# Automatic Construction of A Folksonomy-based Visual Ontology

Hidetoshi Kawakubo, Yuuta Akima and Keiji Yanai
*Department of Computer Science, The University of Electro-Communications, Tokyo*
*1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 Japan*
*Email: {kawaku-h,akima-y,yanai}mm.cs.uec.ac.jp*

***Abstract*—Recently, Folksonomy attracts attentions as a new method to index large-scale image databases. In the Folksonomy-style image databases, they allows users to attach keywords to images as "tags". Since tag words are uncontrolled, they have various and many kinds of tags associated with images. This is much different from conventional image databases. In this paper, we propose a novel method to extract hierarchical structure on relations between tags from Folksonomy. The tag structure we extract can be used as an ontology for image database search which reflects both textual and visual relations between tags. In the proposed method, at first, we collect millions of tag-attached-images from Flickr which is the world-largest Folksonomy-style image database, and remove noise images from them. Next, we estimate concept vectors for highly-frequent tags based on only visual features, only tag word features and combined features of both visual and textual features, and compute JS divergence and entropy for three kinds of concept vectors. Finally we estimate hierarchical structures between tags regarding three kinds of concept vectors. In the experiments, we show the obtained hierarchical structure, and it includes interesting relations which sometimes are difficult to be discovered by human. In addition, as its application, we used and evaluated the obtained ontology for query expansion of text-tag-based image search over Flickr. These results indicate that the proposed method is promising and the structure is expected to help image search and some other applications.**

***Keywords*-visual ontology, Folksonomy, PLSA, DAG, social media, Flickr**

## I. Introduction

Recently, a large number of images have become available on the Web. Especially, photo sharing sites such as Flickr and Panoramio gather a huge number of consumer photos uploaded by many people. Such sites called "social media" allow users to add keywords as "tags" to uploaded photos to make easy to search photo databases. Since the vocabulary of tags user can add to photos is not restricted, various and many kinds of words are used as tags. To index images with uncontrolled tags is called "Folksonomy" in contrast to "Taxonomy" where the vocabulary of keywords attached to images is controlled. Folksonomy is common methods to index large-scale social image databases in these days.

Uncontrolled vocabulary in Folksonomy is its advantage and its drawback at the same time. Uncontrolled vocabulary enables anyone to tag photos easily, while keywords tagged to photos vary depending on users. For example, someone might tag a beach photo with "sea", while the other one might tag the same photo with "ocean". In this case, "ocean" should be identified with a subordinate concept of "ocean". To do that, "ontology" is needed, which describes relations between words including a hierarchical structure. Among existing ontologies, "WordNet" is the most common. However, since WordNet is based on the taxonomic hierarchy constructed by hand, it does not describe co-occurrence of word concepts in the photo context.

In this paper, we propose a novel method to construct an ontology which describes hierarchical structure on relations between tags from Folksonomy. The tag structure we extract can be used as an ontology for image database search which reflects both textual and visual relations between tags. In the proposed method, at first, we collect millions of tag-attached-images from Flickr which is the world-largest Folksonomy-style image database, and remove noise images from them. Next, we estimate concept vectors for highly-frequent tags based on only visual features, only tag word features and combined features of both visual and textual features, and compute JS divergence and entropy for three kinds of concept vectors. Finally we estimate hierarchical structures between tags regarding three kinds of concept vectors.

In the experiments, we show the obtained hierarchical structure, and it includes interesting relations which sometimes are difficult to be discovered by human. In addition, as its application, we used and evaluated the obtained ontology for query expansion of text-tag-based image search over Flickr. These results indicate that the proposed method is promising and the structure is expected to help image search and some other applications.

## II. Related Work

It is very challenging work to build a semantic hierarchical-structured database automatically from the Web. In this section, we introduce some related works regarding both text-based and visual-based hierarchical databases on relations between word concepts.

### *A. Text-based Hierarchical Structure*

WordNet is the most common hierarchical word dictionary which can be used as an ontology [1]. WordNet was built manually by the experts, and contains several kinds of semantic relations such as part-of and instance-of between words. There are many studies which employs WordNet as an ontology. However, WordNet ontology is not always effective for image databases. In addition, WordNet does not contain newly-appeared words and proper nouns such as iPhone.

While WordNet is built manually, several automatic methods to build hierarchical structure of word concepts have been proposed so far. Tang et al. [2] built hierarchical structure of word concepts using Folksonomy-based data for movie rating. Tang et al. defined four kinds of distance measures, and used them to build hierarchical structure. Their obtained hierarchical structure represents connections of contents of data rather than the meaning of word concepts.

Plangprasopchok et al. [3] proposed a method to extract hierarchical relations from Flickr for treating images. However, they did not use visual features of images and use only "collect/set" information provided by Flickr. Although their result of the hierarchical structure is quite interesting, their method is not automatic and requires some manual supervision.

### B. Image-based Hierarchical Structure

Bart et al. [4] and Sivic et al. [5] proposed unsupervised methods to build tree structures of word concepts based on the similarity of visual features. However, their experiments were carried out with small hand-made image databases, and they did not use large-scale Folksonomy.

Wu et al. [6] proposed distance measure of word concepts called "Flickr Distance" employing probabilistic latent topics and JS divergence. In our method, we also use JS divergence as distance measure and construct a hierarchical database, while they consider only distances between concepts,

Deng et al. [7] built image datasets over the WordNet hierarchical structure using Amazon Mechanical Turk [8]. They try to build a high-quality image database quality using crowd sourcing, while our objective is automatic construction of a visual ontology from Folksonomy.

## III. Proposed Method

In this section, we describe a novel method to construct a visual ontology which describes both visual and semantical relations between word concepts. The ontology obtained in this paper is represented in a directed acyclic graph (DAG) where nodes and edges correspond to word concepts and relations, respectively. The relation obtained in the ontology includes several kinds of relations such as co-occurrence, part-of and instance-of. In this paper, we do not classify a kind of relations represented by edges, which will be one of our future works.

We estimate three kinds of hierarchical structures represented by DAGs with visual features, tag features, and combined features. The visual DAG is reflected in visual similarity between concepts, while the tag-based DAG is reflected in semantic similarity which is originally estimated from tag co-occurrency. The combined DAG has their intermediate properties taking into account both visual ans semantic similarities.

The proposed method consists of four following procedures:

(1) Generate visual, tag and combined feature vectors, and compute PLSA topic vectors.
(2) Remove noise images for each concepts.
(3) Compute concept vectors, JS divergence between tag concepts and entropy of each tag concepts.
(4) Build a hierarchical structure represented by a Directed Acyclic Graph (DAG).

### A. Representing Images in Three Ways

In the experiment, we collected about 2,000,000 tagged images from Flickr. From all of them, we extracted the following three kinds of features.

The first representation of an image is bag-of-features (BoF) [9] using SIFT descriptor [10]. We set the size of a codebook of BoF representation as 1,000. Each image is expressed in a 1,000 dimensional vector of appearance frequency histogram of representative local features. We call this a visual feature.

The second representation of an image is bag-of-words (BoW) of tags attached to each image. In Flickr, more than two different tags are usually given to each image. Since BoW represents occurrence of tags for each image, its elements are always one or zero. We use 4,345 kinds of tags to build BoW vectors. We call this a tag feature.

The third representation is a combination of the first and the second representation using Probabilistic Latent Semantic Analysis (pLSA)[11]. This representation method is based on Lienhart et al.'s method [12] called Multimodal pLSA. Original pLSA is a generation model contain latent topic variables proposed by Hofmann [11], which is represented by the following equation:

$$P(d_i, w_j) = \sum_{k=1}^{K} P(d_i|z_k)P(w_j|z_k)P(z_k) \qquad (1)$$

where $d_i$, $w_j$ and $z_k$ represent an image, a (visual) word and a latent topic, respectively. We use pLSA as a method for dimension reduction of feature vectors. $P(z_k|d_i)$ for $k = 1..K$ can be regarded as a compressed feature vector of image $d_i$.

Each image $d_i$ is represented by the combination of $P(z_k^{vis}|d_i)$ and $P(z_k^{tag}|d_i)$, where $z_k^{vis}$ and $z_k^{tag}$ are a latent topic about visual vectors and tag vectors, respectively. Because we set the number of topics for both visual vectors and tag vectors as 100 in this experiment, each image is expressed with a 200 dimensional vector. We call this a combined feature.

To obtain concept vectors for three kinds of representation in the third step, we calculate PLSA topic vectors of a combined feature $P(z_k^{comb}|d_i)$ for each image in addition to PLSA topic vectors of a visual feature and a tag feature.

### B. Removal of Noise Images

We use PLSA topic vectors of visual features for removal of noise images. Noise image removal in this paper is based on the idea that firstly we estimate the probability of a given concept over latent topics $P(Concept|z_k)$ by comparing topic distributions of images belonging to the concept and ones not belonging to the concept, and next we evaluate the probability of the concept over each image $P(Concept|Image)$ using estimated with the probability of the given concept over topics $P(Concept|z_k)$.

To obtain the probability of whether image $d_i$ belongs to a certain concept $Concept$ $P(Concept|d_i)$, at first we calculate $P(Concept|z_k)$ which represents a ratio of images belonging to the given concept $Concept$ regarding the latent topic $z_k$. $P(Concept|z_k)$ is calculated as follows:

$$\begin{aligned} P^{z_k}_{Concept} &= \frac{1}{|D_{Concept}|} \sum_{d_C \in D_{Concept}} P(d_C|z_k^{vis}) \\ P^{z_k}_{NotConcept} &= \frac{1}{|D_{NotConcept}|} \sum_{d_{NC} \in q D_{NotConcept}} P(d_{NC}|z_k^{vis}) \\ P(Concept|z_k) &= \frac{P^{z_k}_{Concept}}{P^{z_k}_{Concept} + P^{z_k}_{NotConcept}} \end{aligned} \qquad (2)$$

where $D_{Concept}$ and $D_{NotConcept}$ represent a set of images belonging to the given concept $Concept$ and a set of images not belonging to the given concept, respectively.

Given $P(Concept|z_k)$ and $P(z_k|d_i^{vis})$, the probability of whether image $d_i$ belongs to the concept $Concept$ $P(Concept|d_i)$ is calculated as follows:

$$P(Concept|d_i) = \sum_{k=1}^{K} P(Concept|z_k) P(z_k|d_i) \qquad (3)$$

We sort images which have $Concept$ as tags in the descending order of $P(Concept|d_i)$, and we use only top 100 images for each concept for calculating a concept vector removing the rest of images as noise images.

### *C. Representing Concepts and Extracting Relations between Concepts*

We represent a concept vector of each concept associated with a tag word as an average of PLSA topic vectors of the images belonging to the concept in the following equations:

$$P(z|Concept) = \frac{\sum_{d \in Concept} p(z|d)}{|Concept|} \qquad (4)$$

As $P(z|d)$, we can use any PLSA topic vectors of visual features, tag features or combined features.

We use JS divergence to estimate distance between concept P, Q (Eq.(5), Eq.(6)). JS divergence (Jensen-Shannon divergence) is a distance measure between two probability distributions. JS divergence can be regarded as a symmetricalized version of KL divergence (Kullback-Leibler divergence). If two concept vectors are similar to each other, the value of JS divergence will be small.

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \qquad (5)$$

$$\begin{aligned} D_{JS}(P||Q) &= \frac{D_{KL}(P||(P/2 + Q/2))}{2} \\ &+ \frac{D_{KL}(Q||(Q/2 + P/2))}{2} \end{aligned} \qquad (6)$$

In addition to JS divergence, we calculate entropy of each concept vector which is needed by the next step to build hierarchical structure (Eq.(7)). Basically, high entropy concepts are distributed over broad topics, and low entropy concepts are concentrated on specific topics.

$$H(P) = -\sum_{z \in Z} P(z|Concept) \log(P(z|Concept)) \qquad (7)$$

### *D. Building of Hierarchical Structure*

As a method to build a hierarchical structure from Folksonomy tags, we adopt Eda et al's method [13] which employs a Directed Acyclic Graph (DAG). Although it is originally a method for tag features, it is applicable for visual features and combined features. Since a DAG is a graph structure where a node can have multiple parents in contrast to trees that allow each node to have only one parent, DAG representation is appropriate as a representation of Folksonomy-based visual ontology. Moreover, since a DAG has no closed chains, we follow edges of DAG from a broader concept to a subordinate concept easily.

Given distance between concepts and entropy of each concept, we can build hierarchical structure of concepts represented by a DAG according to the following procedures:

For each concept $C$,

(1) Select $k$ concepts at most to which the distance from $C$ is less than $threshold$.
(2) Compare entropy values of $C$ and selected concepts, and classify the concepts with smaller entropy as child nodes and the concepts with larger entropy as parent nodes.

In this paper, we set $k$ as 20, $threshold$ as 0.05, 0.15 and 0.10 for visual features, tag features and combined features, respectively.

## IV. Experiments

In the experiments, we collected about 2,000,000 tagged images from Flickr, selected the top 5,000 frequent tags as "concepts" for building a concept hierarchy, and built a hierarchical structure by the proposed method. At first, we evaluate removal of noise images, and then we show part of obtained visual ontology.

### *A. Results of Noise Removal*

Fig.1 shows comparison on accuracy rate of randomly-selected 100 images before noise removal and top 100 images regarding $P(Concept|d_i)$ for 10 concepts after noise removal.

The average accuracy rate of randomly-selected images is 70%, while the average accuracy rate of images selected by proposal noise removal method is 92%. This shows that noise image removal worked well. Fig.2 shows parts of the dataset after the noise image removal.

### *B. Hierarchical Structures of Noun Concepts*

We obtained hierarchical structures about 2,657 concepts using 2,000,000 tagged images. Here, we explain part of obtained DAGs regarding some concepts. At first, we show the DAG which starts from "sand" node. Note that the DAGs shown here are a part of the DAGs extracted for explanation due to space limitation.

Fig.3 shows the hierarchical structure under "sand" node estimated by visual features. This structure includes "water" and "sea" which are expected to be related to "sand" as subordinate concepts of "sand". However, it is not natural for human that "water" is regarded as the subordinate concept of "lake". Fig.6 shows the top 50 images of "sand", "lake" and "water" regarding $P(Concept|Image)$. These three

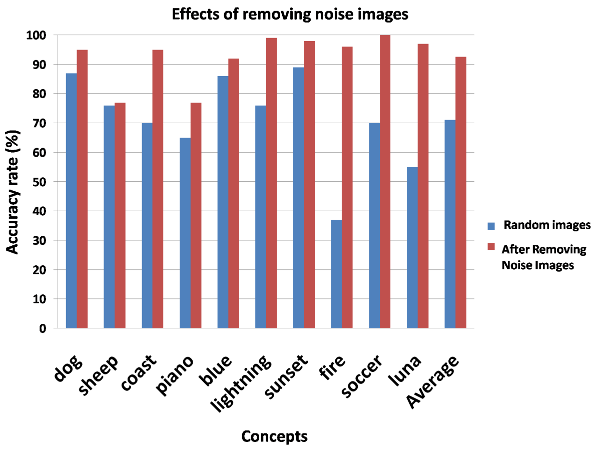

Figure 1. Results of noise removal.

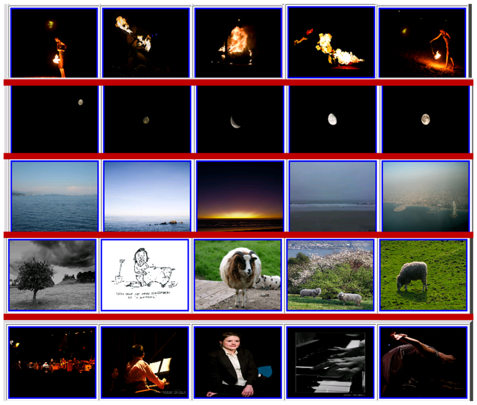

Figure 2. Sample images after noise removal for "fire", "luna", "coast", "sheep" and "piano".

concepts are similar to each other visually. Since "water" images are relatively less various among them, "water" becomes a subordinate concept of "sand" and "lake" in terms of visual features.

Fig.4 shows the hierarchical structure under "sand" estimated by tag features. Tags frequently-used tend to become superordinate concepts. Fig.4 shows the subordinate structure that starts from "sand" estimated by combined features, which can be regarded as being taken into account both visual relations and semantic tag relations between concepts.

Although "beach" does not appear in the DAG of tag features (Fig.4), it appears in the DAGs of both visual features and combined features. On the other hand, "water" which appears in the visual DAG does not appear in the combined DAG, since relation between "water" and "sand" is not so strong regarding the tag-based structure. In addition, although "lake" appears in the visual DAG (Fig.3), the semantic relationship between "sand" and "lake" is not so strong. In this sense, disappearance of "lake" in the combined DAG can be regarded as a reasonable result.

Next, we explain about the DAG started from the "mountain" node shown in Fig.7. Fig.7 shows some representative images of "mountain". In the DAG of "mountain" built

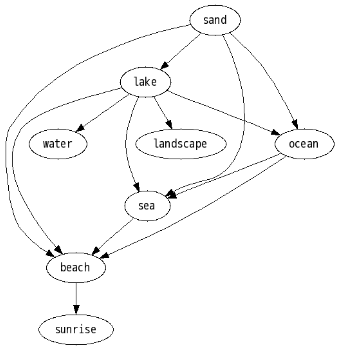

Figure 3. The DAG under "sand" node by visual features

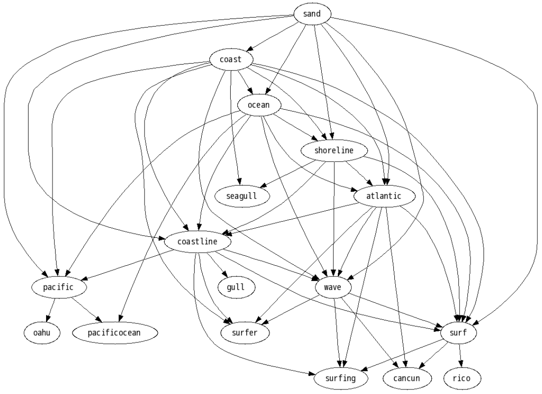

Figure 4. The DAG under "sand" node by tag features

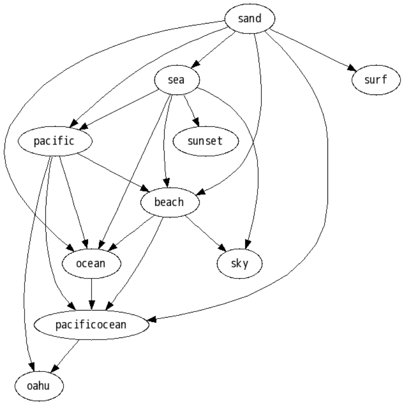

Figure 5. The DAG under "sand" node by combined features

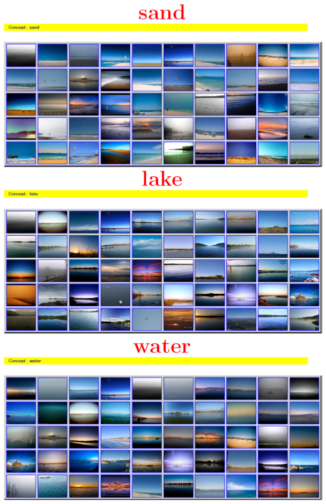

Figure 6. Top 50 images of "sand", "lake" and "water"

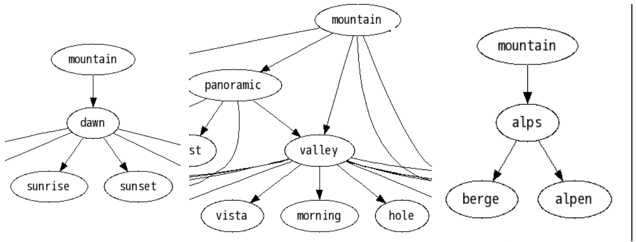

Figure 7. The DAG under “mountain” node (by visual features (on the left), by tag features (on the center) and by combined features (on the right))

by visual features, “mountain” have subordinate concepts “dawn”, “sunrise” and “sunset”.

Fig.9 shows representative images of “dawn”. These images depict that the sun rises from or go down to the horizon or the mountain, and they resemble “mountain” images very much visually. Since most of “dawn” images represents horizons and are similar to each other, the entropy of “dawn” becomes low. Therefore, the “dawn” node was placed as a subordinate node of the “mountain” node.

Meanwhile, in the hierarchical structure of “mountain” built using tag features, “mountain” have subordinate concepts “valley”, “vista”, ”panoramic” and “morning”. These concepts resemble the “mountain” semantically, but these do not always resemble “mountain” visually (Fig.10).

Furthermore, in the hierarchical structure of “mountain” built using combined features, subordinate concepts “berge”, “alps” and “alpen” are extracted which do not appear in the above-mentioned hierarchical structures. These concepts resemble the “mountain” semantically and visually (Fig.11). We expect that this ontology enables us to search for images tagged with only “alps” by the query keyword “mountain”.

Figure 8. Representative images of “mountain”

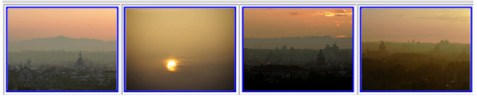

Figure 9. Representative images of “dawn”

Figure 10. Representative images of “vista”

Figure 11. Representative images of “alps”

As the result related to proper nouns, we explain the hierarchical structures above the “Pontevedra” node. “Pontevedra” is a name of the city of the northwest of Spain. In the result by tag features (Fig.13), broader concepts such as “Galiza” and “acoruña” appeared. Note that “Galiza” is the state which “Pontevedra” belongs to. “acoruña” means “A Coruña”, the city of the northwest of Spain, which is larger than “Pontevedra”.

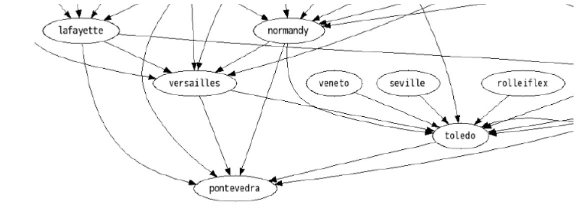

Figure 12. A part of the DAG above “Pontevedra” node by visual features

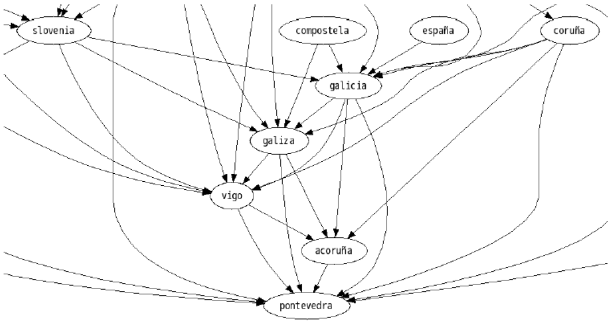

Figure 13. A part of hierarchical structures above “Pontevedra” node by tag features

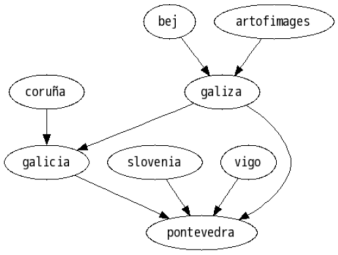

Figure 14. The DAG above “Pontevedra” node by Combined features

Figure 15. Sample images of “Pontevedra”

Figure 16. Sample images of “Toledo”

“Toledo”, the city of the center of Spain, appears in the result only by visual features (Fig.12). Although “Toledo” may resemble “Pontevedra” visually (Fig.15, Fig.16), the semantic relation is weak.

The result by combined features is close to the one by tag features than by visual features (fig:resultPontevedraCombined). It is one of future works to adjust the weight of the balance to combine visual features and tag features automatically.

### C. Hierarchical Structures of Adjective Concepts

In the last, as an example of DAG which starts from an adjective concept, we explain about hierarchical structures under “cute” node. Fig.17 shows the DAG by visual features.

"Statue" has weak semantic relationship with "cute", while "puppy" and "child" are apparently subordinate concepts of "cute". In the result by combined features, "statue" disappears and the DAG was improved. However, the subordinate node "puppy" also disappeared.

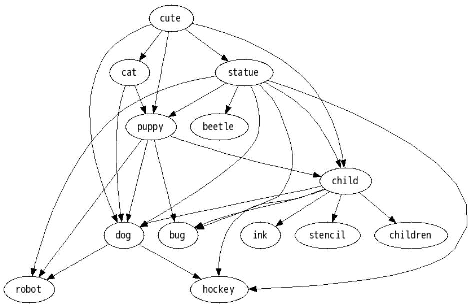

Figure 17. The DAG under "cute" node by visual features

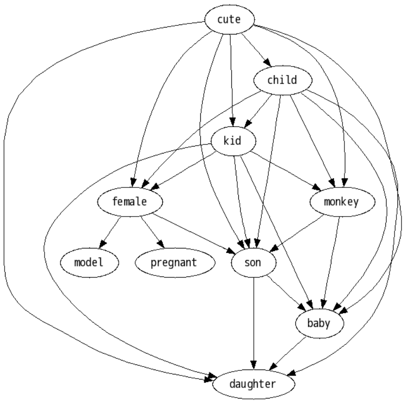

Figure 18. The DAG under "cute" node by tag features

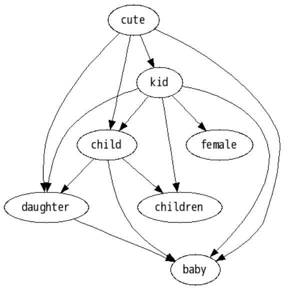

Figure 19. The DAG under "cute" node by combined features

### D. Query Expansion Using Ontology By Combined features

As a simple experiment to evaluate effectiveness of our ontology, we made an experiment on query expansion of text-tag-based image search over Flickr, which is possible one of applications of the obtained ontology.

We evaluated the number of relevant images among the top 10 returned images from Flickr with search by original keywords and expanded keyword sets. As expanded methods, we tried to add one of three subordinate words with the original word and to add one of three superordinate words with the original one. We sent Flickr the expanded query word sets with "AND search" by changing additional words three times for each words regarding both sub- and superordinate words, while we sent Flickr the original word to get the baseline result. We compared the result by the original word with the results by the subordinate words and the ones by the subordinate words.

We evaluated how many irrelevant images we found among the top ten images returned from Flickr. Tab.I shows the result of original and query expansion search. By using subordinate concepts or superordinate words as additional query words, search results were improved.

Table I
THE RESULTS OF QUERY EXPANSION. THE AVERAGE NUMBER OF IRRELEVANT IMAGES IN THE TOP 10 IMAGES RETURNED FROM FLICKR.

| | the average number of irrelevant images |
|---|---|
| only original query word | 2.4 |
| with subordinate word | 1.9 |
| with superordinate word | 1.9 |

## V. CONCLUSION

In this paper, we proposed a novel method to build a large-scale visual ontology automatically. In the experiments, we constructed three kinds of ontologies based on visual, tag and combined features, and we confirmed that the ontology built from combined features which took into account both visual similarity and semantical similarity was more natural that the two other ontologies.

As future work, we explore applications of the obtained ontology. One of the possible applications is query expansion. We plan to study how to incorporate a visual ontology into a text-based image search system effectively.

## REFERENCES

[1] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, The MIT Press, 2000.

[2] J. Tang, H. Leung, Q. Luo, D. Chen, and J. Gong, "Towards ontology learning from folksonomies," in *Proc. of International Joint Conferences on Artificial Intelligence*, 2009.

[3] A. Plangprasopchok and K. Lerman, "Constructing folksonomies from user-specified relations on Flickr," in *Proc. of the 18th International Conference on World Wide Web*, 2009, pp. 781–790.

[4] E. Bart, I. Porteous, P. Perona, and M. Welling, "Unsupervised learning of visual taxonomies," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[5] J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, and A.A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[6] L. Wu, X. S. Hua, N. Yu, W. Y. Ma, and S. Li, "Flickr distance," in *Proc. of ACM International Conference Multimedia*, 2008.

[7] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2009, pp. 710–719.

[8] Amazon Mechanical Turk, *http://aws.amazon.com/mturk/*.

[9] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, 2004, pp. 59–74.

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

[11] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 43, pp. 177–196, 2001.

[12] R. Lienhart, S. Romberg, and E. Horster, "Multilayer pLSA for multimodal image retrieval," in *Proc. of ACM International Conference on Image and Video Retrieval*, 2009.

[13] T. Eda, M. Yoshikawa, T. Uchiyama, and T. Uchiyama, "The Effectiveness of Latent Semantic Analysis for Building Up a Bottom-up Taxonomy from Folksonomy Tags," *World Wide Web*, vol. 12, no. 4, pp. 421–440, 2009.