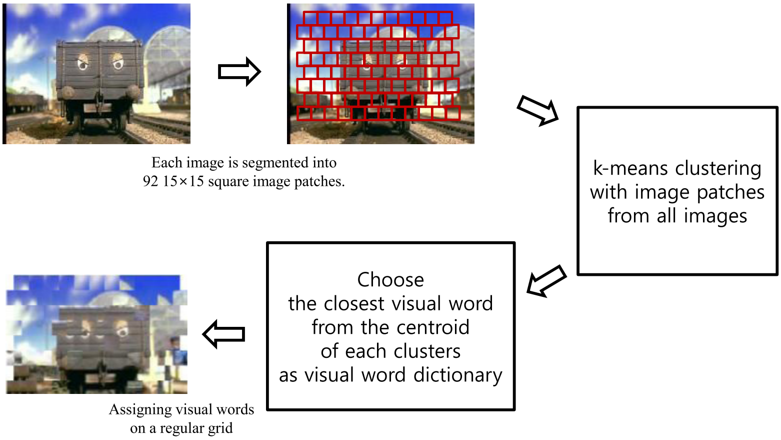

Fig. 5. The process to build a visual word dictionary and to convert original images into ones to be trained. All of the image patches segmented are grouped into 10,000 clusters and converted by the closest visual word from original image patches.

1. Summing up the weights of hyperedges having the text T_q.
2. Choose the index of visual word that make conditional likelihood be maximum at the j-th region on the grid.
3. Combining the image patches with the corresponding index at the j-th region.

5 Experimental Results

5.1 Data and Experimental Setups

As mentioned briefly in Section 1 and Fig. 1, we captured 1007 image-text pairs from 26 video clips of Thomas and Friends season 1. We used a capture tool to collect image-text pairs automatically whenever a subtitle appeared. Table 1 shows the distribution across 26 video clips. And the experimental setting is shown in Table 2.

5.2 Experimental Results

During the incremental learning, HNs were trained in sequence and retrieved top-N closest images using the sum of absolute difference in RGB scale between the generated visual query and original images to perform an image-to-image retrieval task. Fig. 6 and Fig. 7 show the results of image order 5 and order 35 each when the cue 'engine' is given. Then, shown in order, are the generated visual query, the closest top-5 images near the visual query in that dataset D_n and all of the original images associated with the cue in D_n. The associated original images are the same, but the generated visual queries are rather different, which cause the top-5 retrieved images to also be different. They include some original images (10/23 cases of nonzero original images, 10/45 in total). The visual queries generated from 5-order HNs are more flexible to incoming new instances than those by 35-order HNs. (25 consecutive difference (Σ $|I_i - I_{i-1}|$) per pixel: $\sigma_5 = 18.7 < \sigma_{35} = 26.6$, $m_5=17.1 \approx m_{35}=18.7$).

In more than 2 words cases, the visual query can be generated. Fig. 8 shows a comparison between the case given text cues 'noise' and 'once' simultaneously and each. Though they are generated from the same HN model, at each case, they reflect the original images well, and the 2 words case do also. Even though there is no instance

Table 1. The frequencies of instances in the incremental dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Total
40	40	35	39	45	36	35	38	44	46	41	38	42	40	36	38	38	37	35	38	41	39	36	36	35	39	1007

Table 2. The information and parameter set for the experiments

Information	Values
Total data	1007 in 26 sets
Total text words	1256
Number of regions on 1 image	92
Number of visual words	10,000

Parameters	Values
Text order	3 (tri-gram)
Image order	5, 35
Sampling rate	10
Image patch size	15 ×15

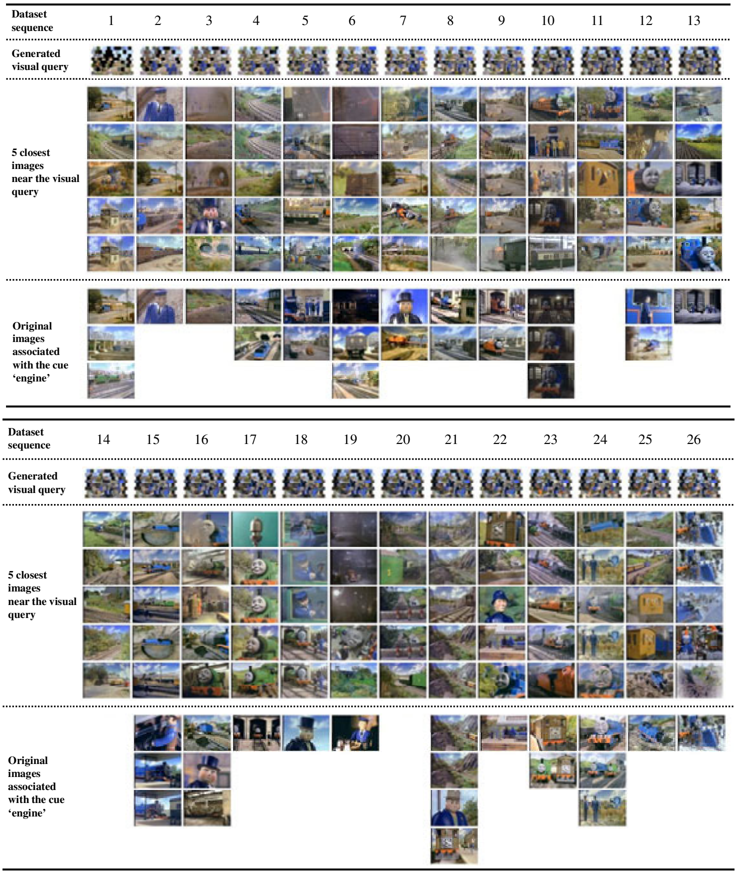
Dataset sequence
1
2
3
4
5
6
7
8
9
10
11
12
13
Generated visual query
5 closest images near the visual query
Original images associated with the cue 'engine'
Dataset sequence
14
15
16
17
18
19
20
21
22
23
24
25
26
Generated visual query
5 closest images near the visual query
Original images associated with the cue 'engine'

Fig. 6. An example in sequential presentation from top left to top right, and bottom left to bottom right. It shows the generated visual queries, the related original images and the retrieved top-5 images. (image order: 5, linguistic cue: engine).

having 'noise' and 'once' together (not even in the same dataset), the visual query with mixed two cases can emerge when given the cues 'noise' and 'once' together. This point is important if the amount of data is very large, because one text can have the visual concept each, which they can work as additive prototypes.

The result of the overall retrieval performance is summarized in Table 3. It is done by checking whether more than one original image is retrieved for each linguistic cue in text dictionary during the incremental learning. If the large portion of the corpus is sparse, unsupervised learning methods confront the difficulty of learning the specific information for the discrimination. To show general characteristics of performance, we may ignore the cases of low frequencies. As a result, then, we get higher accuracy to retrieve relevant images.

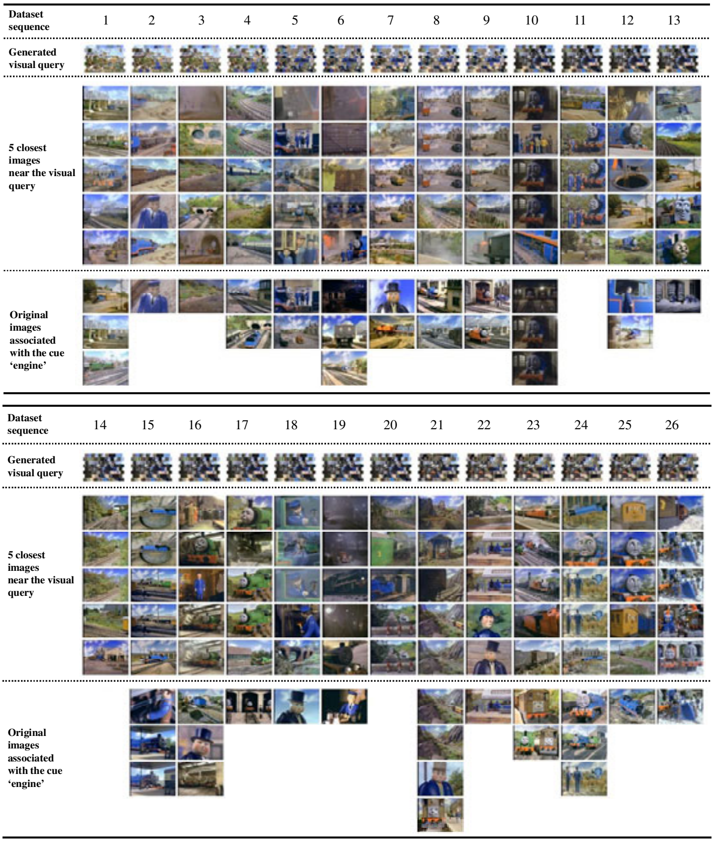
Dataset sequence
1 2 3 4 5 6 7 8 9 10 11 12 13
Generated visual query
5 closest images near the visual query
Original images associated with the cue 'engine'
Dataset sequence
14 15 16 17 18 19 20 21 22 23 24 25 26
Generated visual query
5 closest images near the visual query
Original images associated with the cue 'engine'

Fig. 7. An example in sequential presentation from top left to top right, and bottom left to bottom right. It shows the generated visual queries, the related original images and the retrieved top-5 images. (image order: 35, linguistic cue: engine).

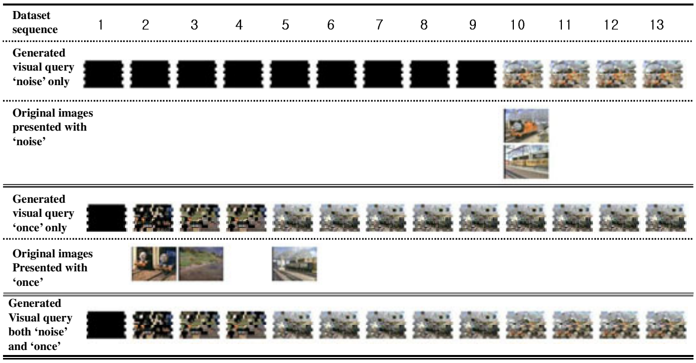
Dataset sequence
1 2 3 4 5 6 7 8 9 10 11 12 13
Generated visual query 'noise' only
Original images presented with 'noise'
Generated visual query 'once' only
Original images Presented with 'once'
Generated Visual query both 'noise' and 'once'

Fig. 8. An example of mixed words giving the cues 'noise' and 'once' by sequence 13. Even though they do not occur in the same instances, generated visual query reflects the original images together well. The reason for black visual queries in the left part comes from presenting no instance to learn 'noise' and 'once' yet.

Table 3. The overall performance of retrieval results in various tasks (order: 35)

Retrieval task		# of cases	Size of retrieved candidates (Top-N)			
			3	5	8	10
All cases	Successful cases	1256	334	462	617	692
	Percentage (%)		26.6%	36.8%	49.1%	55.1%
Cases of freq. >= 3	Successful cases	528	253	338	414	438
	Percentage (%)		47.9%	64.0%	78.4%	83.0%
Cases of freq. >= 5	Successful cases	380	215	286	336	343
	Percentage (%)		56.6%	75.3%	88.4%	90.3%
Cases of freq. >= 7	Successful cases	288	187	237	270	272
	Percentage (%)		64.9%	82.3%	93.8%	94.4%
Cases of freq. >= 10	Successful cases	208	154	190	203	204
	Percentage (%)		74.0%	91.4%	97.6%	98.1%

6 Concluding Remarks

We separated text-to-image retrieval task into two steps as follows: 1) text-to-image conversion and 2) image-to-image retrieval. And we proposed a method to generate visual query based on cross-modal associative learning by incremental hypernetwork models with the focus on the text-image conversion reflecting the related images from an image-text corpus. Experimental results show that the visual query generated by this method can be used for the image-to-image retrieval task. In this study, we just estimate with the small number of bases of the specific order (k-order hyperedges) without explicit learning process. We will go on to establish proper learning processes with unsupervised HNs and apply proper CBIR methods to the second step.

Acknowledgements

This work was supported in part by IT R&D Program of MKE/KEIT (KI002138, MARS), by NRF Grant of MEST (314-2008-1-D00377, Xtran), and by the BK21-IT program funded by Korean Government (MEST). The authors thank Jung-Woo Ha and Byoung-Hee Kim for helpful discussion. And we also thank Minsu Cho and Prof. Kyoung Mu Lee for image processing.

References

[1] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), Article 5, 40(2) (2008)

[2] The Stanford Encyclopedia of Philosophy, http://plato.stanford.edu

[3] Tsymbal, A.: The problem of concept drift: definitions and related work, Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin, Ireland (2004),
http://www.cs.tcd.ie/publications/
tech-reports/reports.04/TCD-CS-2004-15.pdf

[4] Ha, J.-W., Kim, B.-H., Kim, H.-W., Yoon, W.C., Eom, J.-H., Zhang, B.-T.: Text-to-image cross-modal retrieval of magazine articles based on higher-order pattern recall by hypernetworks. In: The 10th Int. Symposium on Advanced Intelligent Systems (ISIS 2009), pp. 274–277 (2009)

[5] Zhang, B.-T.: Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. IEEE Computational Intelligence Magazine 3(3), 49–63 (2008)

[6] Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: The 26th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 119–126 (2003)

[7] Pan, J.-Y., Yang, H.-J., Faloutsos, C., Duygulu, P.: Automatic Multimedia Cross-modal Correlation Discovery. In: The 10th ACM SIGKDD Conf. on Knowledge discovery and data mining, pp. 653–658. Association for Computing Machinery, New York (2004)

[8] Snoek, C.G.M., Worring, M.: Concept-based video retrieval. Foundations and Trends in Information Retrieval 2(4), 215–322 (2009)

[9] Yan, R., Hauptmann, A.G.: A review of text and image retrieval approaches for broadcast news video. Information Retrieval 10(4-5), 445–484 (2007)

[10] Li, D., Dimitrova, N., Li, M., Sethi, K.: Multimedia content processing through cross-modal association. In: Proc. of the 11th ACM Int. Conf. on Multimedia (MM 2003), pp. 604–611 (2003)

[11] van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010) (in Press)

[12] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In: Proc. 11th Int. Conf. on Computer Vision (ICCV 2007), pp. 1–8 (2007)

[13] Joly, A., Buisson, O.: Logo Retrieval with a Contrario Visual Query Expansion. In: Proc. 7th ACM Int. Conf. on Multimedia, pp. 581–584 (2009)

[14] Jiang, Y.-G., Ngo, C.-W.: Bag-of-Visual-Words Expansion using Visual Relatedness for Video Indexing. In: Proc. 31st Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 769–770 (2008)

[15] Feng, S., Manmatha, R., Lavrenko, V.: Multiple Bernoulli Relevance Models for Image and Video Annotation. In: CVPR 2004 (2), pp. 1002–1009 (2004)

[16] Block, N.: Imagery. MIT Press, Cambridge (1981)

[17] Glasgow, J., Papadias, D.: Computational imagery. Cognitive Science 16, 355–394 (1992)

[18] Roy, D., Hsiao, K.-Y., Mavridis, N.: Mental Imagery for a Conversational Robot. IEEE Transactions on Systems, Man, and Cybernetics, Part B 34, 1374–1383 (2004)