IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 7, 2017, accepted December 13, 2017, date of publication December 22, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2786218

A Language-Independent Ontology Construction Method Using Tagged Images in Folksonomy

SHOTA HAMANO, (Student Member, IEEE), TAKAHIRO OGAWA, (Member, IEEE), AND MIKI HASEYAMA, (Senior Member, IEEE)
Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Shota Hamano (hamano@lmd.ist.hokudai.ac.jp)

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP17H01744.

ABSTRACT This paper presents a language-independent ontology (LION) construction method that uses tagged images in an image folksonomy. Existing multilingual frameworks that construct an ontology deal with concepts translated on the basis of parallel corpora, which are not always available; however, the proposed method enables LION construction without parallel corpora by using visual features extracted from tagged images as the alternative. In the proposed method, visual similarities in tagged images are leveraged to aggregate synonymous concepts across languages. The aggregated concepts take on intrinsic semantics of themselves, while they also hold distinct characteristics in different languages. Then relationships between concepts are extracted on the basis of visual and textual features. The proposed method constructs a LION whose nodes and edges correspond to the aggregated concepts and relationships between them, respectively. The LION enables successful image retrieval across languages since each of the aggregated concepts can be referred to in different languages. Consequently, the proposed method removes the language barriers by providing an easy way to access a broader range of tagged images for users in the folksonomy, regardless of the language they use.

INDEX TERMS Concept relationship, hierarchical structure, image folksonomy, image retrieval, synonymous concept, tagged image, tag refinement.

I. INTRODUCTION

An ontology represents concepts and the semantic relationships between them in an organized manner. Ontologies are consistent with human perception and thus achieve remarkable success in various fields such as natural language processing [1], [2], object recognition [3], [4] and information retrieval [5], [6]. The hierarchical structure of concepts provides intuitive information about concepts and enables users to obtain an overview of the concepts.

Lexical ontologies such as WordNet [7], SUMO [8] and LSCOM [9] are manually constructed by continuous and intensive efforts by experts. Although their quality is high, maintenance of these ontologies is time-consuming and laborious work. Thus, manually constructed ontologies often fail to catch up with current trends in the world, where a number of newborn concepts are generated day after day, especially in social networking services such as Flickr[1] and Twitter.[2] To address this problem, various methods for automated construction of ontologies have been proposed [10]–[16]. Concepts are generally identified in documents or tags assigned to images in a folksonomy. Traditionally, word co-occurrence in textual documents is used to detect hypernymous concepts [10]. Even hash tags are collaboratively used with ontologies for trend analysis on Twitter [14]. Visual features are also utilized for detection of salient objects in images [11], estimation of concepts corresponding to given images [12] or image retrieval [15], [16].

However, these methods focus only on tags in one language, which is mostly English. This means that tags in other languages are treated as noisy tags to be removed in preprocessing, though the number of tags in other languages is not negligible. According to the analysis in [17], 170 non-English languages are used in the Flickr folksonomy, and they account for 25% of the total number of tags. Previous methods that deal with only English language are of no benefit to non-English speaking users, who are a relatively large proportion of users in the Flickr folksonomy. Also, tags in these languages are discarded despite their potential for contribution to accurate ontology construction. Retrieval

[1]https://www.flickr.com/
[2]https://twitter.com/

results enriched with multilingual sources of tagged images reportedly enhance retrieval performance and users' satisfaction [18], [19]. Thus, multilingual frameworks for ontology construction are absolutely necessary.

Following the first reported method [20], several multilingual ontology construction methods have been proposed. In these methods, concepts in different languages are matched manually on the basis of parallel corpora. For example, BabelNet [21] provides a multilingual ontology with concepts in various languages translated from English concepts in WordNet. The constructed multilingual ontologies are applied to specific domains such as medicine [22]–[24], health [25], [26] and literature [27]. Also, multilingual ontologies contribute to sentiment analysis [19], [28]–[30], named entity resolution [31], [32] and information retrieval [33], [34]. These methods aim to capture cultural differences by considering concepts in various languages. There are other methods for multilingual ontology construction [35]–[39] that require large parallel corpora to match synonymous concepts across languages. Another approach to match synonymous concepts across languages includes multilingual concept representation [40]–[42]. These methods learn a common space where synonymous concepts in different languages are located close to each other. However, the learning process in these methods still requires large parallel corpora to obtain effective concept representation.

As stated above, existing methods that construct a multilingual ontology are entirely based on rich language resources. However, this is a crucial problem since such corpora are not always available. The amount and quality of non-English language resources are significantly limited compared to English [43], [44]. Hence, multilingual ontologies should be constructed without resort to parallel corpora to overcome varied situations according to languages. Also, mining concepts universal among languages has not been addressed yet, while previous methods focus on observing cultural differences. Therefore, we aimed to construct an ontology that is universal among languages, *i.e.*, a **L**anguage-**I**ndependent **ON**tology (LION) without parallel corpora.

In this paper, we propose a LION construction method that uses tagged images in an image folksonomy as the alternative to parallel corpora. In the proposed method, visual similarities in tagged images are leveraged to aggregate synonymous concepts across languages. We use tagged images as the alternative to parallel corpora since images with synonymous tags are similar regardless of the languages in which tags are represented [45]–[49]. The aggregated concepts take on intrinsic semantics of themselves, while they also hold distinct characteristics in different languages. Then we extract relationships between concepts on the basis of visual and textual features. We finally construct a directed acyclic graph (DAG) as a LION whose nodes and edges correspond to the aggregated concepts and relationships between them, respectively. The LION enables successful image retrieval across languages since each of the aggregated concepts is referred to in any language. Consequently, the proposed method takes away the language barriers by providing an easy way to access a broader range of tagged images for users in the folksonomy, regardless of the language they use. The main contribution of this paper is derivation of a framework for LION construction without parallel corpora. In this paper, we make the most use of the Flickr folksonomy, which hosts a huge number of tagged images in various languages.

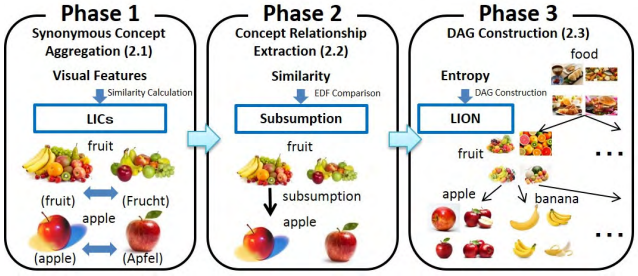


FIGURE 1. Overview of the proposed method for LION construction. Phase 1: The proposed method first uses visual features to aggregate synonymous concepts across languages. "(Fruit)" and "(apple)" represent concepts in English and "(Frucht)" and "(Apfel)" represent concepts in German. "Fruit" and "apple" represent LICs identified on the basis of visual similarities. These LICs are to share the characteristics of the concepts in English and German. Phase 2: Similarities between LICs are calculated on the basis of visual and textual modalities. Comparing EDFs of these similarities, the more effective one is selected for subsumption score calculation. Phase 3: Finally, a LION is constructed on the basis of entropy, which quantifies the semantic broadness of each LIC. Since each node holds the characteristics of the corresponding concepts in different languages, the semantic relationships between concepts in each language are also encoded in the constructed LION.

In the preliminary work [50], we have already proposed a method that extracts relationships between language-independent concepts (LICs) and applied it to tag refinement. That method [50] has been improved by considering the correlations between textual features in different languages. Additional experiments in image retrieval were conducted to thoroughly analyze the effectiveness of the proposed method.

II. LANGUAGE-INDEPENDENT ONTOLOGY CONSTRUCTION USING TAGGED IMAGES

In this section, we explain the proposed method that constructs a LION using tagged images in the Flickr folksonomy. As shown in Fig. 1, the proposed method consists of the following three phases.

Synonymous Concept Aggregation

The proposed method aggregates synonymous concepts across languages to identify LICs based on visual features extracted from tagged images. This approach is different from those in previous methods in that no parallel corpora are required. The proposed method makes the most use of the Flickr folksonomy to enrich concept representation considering distinct characteristics of concepts in different languages. This is the main contribution of this paper in that we make the most use of the Flickr folksonomy, which hosts a huge number of images with tags in various languages.

Concept Relationship Extraction
The proposed method explores relationships between LICs identified by synonymous concept aggregation. In this phase, we obtain language-independent textual features via canonical correlation analysis (CCA) [51] and utilize them in addition to visual features. The derived features are effective in concept relationship extraction since these features are obtained by maximizing the correlation between concepts in different languages. The proposed method properly compares the modalities to select the more effective one in similarity calculation between concepts. Concept relationships are extracted on the basis of the similarities.

DAG Construction
The proposed method constructs a DAG as a LION whose nodes and edges correspond to LICs and relationships between them, respectively. Semantic broadness of concepts is taken into consideration in accurate ontology construction. Concepts in the constructed ontology take on intrinsic semantics of themselves, while they also hold distinct characteristics in different languages. This is what makes the ontology universal across languages, which removes the existing language barriers.

Following these phases, the proposed method effectively utilizes tagged images in the Flickr folksonomy for LION construction. In this paper, we focus on exploration of relationships between concepts in two languages, $\texttt{LANG}^{(1)}$ and $\texttt{LANG}^{(2)}$, and define $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ as the sets of concepts in $\texttt{LANG}^{(1)}$ and $\texttt{LANG}^{(2)}$, respectively.

A. SYNONYMOUS CONCEPT AGGREGATION

In the first phase, we aggregate synonymous concepts across languages based on visual similarities in tagged images. Different from previous approaches, the proposed method leverages visual features to match synonymous concepts in different languages. We use visual similarities as the alternative to parallel corpora since images with synonymous tags are considered to be represented by similar visual features.

Toward representation of concepts $c_i^{(l)}$ ($l = 1, 2$; $i = 1, 2, \ldots, |\mathcal{C}^{(l)}|$), we extract visual features $\boldsymbol{\phi}_{i,n}^{(l)}$ ($n = 1, 2, \ldots, N_i^{(l)}$; $N_i^{(l)}$ being the number of images with a tag corresponding to $c_i^{(l)}$) from tagged images, where $|\cdot|$ denotes the cardinality of the set. Note that l corresponds to $\texttt{LANG}^{(l)}$. In this paper, we used 4096-dimensional features extracted from the fc6 layer in AlexNet [52]. We then apply locality-constrained linear coding (LLC) [53] to obtain new features $\boldsymbol{V}_i^{(l)} = [\boldsymbol{v}_{i,1}^{(l)} \boldsymbol{v}_{i,2}^{(l)} \cdots \boldsymbol{v}_{i,N_i^{(l)}}^{(l)}]$ as follows:

$$\min_{\boldsymbol{V}^{(l)}} \sum_{n=1}^{N_i^{(l)}} \left(\|\boldsymbol{\phi}_{i,n}^{(l)} - \boldsymbol{B}\boldsymbol{v}_{i,n}^{(l)}\| + \lambda \|\boldsymbol{d}_{i,n}^{(l)} \odot \boldsymbol{v}_{i,n}^{(l)}\| \right)$$
$$\text{s.t. } \mathbf{1}^\top \boldsymbol{v}_{i,n}^{(l)} = 1, \quad \forall n, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\odot$ is an operator for the Hadamard product. The codebook $\boldsymbol{B}$ is generated by applying k-means clustering [54] to visual features $\boldsymbol{\phi}_{i,n}^{(l)}$, and

$$d_{i,n,k}^{(l)} = \exp\left(\frac{\|\boldsymbol{\phi}_{i,n}^{(l)} - \boldsymbol{b}_k\|}{\sigma} \right), \tag{2}$$

where σ is a parameter to control the locality. Finally, we obtain representative features $\boldsymbol{v}_i^{(l)}$ for $c_i^{(l)}$ by applying max-pooling to visual features $\boldsymbol{v}_{i,n}^{(l)}$.

Based on these features, we calculate visual similarities $s_v(c_i^{(1)}, c_j^{(2)})$ as follows:

$$s_v(c_i^{(1)}, c_j^{(2)}) = \frac{\boldsymbol{v}_i^{(1)\top} \boldsymbol{v}_j^{(2)}}{\|\boldsymbol{v}_i^{(1)}\| \, \|\boldsymbol{v}_j^{(2)}\|}. \tag{3}$$

Note that any similarity metric can be used as an alternative to Eq. (3) in the proposed framework. In this paper, we adopt the cosine metric due to its simplicity and effectiveness, which is shown later in the experiments. Finally, we match $c_j^{(2)}$ to $c_i^{(1)}$ that satisfies

$$\hat{c}_i^{(2)} = \arg \max_{c_j^{(2)}} s_v(c_i^{(1)}, c_j^{(2)}). \tag{4}$$

The proposed method aggregates concepts $c_i^{(1)}$ and $\hat{c}_i^{(2)}$ to identify LICs c_i, and $\mathcal{C}$ is defined as the set of LICs identified in this phase. These LICs c_i are considered to be more universal than a concept in one language, while distinct characteristics of $c_i^{(1)}$ and $\hat{c}_i^{(2)}$ are retained.

We effectively utilize visual features in synonymous concept aggregation across languages without parallel corpora. The aggregated LICs are interpreted as enriched representation of distinct characteristics of concepts in different languages. Consequently, this phase enables concept relationship extraction across languages, as described in the following subsection, which plays a crucial role in LION construction. Since a number of newborn concepts being generated daily, parallel corpora hardly keep up with the trend in the folksonomy. Therefore, we focus on tagged images, with which concepts are represented. The proposed framework best utilizes the Flickr folksonomy, which is affluent in tagged images, by leveraging them as an alternative to parallel corpora in synonymous concept aggregation.

B. CONCEPT RELATIONSHIP EXTRACTION

We extract textual features $\boldsymbol{\tau}_i^{(1)}$ and $\hat{\boldsymbol{\tau}}_i^{(2)}$ for concepts corresponding to concepts $c_i^{(1)}$ and $\hat{c}_i^{(2)}$, respectively. In this paper, we adopt the GloVe model [55] trained using monolingual corpora for textual feature calculation. Textual features for synonymous concepts are not necessarily similar since the concept representations are learned via different corpora. In the proposed method, we consider correlations between languages by applying CCA to the textual features. CCA is a technique that explores correlations between two sets of features, and it has been reported to be effective for multilingual

concept representation [42]. We obtain projection matrices $\boldsymbol{U}^{(1)}$ and $\boldsymbol{U}^{(2)}$ by solving the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)}} \quad & \boldsymbol{U}^{(1)\top}\boldsymbol{T}^{(1)}\boldsymbol{T}^{(2)\top}\boldsymbol{U}^{(2)} \\ \text{s.t. } & \boldsymbol{U}^{(1)\top}\boldsymbol{T}^{(1)}\boldsymbol{T}^{(1)\top}\boldsymbol{U}^{(1)} = \boldsymbol{I}, \\ & \boldsymbol{U}^{(2)\top}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{U}^{(2)} = \boldsymbol{I}, \end{aligned} \tag{5}$$

where $\boldsymbol{T}^{(l)} = [\boldsymbol{\tau}_1^{(l)}\boldsymbol{\tau}_2^{(l)}\dots\boldsymbol{\tau}_{|\mathcal{C}|}^{(l)}]$, and $\boldsymbol{I}$ is the identity matrix. Using the above matrices, we project textual features into the space where the correlations between languages are maximized to obtain canonical textual features as follows:

$$\boldsymbol{t}_i^{(1)} = \boldsymbol{U}^{(1)\top}\boldsymbol{\tau}_i^{(1)}, \quad \boldsymbol{t}_j^{(2)} = \boldsymbol{U}^{(2)\top}\boldsymbol{\tau}_j^{(2)}. \tag{6}$$

In the obtained space, not only semantically similar concepts but also synonymous concepts across languages are located close to one another. Therefore, the proposed method accurately extracts relationships between LICs by utilizing canonical textual features.

Now, we represent LICs c_i by $\boldsymbol{v}_i = [\boldsymbol{v}_i^{(1)\top}\hat{\boldsymbol{v}}_i^{(2)\top}]^\top$ and $\boldsymbol{t}_i = [\boldsymbol{t}_i^{(1)\top}\hat{\boldsymbol{t}}_i^{(2)\top}]^\top$, where $\hat{\boldsymbol{t}}_i^{(2)}$ is textual features for $\hat{c}_i^{(2)}$. By concatenating features corresponding to $c_i^{(1)}$ and $\hat{c}_i^{(2)}$, we can effectively represent LICs since they not only capture the universal semantics across languages but also reflect different aspects of the concepts in the two languages.

We then calculate similarities between concepts c_i and c_j based on each modality $m \in \{v, t\}$. In this paper, textual similarities $s_t(c_i, c_j)$ are defined in the same manner as visual similarities:

$$s_t(c_i, c_j) = \frac{\boldsymbol{t}_i^\top \boldsymbol{t}_j}{\|\boldsymbol{t}_i\|\,\|\boldsymbol{t}_j\|}. \tag{7}$$

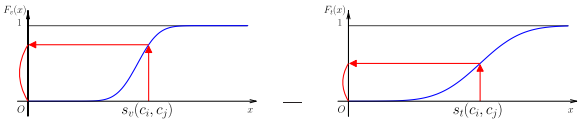
We leverage the more effective modality to determine the similarity between concepts c_i and c_j. Here, we sort $s_m(c_i, c_j)$ in ascending order and denote them as $s_m[h]$ $(h = 1, 2, \dots, H)$. H is equal to $|\mathcal{C}|^2$. Next, we build an empirical distribution function (EDF) [56] $F_m(x)$ defined as

$$F_m(x) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{I}[s_m(h) \le x], \tag{8}$$

where $\mathbb{I}[\cdot]$ is the indicator function, which returns 1 if the condition is satisfied, and 0 otherwise. As shown in Fig. 2, the derived EDF captures the statistical distribution of similarities [57]. The proposed method enables proper comparison of modalities by equalizing occurrence probability of similarities in a constant interval of the longitudinal axis. Based on the EDFs, we can select the modality that is the more effective of the two. The proposed method adaptively calculates similarities between concepts on the basis of the selected modality as follows:

$$s(c_i, c_j) = \max_{m\in\{v,t\}} F_m\left(s_m(c_i, c_j)\right). \tag{9}$$

In this way, we can select an optimal modality for concept relationship extraction since EDFs enable comparison of the effectiveness of the modalities by considering the statistical distribution of similarities [57]. This calculation scheme is reasonable since the modality in which concepts are more similar depends on tagged images. For example, concepts with broad semantics such as "animal" exhibit higher similarity from the viewpoint of textual features since visual features significantly differ according to what kind of animals are in the images.



$$\begin{aligned} s(c_i, c_j) &= \max_{m\in\{v,t\}} F_m\left(s_m(c_i, c_j)\right) \\ &= F_v\left(s_v(c_i, c_j)\right) \end{aligned}$$

FIGURE 2. Example of similarity calculation between concepts based on EDFs. In this example, visual features are selected for similarity calculation.

Then we extract relationships between LICs based on the modality selected above. $s(c_i, c_j)$ is interpreted as a score representing co-occurrent relationships between c_i and c_j [16]. The proposed method extracts concept co-occurrent relationships based on scores in Eq. (9). We also define a subsumption score $p(c_i|c_j)$ to quantify the extent that c_i subsumes c_j using co-occurrent scores as follows:

$$p(c_i|c_j) = \frac{s(c_i, c_j)}{\sum_{c\in\mathcal{C}} s(c_i, c)}. \tag{10}$$

When $p(c_i|c_j)$ is greater than $p(c_j|c_i)$, we regard c_j as an LIC subsumed by c_i. This definition is based on the assumption that c_i can be used wherever c_j is used if c_i subsumes c_j, but not vice versa [10], [16].

In this phase, we adaptively select the more effective modality in similarity calculation to realize collaborative use of multimodal features to extract concept relationships. Accurate extraction of concept relationships contributes to performance improvement in the subsequent ontology construction.

C. DAG CONSTRUCTION

Finally, we construct a DAG as an ontology. A DAG is a directed graph with no cycles whose nodes are allowed to have multiple parents. DAGs are suitable for ontologies since they represent the hierarchical structure of concepts at arbitrary levels of semantic broadness [16], [58]. Therefore, we adopt a DAG for the structure representing an ontology. The DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed based on Algorithm 1. $\mathcal{V}$ and $\mathcal{E}$ are the sets of concept nodes and directed edges, respectively. An edge e_{c_i,c_j} from c_i to c_j indicates a subsumption relationship $c_i \to c_j$. The weight of each edge $w_{c_i\to c_j}$ is defined as the subsumption score $p(c_i|c_j)$. To consider the semantic broadness of each concept, we construct the DAG $\mathcal{G}$ based on the entropy defined as follows:

$$H(c_i) = -\sum_{j=1}^{|\mathcal{C}|} p(c_i|c_j)\log(p(c_i|c_j)). \tag{11}$$

Algorithm 1 DAG construction algorithm $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Input: Concepts c_i with their entropy $H(c_i)$ and weighted relationships $w_{c_i \to c_j}$
Output: A DAG of concepts $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

```
for c_i ∈ C in descending order of entropy do
    V ⇐ c_i
    for c_j ∈ C in descending order of p(c_i|c_j) do
        if H(c_i) ≥ H(c_j) then
            E ⇐ e_{c_i,c_j}, V ⇐ c_j
        end if
    end for
end for
output G
```

The premise here is that a concept with high entropy is expected to have broad semantics, while a concept with low entropy is likely to have specific semantics.

The proposed method realizes LION construction using multilingual sources of tagged images and the hierarchical structure of concepts. Consequently, the derived ontology removes the language barriers to help users obtain desired images across languages.

D. LIMITATION OF THE PROPOSED METHOD

Since the proposed method uses visual features extracted from tagged images to aggregate synonymous concepts across languages, the performance depends on them. With inaccurate aggregation of concepts across languages, the identified LICs affect the subsequent phases for relationship extraction and DAG construction. In fact, there are two cases where it is difficult to apply the proposed method. First, concepts with multiple semantics are not effectively captured. Since we describe each concept by one feature vector, polysemic concepts are not appropriately represented in the constructed LION. Second, images containing many objects affect the quality of semantic representation of concepts. For example, if an image containing "mountain", "sky" and "cloud" with a tag "mountain", visual features extracted from this image may not be appropriate in semantic representation of "mountain". The issue of how to capture multiple semantic of concepts and multiple objects in images will be addressed in our future work.

III. EXPERIMENTAL RESULTS

In this section, we evaluate the quality of the constructed LION to verify the effectiveness of the proposed method. In this experiment, we used Flickr images for the dataset. We collected images by Flickr API[3] in such a way that each image has tags in at least one of English (EN), German (DE), French (FR), Russian (RU), Japanese (JA) and Chinese (ZH), and Korean (KO) languages. We extracted 4096-dimensional features [59] from the fc6 layer in AlexNet for visual features by the deep learning framework Caffe [60].

[3]https://www.flickr.com/services/api/

TABLE 1. Number of images collected from Flickr and training corpora for GloVe.

Language	Images	Vocabulary	Word tokens
EN	30143	1778575	1777497491
DE	29684	1505149	618569591
FR	29694	754100	468980630
RU	29003	595926	328117251
JA	28854	367607	256158273
ZH	25874	137567	229237678
KO	23115	476899	44383960

TABLE 2. Statistics of the constructed ontologies. Concepts and subsumption relationships correspond to nodes and edges in DAGs, respectively. Ontologies derived from two languages by the proposed method have the same number of nodes and edges as the language that has the less concepts.

Language	Concepts	Cooccurrence	Subsumption
EN	20542	1455730	223609
DE	19187	1332485	201119
FR	19143	952107	206597
RU	15006	801175	196998
JA	12431	769979	189721
ZH	10414	641505	164801
KO	8872	502171	123806

Note that the proposed method needs no training process since an off-the-shelf deep model pretrained with a large image dataset is utilized. The number of codes for LLC was set to 1000. For textual features, we independently trained GloVe for each language using monolingual corpora provided by Polyglot project[4] to extract 200-dimensional features. The details of the dataset are shown in Table 1. We constructed LIONs using concepts in two languages $(\mathtt{LANG}^{(1)}, \mathtt{LANG}^{(2)}) \in \{\text{EN, DE, FR, RU, JA, ZH, KO}\} \times (\{\text{EN, DE, FR, RU, JA, ZH, KO}\} \setminus \mathtt{LANG}^{(1)})$, where $\times$ denotes the Cartesian product. Statistics of the derived ontologies are shown in Table 2. A part of LION for EN and JA is shown in Fig. 3. Two subgraphs on "animal" and "plant" with their subsuming concepts are illustrated with the corresponding images. It can be seen that the constructed LION is consistent with human perception. We separately evaluated the quality of synonymous concept aggregation and concept relationship extraction as shown in the following subsections to properly assess the effectiveness of each procedure in the proposed method. In this experiment, we selected 150 non-abstract concepts such as "airplane", "fish" and "road" from $\mathcal{C}^{(1)}$ for a concept set $\mathcal{C}^{(1)}_{\text{test}}$ in the following evaluations.

A. SYNONYMOUS CONCEPT AGGREGATION

We first evaluated the performance of the proposed method (**Ours**) in synonymous concept aggregation across languages. For comparative methods, we adopted the following similarity measures.

[4]https://sites.google.com/site/rmyeid/projects/polyglot

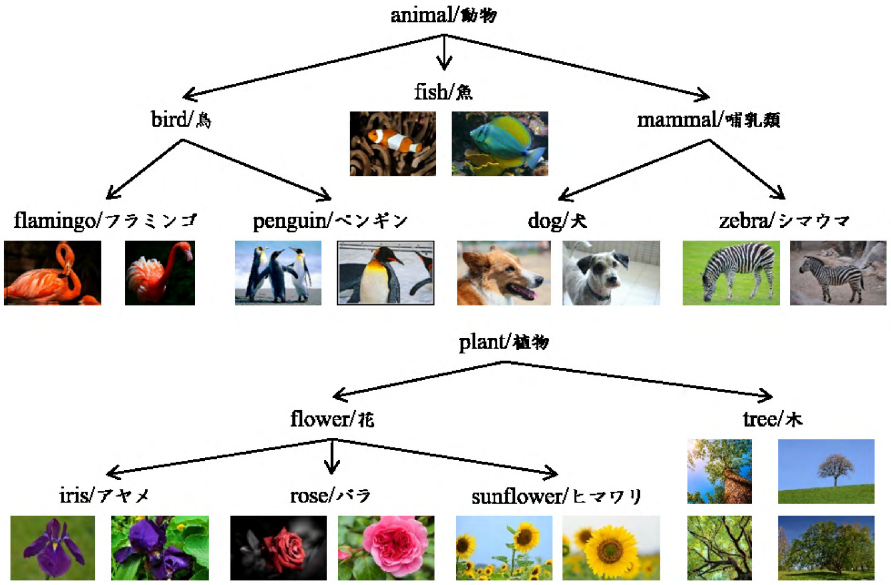


FIGURE 3. Part of the constructed LION for EN and JA. Subgraphs on "animal" and "plant" are illustrated. In this figure, Japanese concepts next to English ones are correctly detected synonymous concepts.

Comparative method 1 (BOF):
This is a method that aggregates synonymous concepts across languages based on similarities defined as follows:

$$s_v(c_i^{(1)}, c_j^{(2)}) = \frac{\tilde{\boldsymbol{v}}_i^{(1)\top} \tilde{\boldsymbol{v}}_j^{(2)}}{\|\tilde{\boldsymbol{v}}_i^{(1)}\| \, \|\tilde{\boldsymbol{v}}_j^{(2)}\|}, \tag{12}$$

where $\tilde{\boldsymbol{v}}_i^{(1)}$ and $\tilde{\boldsymbol{v}}_j^{(2)}$ are respectively bag-of-features (BOF) [61] representation of $c_i^{(1)}$ and $c_j^{(2)}$ based on the AlexNet fc6 features obtained by the k-means clustering technique [54]. The codebook for vector quantization was generated using all of the images in the dataset, and the number of codes in the codebook was set to 1000.

Comparative method 2 (Mean):
This is a method that aggregates synonymous concepts across languages based on similarities defined as follows:

$$s_v(c_i^{(1)}, c_j^{(2)}) = \frac{\bar{\boldsymbol{v}}_i^{(1)\top} \bar{\boldsymbol{v}}_j^{(2)}}{\|\bar{\boldsymbol{v}}_i^{(1)}\| \, \|\bar{\boldsymbol{v}}_j^{(2)}\|}, \tag{13}$$

where $\bar{\boldsymbol{v}}_i^{(l)}$ is

$$\bar{\boldsymbol{v}}_i^{(l)} = \frac{1}{N_i^{(l)}} \sum_{n=1}^{N_i^{(l)}} \boldsymbol{\phi}_{i,n}^{(l)}. \tag{14}$$

For a criterion to evaluate the quality of synonymous concept aggregation, we adopted accuracy defined as

$$\text{Accuracy} = \frac{1}{|\mathcal{C}_{\text{test}}^{(1)}|} \sum_{i=1}^{|\mathcal{C}_{\text{test}}^{(1)}|} \mathbb{I}[c_{\text{test},i}^{(1)} = c_{\text{GT},i}^{(2)}], \tag{15}$$

TABLE 3. Accuracy of synonymous concept aggregation across languages. Average values over languages are presented.

LANG$^{(1)}$	Ours	BOF	Mean
EN	**0.633**	0.450	0.553
DE	**0.593**	0.543	0.570
FR	**0.602**	0.540	0.583
RU	**0.594**	0.513	0.547
JA	**0.607**	0.507	0.553
ZH	**0.613**	0.533	0.560
KO	**0.597**	0.507	0.560

where $c_{\text{GT},i}^{(2)} \in \mathcal{C}^{(2)}$ is the ground truth concept for $c_{\text{test},i}^{(1)} \in \mathcal{C}_{\text{test}}^{(1)}$. The ground truth for synonymous concept aggregation was obtained by Google Translate.[5]

Aggregation results are shown in Table 3. For all of the languages, the proposed method achieved higher accuracy than the comparative methods. This demonstrates that the proposed method is not only the simplest but also the most effective of all of the synonymous concept aggregation methods.

B. CONCEPT RELATIONSHIP EXTRACTION

We evaluated the quality of concept relationship extraction. For comparison, we adopted the state-of-the-art method (**FBVO**) proposed in [16]. **FBVO** is a method that extracts concept relationships based on collaborative use of visual and textual features by linear combination of similarities. In this method, concept subsumption relationships are extracted using similarities defined as follows:

$$s(c_i, c_j) = \lambda s_v(c_i, c_j) + (1 - \lambda) s_t(c_i, c_j), \tag{16}$$

[5]https://translate.google.com/

where $\lambda \in [0, 1]$ is a weighting parameter to control the influence of each modality. We adopted the optimal λ that achieved the highest performance. **FBVO** is different from **Ours** in that the weight for each modality is common among all concepts. We also adopted a supervised learning-based method (**JOCL**) proposed in [13], which takes an approach different from ours. **JOCL** uses the confidence scores of each concept obtained when support vector machines (SVMs) [62] classify images corresponding each concept as features for concept relationship extraction. We evaluated these methods by precision (P) and recall (R) defined as follows:

$$\mathrm{P@}K = \frac{1}{|\mathcal{C}_{\mathrm{test}}|} \sum_{c \in \mathcal{C}_{\mathrm{test}}} \frac{|U_K(c) \cap U_{\mathrm{GT}}(c)|}{K}, \tag{17}$$

$$\mathrm{R@}K = \frac{1}{|\mathcal{C}_{\mathrm{test}}|} \sum_{c \in \mathcal{C}_{\mathrm{test}}} \frac{|U_K(c) \cap U_{\mathrm{GT}}(c)|}{|U_{\mathrm{GT}}(c)|}, \tag{18}$$

where $U_K(c)$ and $U_{\mathrm{GT}}(c)$ denote the set of top K extracted subsumption relationships of concept c obtained by the constructed ontology and the ground truth relationships generated by WordNet, respectively. In this experiment, we regarded synsets in WordNet as LICs, not mere English words.

The results are shown in Table 4. As shown in this table, **Ours** and **FBVO** outperformed **JOCL**. Images in folksonomies, in general, are assigned irrelevant tags or missing relevant tags. From the view of **JOCL**, SVMs are trained with "mislabeled" samples, which degrade classification performance. Accordingly, it is considered that **JOCL** suffers from its unsatisfactory performance when applied to an image folksonomy with noisy tags. Besides, this table shows that similarity calculation based on EDFs, adopted in **Ours**, is more effective than that based on linear combination in **FBVO**. We further confirmed that synonymous concept aggregation prior to concept relationship extraction leads to an improvement in performance by leveraging multilingual sources of annotated images.

TABLE 4. Precision and recall on concept relationship extraction at different position $K \in \{10, 20\}$. The postfix "Multi" indicates that concept relationships were extracted after synonymous concept aggregation. Average values over languages are presented for $\mathrm{LANG}^{(1)}$-Multi. (a) Precision. (b) Recall.

(a)

	Ours		FBVO		JOCL	
$\mathrm{LANG}^{(1)}$	P@10	P@20	P@10	P@20	P@10	P@20
EN	0.292	0.306	0.272	0.266	0.177	0.204
EN-Multi	**0.300**	**0.308**	0.278	0.269	0.176	0.198
DE	0.301	0.312	0.264	0.263	0.175	0.198
DE-Multi	**0.303**	**0.313**	0.263	0.267	0.183	0.204
FR	0.295	0.305	0.255	0.257	0.172	0.190
FR-Multi	**0.304**	**0.306**	0.258	0.264	0.175	0.206
RU	0.285	0.297	0.262	0.279	0.168	0.188
RU-Multi	**0.298**	**0.310**	0.270	0.283	0.170	0.196
JA	0.277	0.295	0.266	0.285	0.181	0.204
JA-Multi	**0.281**	**0.298**	0.268	0.283	0.192	0.210
ZH	0.272	**0.268**	0.269	0.251	0.178	0.196
ZH-Multi	**0.288**	0.266	0.272	0.248	0.183	0.213
KO	**0.290**	0.300	0.270	0.284	0.168	0.194
KO-Multi	0.288	**0.307**	0.270	0.288	0.176	0.205

(b)

	Ours		FBVO		JOCL	
$\mathrm{LANG}^{(1)}$	R@10	R@20	R@10	R@20	R@10	R@20
EN	0.459	0.511	0.459	0.463	0.387	0.472
EN-Multi	**0.461**	**0.519**	0.457	0.467	0.411	0.493
DE	0.485	0.508	0.472	0.491	0.392	0.468
DE-Multi	**0.491**	**0.510**	0.476	0.488	0.402	0.480
FR	0.469	0.502	0.457	0.493	0.388	0.460
FR-Multi	**0.473**	**0.510**	0.461	0.500	0.399	0.487
RU	0.467	0.508	0.423	0.470	0.395	0.466
RU-Multi	**0.488**	**0.516**	0.459	0.488	0.414	0.478
JA	**0.488**	**0.516**	0.457	0.467	0.398	0.462
JA-Multi	0.479	0.508	0.451	0.469	0.405	0.469
ZH	0.453	0.499	0.453	0.478	0.388	0.461
ZH-Multi	**0.459**	**0.502**	0.451	0.482	0.396	0.481
KO	0.456	0.497	0.443	0.470	0.378	0.448
KO-Multi	**0.474**	**0.512**	0.455	0.488	0.401	0.460

IV. APPLICATIONS

In this section, we verify the applicability of the LION to tag refinement and tag-based image retrieval.

A. TAG REFINEMENT

An image folksonomy such as Flickr is a framework where users can freely assign tags to their uploaded images [63], [64]. Due to its very nature, the folksonomy suffers from noisy and incomplete tags [65]. This situation arises because users tend to assign tags to images based on their knowledge and experience, which often fall short of high consistency to describe the semantic contents of the images. These tags degrade tag-based retrieval performance; noisy tags lower precision, and incomplete tags lower recall. Solutions to this problem include tag refinement techniques [66]–[78]. Tag refinement methods remove noisy tags from images and complement images with missing relevant tags to improve performance in tag-based applications such as image retrieval. Previous tag refinement methods use visual features and textual features to consider the relationships between tags [66], [67] or between images [68], [69]. Other methods [70]–[78] also take into consideration the relationships between tags and images simultaneously. These methods formulate tag refinement as problems of matrix completion [70]–[73], matrix factorization [74]–[76] and graph analysis [77], [78]. Based on these mathematical techniques, the reliability of each tag is evaluated to remove rare or irrelevant tags and complement missing relevant tags. However, these methods focus only on tags in one language, which is mostly English. This means that tags in other languages are treated as noisy tags to be removed, though the number of tags in non-English languages is not negligible [17]. Tags in these languages are discarded despite their potential for contribution to improvement of tag-based applications [18], [19]. Also, non-English speaking users, accounting for a relatively large proportion of users in Flickr folksonomy, cannot benefit from the previous tag refinement methods. Thus, multilingual frameworks for tag refinement are absolutely necessary.

We applied LION to tag refinement. The hierarchies revealed by LION provide hypernymous and hyponymous

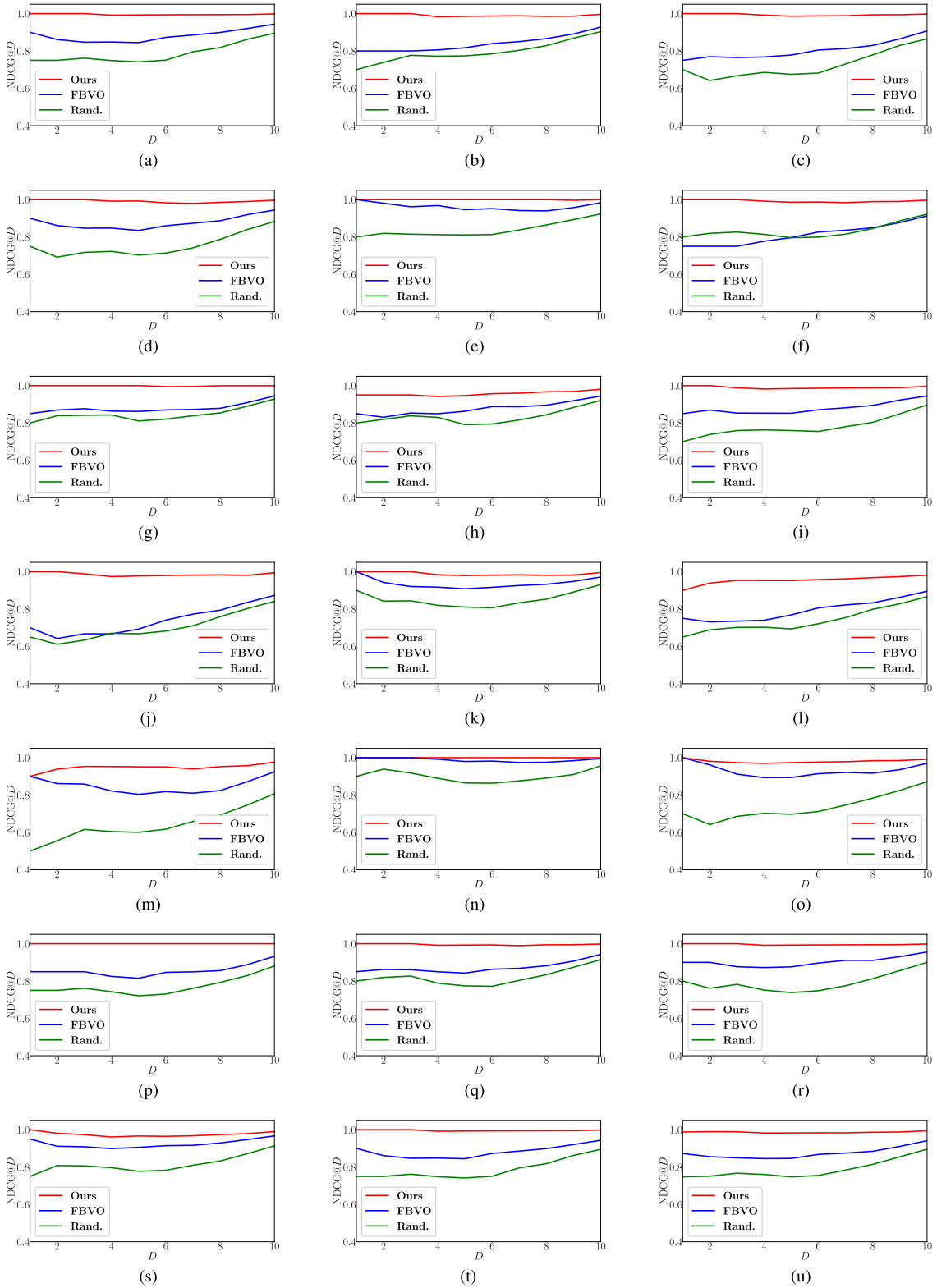


FIGURE 4. NDCG at different position *D* from 1 to 10. (a) Participant 1. (b) Participant 2. (c) Participant 3. (d) Participant 4. (e) Participant 5. (f) Participant 6. (g) Participant 7. (h) Participant 8. (i) Participant 9. (j) Participant 10. (k) Participant 11. (l) Participant 12. (m) Participant 13. (n) Participant 14. (o) Participant 15. (p) Participant 16. (q) Participant 17. (r) Participant 18. (s) Participant 19. (t) Participant 20. (u) Average over participants.

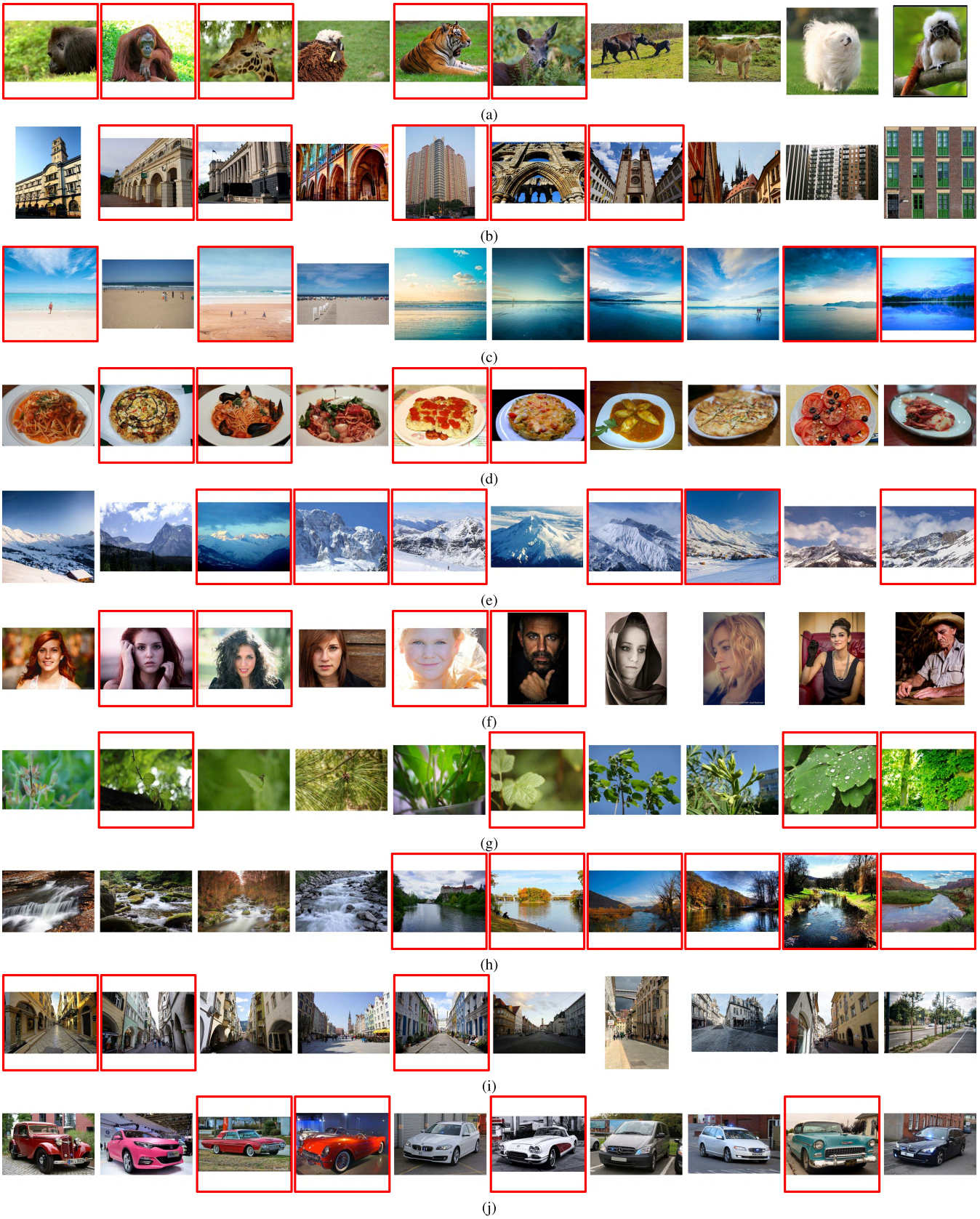


FIGURE 5. Image retrieval results by the proposed method when each keyword was input. Images highlighted with red frames are those with tags in languages different from LANG(1), which were not to be retrieved by the comparative methods that deal with keywords in only one language. (a) "animal". (b) "architecture". (c) "beach". (d) "food". (e) "mountain". (f) "people". (g) "plant". (h) "river". (i) "street". (j) "vehicle".

TABLE 5. Number of test images for each language pair.

Language	EN	DE	FR	RU	JA	ZH	KO
EN		1031	1172	987	862	913	611
DE			1349	788	727	671	445
FR				826	570	432	512
RU					489	538	398
JA						1288	1032
ZH							1201
KO							

tags of each tag. Thus, we consult the extracted tag hierarchies to remove irrelevant tags and complement hypernymous tags of the tags already assigned to the images. Given a tagged image, we calculate a confidence score of each tag defined as follows:

$$p(\boldsymbol{v}_{i,n}|c_i') = \frac{1}{N_i}\sum_{v=1}^{N_i} g_\gamma\left(\boldsymbol{v}_{i,n}, \boldsymbol{v}_{i,v}\right), \tag{19}$$

where $g_\gamma(\cdot,\cdot)$ is defined as

$$g_\gamma(\boldsymbol{v}_{i,n}, \boldsymbol{v}_{j,v}) = \exp\left(-\frac{\|\boldsymbol{v}_{i,n}^{(1)} - \boldsymbol{v}_{j,v}^{(2)}\|^2}{2\gamma^2}\right), \tag{20}$$

and γ is the median of $\|\boldsymbol{v}_{i,n}^{(1)} - \boldsymbol{v}_{j,v}^{(2)}\|$. To best leverage the hierarchical structure of the LION, we redefine c_i' by expanding each LIC c_i with visual features of images that represent the child node LICs of c_i. We removed already assigned tags with lower confidence scores than the average. Also, we complemented hypernymous tags of the already assigned tags by consulting the derived hierarchies revealed by LION.

To perform tag refinement, we randomly removed 40% of the assigned tags from all images in the dataset in such a way that each image has at least one tag removed and keeps at least one tag after tag removal. We split the dataset into a training set and a test set. As shown in Table 5, the number of test images is different according to the language pair. The test sets are relatively small compared to the total number of images since we selected the images with tags in multiple languages for proper evaluation. In this experiment, we took the originally assigned tags for the ground truth due to the high cost of manual judgments for tag refinement. We evaluated tag refinement accuracy in terms of Average Precision (AP), Average Recall (AR) and coverage (C) defined as follows:

$$\mathrm{AP}@K = \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\frac{R(n)}{K}, \tag{21}$$

$$\mathrm{AR}@K = \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\frac{R(n)}{R_{\mathrm{GT}}(n)}, \tag{22}$$

$$\mathrm{C}@K = \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\mathbb{I}[R(n) > 0], \tag{23}$$

where N_{test} is the number of test images, and $R_{\mathrm{GT}}(n)$ and $R(n)$ are the number of ground truth tags for the nth image and the number of correctly recovered tags for the nth image, respectively. We compared the proposed method and our previous method [50] (**Prev.**) in terms of tag refinement accuracy.

The tag refinement performance of each method is shown in Table 6. From the results, we can confirm that the proposed method outperforms the previous method. The difference between **Ours** and **Prev.** is whether CCA is applied to textual features or not. Therefore, the improvement is attributed to the canonical textual features, which consider correlations between textual features in different languages.

B. TAG-BASED IMAGE RETRIEVAL

With expansion of image folksonomies, the number of images on the Web has been dramatically increasing [79]. Image retrieval techniques play a key role in efficiently obtaining desired images. Typical retrieval systems return lists of images relevant to input queries. However, it is difficult for users to identify their desired images if they cannot input proper queries [80]. Thus, image retrieval on the Web has recently shifted from the conventional query-response systems to exploratory ones [81]. A typical exploratory retrieval starts when a user has interest in finding information on a topic with vague or broad queries [82].

We evaluated the LION from the viewpoint of applicability. We applied the constructed LION to a tag-based image retrieval system. Following the previous subsection, we expand each concept c with visual features of images that represent the child node LICs of c. Then we calculate the confidence score of c in the same manner as Eq. (19). Given a keyword, the system returns the top ten relevant images in descending order of confidence scores. In this experiment, we randomly selected two languages, $\mathtt{LANG}^{(1)}$ and $\mathtt{LANG}^{(2)}$, and used ten concepts corresponding to the keywords "animal", "architecture", "beach", "food", "mountain", "people", "plant", "river", "street" and "vehicle". We had 20 participants judge the relevance of each image on three scales: "0: Irrelevant", "1: Relevant" and "2: Very Relevant." Based on these values, we evaluated the applicability of the methods to image retrieval using the normalized discounted cumulative gain (NDCG) [83] defined as follows:

$$\mathrm{NDCG}@D = \frac{1}{Z_D}\sum_{d=1}^{D}\frac{2^{r_d}-1}{\log_2(d+1)}, \tag{24}$$

where Z_D is a normalization constant so that $\mathrm{NDCG}@D \in [0,1]$ and r_d is the relevance score at rank d. For calculation of Z_D, the ground truth of r_d was obtained through majority voting of the participants' labeling. We adopted the method [16] (**FBVO**) and random retrieval (**Rand.**) for comparative methods. Figure 4 shows NDCG@D for each participant and the average obtained by the proposed method and the comparative methods. This figure verifies that the proposed method outperforms the state-of-the-art **FBVO** for all participants. Retrieval results by the proposed method are shown in Fig. 5. Images highlighted with red frames are those with tags in languages different from $\mathtt{LANG}^{(1)}$. The proposed method successfully provided these images by aggregating synonymous concepts across languages on the basis of visual features. These figures indicate that synonymous concept

TABLE 6. Average values of AP@2, AR@2 and C@2 over languages for tag refinement. (a) AP@2. (b) AR@2. (c) C@2.

(a)

Language	Ours	Prev.
EN	**0.35**	0.28
DE	**0.32**	0.27
FR	**0.33**	0.29
RU	**0.32**	0.28
JA	**0.34**	0.27
ZH	**0.31**	0.29
KO	**0.33**	0.27

(b)

Language	Ours	Prev.
EN	**0.75**	0.71
DE	**0.77**	0.72
FR	**0.73**	0.69
RU	**0.70**	0.67
JA	**0.76**	0.70
ZH	**0.74**	0.71
KO	**0.72**	0.70

(c)

Language	Ours	Prev.
EN	**0.77**	0.64
DE	**0.69**	0.51
FR	**0.68**	0.56
RU	**0.74**	0.58
JA	**0.71**	0.53
ZH	**0.75**	0.62
KO	**0.72**	0.52

aggregation plays a crucial role in improving the image retrieval accuracy.

V. CONCLUSIONS

In this paper, we have proposed a LION construction method using tagged images in the Flickr folksonomy. The proposed method identifies LICs based on visual similarities between concepts. In this way, the proposed method enables extraction of relationships between LICs in LION construction. The constructed LION provides an easy way to access concepts in any language. Experimental results verified that the proposed method achieves high quality of the LION, and LION is in its element when applied to tag refinement and tag-based image retrieval.

REFERENCES

[1] J. Monti, M. Monteleone, M. P. Di Buono, and F. Marano, "Natural language processing and big data—An ontology-based approach for cross-lingual information retrieval," in *Proc. IEEE Int. Conf. Social Comput.*, Sep. 2013, pp. 725–731.

[2] A. Groza and P. O. Maria, "Mining arguments from cancer documents using natural language processing and ontologies," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process.*, Sep. 2016, pp. 77–84.

[3] W. Damak, I. Rebai, and I. K. Kallel, "Semantic object recognition by merging decision tree with object ontology," in *Proc. IEEE Int. Conf. Adv. Technol. Signal Image Process.*, Mar. 2014, pp. 65–70.

[4] H. Fukuda, S. Mori, Y. Kobayashi, Y. Kuno, and D. Kachi, "Object recognition based on human description ontology for service robots," in *Proc. Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2014, pp. 4051–4056.

[5] Y. I. A. Khalid and S. A. Noah, "A framework for integrating DBpedia in a multi-modality ontology news image retrieval system," in *Proc. IEEE Int. Conf. Semantic Technol. Inf. Retr.*, Jun. 2011, pp. 144–149.

[6] G. Besbes and H. Baazaoui-Zghal, "Fuzzy ontology-based medical information retrieval," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2016, pp. 178–185.

[7] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proc. ACM Int. Conf. Formal Ontol. Inf. Syst.*, 2001, pp. 2–9.

[9] M. Naphade *et al.*, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, Jul./Sep. 2006.

[10] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 206–213.

[11] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 167–187, Feb. 2008.

[12] J. Ha, K. Kim, and B. Zhang, "Automated construction of visual-linguistic knowledge via concept learning from cartoon videos," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 522–528.

[13] S.-F. Tsai, H. Tang, F. Tang, and T. S. Huang, "Ontological inference framework with joint ontology construction and learning for image understanding," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 426–431.

[14] A. Chianese, F. Marulli, and F. Piccialli, "Cultural heritage and social pulse: A semantic approach for CH sensitivity discovery in social media data," in *Proc. IEEE Int. Conf. Semantic Comput.*, Feb. 2016, pp. 459–464.

[15] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[16] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and A. Ghoneim, "Folksonomy-based visual ontology construction and its applications," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 702–713, Apr. 2016.

[17] A. Koochali, S. Kalkowski, A. Dengel, D. Borth, and C. Schulze, "Which languages do people speak on Flickr?: A language and geo-location study of the YFCC100M dataset," in *Proc. ACM Workshop Multimedia COMMONS*, 2016, pp. 35–42.

[18] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 976–983.

[19] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S. F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 159–168.

[20] G. Ngai, M. Carpuat, and P. Fung, "Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment," in *Proc. ACM Int. Conf. Comput. Linguistics*, 2002, pp. 1–7.

[21] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, Dec. 2012.

[22] N. Collier *et al.*, "A multilingual ontology for infectious disease surveillance: Rationale, design and challenges," *Language Resour. Eval.*, vol. 40, nos. 3–4, pp. 405–413, 2006.

[23] M. Simonet *et al.*, "Multilingual enrichment of an ontology of cardiovascular diseases," in *Proc. IEEE Conf. Comput. Cardiol.*, Sep. 2006, pp. 909–912.

[24] A. Annane, V. Emonet, F. Azouaou, and C. Jonquet, "Multilingual mapping reconciliation between English-French biomedical ontologies," in *Proc. ACM Int. Conf. Web Intell., Mining Semantics*, 2016, pp. 1–13.

[25] J. Salim, S. F. M. Hashim, and A. Aris, "A framework for building multilingual ontologies for Islamic portal," in *Proc. IEEE Int. Symp. Inf. Technol.*, Jun. 2010, pp. 1302–1307.

[26] S. Albukhitan and T. Helmy, "Multilingual food and health ontology learning using semi-structured and structured Web data sources," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Dec. 2012, pp. 231–235.

[27] R. Lourdusamy and M. F. Joseph, "A multilingual ontology for Tamil literary works," in *Proc. IEEE Int. Conf. Adv. Comput. Commun. Syst.*, Jan. 2016, pp. 1–4.

[28] N. Pappas *et al.*, "Multilingual visual sentiment concept matching," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 151–158.

[29] B. Jou, M. Y. Qian, and S.-F. Chang, "SentiCart: Cartography and geo-contextualization for multilingual visual sentiment," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 389–392.

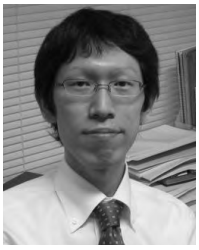
[30] H. Liu *et al.*, "Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 417–420.

[31] C. Krstev, D. Vitas, D. Maurel, and M. Tran, "Multilingual ontology of proper names," in *Proc. Language Technol. Conf.*, 2005, pp. 116–119.

[32] A. Savary, L. Manicki, and M. Baron, "Populating a multilingual ontology of proper names from open sources," *J. Language Model.*, vol. 1, no. 2, pp. 189–225, 2013.

[33] J. Guyot, S. Radhouani, and G. Falquet, "Ontology-based multilingual information retrieval," in *Proc. Workshop Cross-Langu. Edu. Function*, 2005, pp. 21–23.
[34] H. Aliane, "An ontology based approach to multilingual information retrieval," in *Proc. IEEE Int. Conf. Inf. Commun. Technol.*, Oct. 2006, pp. 1732–1737.
[35] P. Yu and H. Wang, "A multilingual ontology-based approach to attribute correspondence identification," in *Proc. IEEE Int. Conf. Elect. Control Eng.*, Sep. 2011, pp. 1896–1900.
[36] D. Embley, S. Liddle, D. Lonsdale, and Y. Tijerino, "Multilingual ontologies for cross-language information extraction and semantic search," in *Proc. ACM Int. Conf. Conceptual Model.*, 2011, pp. 147–160.
[37] C. Meilicke *et al.*, "MultiFarm: A benchmark for multilingual ontology matching," *Web Semantics, Sci., Services Agents World Wide Web*, vol. 15, pp. 62–68, Sep. 2012.
[38] N. Paulins, I. Arhipova, and S. Balina, "Multilingual information delivery based on a domain ontology," in *Proc. ACM Int. Conf. Comput. Syst. Technol.*, 2014, pp. 430–436.
[39] M. T. M. Ankon, S. N. Tumpa, and M. M. Ali, "A multilingual ontology based framework for Wikipedia entry augmentation," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.*, Dec. 2016, pp. 541–545.
[40] S. Gouws, Y. Bengio, and G. Corrado, "BilBOWA: Fast bilingual distributed representations without word alignments," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 748–756.
[41] A. Bérard, C. Servan, O. Pietquin, and L. Besacier, "MultiVec: A multilingual and multilevel representation learning toolkit for NLP," in *Proc. Langu. Resour. Eval. Conf.*, 2016, pp. 4188–4192.
[42] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 462–471.
[43] P. Koehn and K. Knight, "Learning a translation lexicon from monolingual corpora," in *Proc. ACL Workshop Unsupervised Lexical Acquisition*, 2002, pp. 9–16.
[44] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," in *Proc. ACL Workshop Human Langu. Technol.*, 2008, pp. 771–779.
[45] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.
[46] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, 2008.
[47] S. Tunga, D. Jayadevappa, and C. Gururaj, "A comparative study of content based image retrieval trends and approaches," *Int. J. Image Process.*, vol. 9, no. 3, pp. 127–155, 2015.
[48] R. Funaki and H. Nakayama, "Image-mediated learning for zero-shot cross-lingual document retrieval," in *Proc. Conf. Empirical Methods Natural Langu. Process.*, 2015, pp. 585–590.
[49] H. Nakayama and N. Nishida, "Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot," *Mach. Transl.*, vol. 31, no. 1, pp. 49–64, 2017.
[50] S. Hamano, T. Ogawa, and M. Haseyama, "Tag refinement based on multilingual tag hierarchies extracted from image folksonomy," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 1327–1331.
[51] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[53] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.
[54] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[55] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Langu. Process.*, 2014, pp. 1532–1543.
[56] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[57] R. Harakawa, T. Ogawa, and M. Haseyama, "Extraction of hierarchical structure of Web communities including salient keyword estimation for Web video retrieval," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1021–1025.
[58] G. W. Furnas and J. Zacks, "Multitrees: Enriching and reusing hierarchical structure," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst.*, 1994, pp. 330–336.
[59] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[60] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[61] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 1–22.
[62] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[63] D. Neal, "Folksonomies: Introduction: Folksonomies and image tagging: Seeing the future?" *Bull. Amer. Soc. Inf. Sci. Technol.*, vol. 34, no. 1, pp. 7–11, 2007.
[64] E. Spyrou and P. Mylonas, "A survey on Flickr multimedia research challenges," *Eng. Appl. Artif. Intell.*, vol. 51, pp. 71–91, May 2016.
[65] J. Dodge *et al.*, "Detecting visual text," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Langu. Technol.*, 2012, pp. 762–772.
[66] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & wordnet," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 706–715.
[67] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 491–500.
[68] D. Liu, X.-S. Hua, and H.-J. Zhang, "Content-based tag processing for Internet social images," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 723–738, 2011.
[69] M. Joseph and M. S. G. Premi, "Contextual feature discovery and image ranking for image object retrieval and tag refinement," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 190–194.
[70] X. Li, Y.-J. Zhang, B. Shen, and B. Di Liu, "Low-rank image tag completion with dual reconstruction structure preserved," *Neurocomputing*, vol. 173, no. 2, pp. 425–433, 2016.
[71] S. Lee, W. De Neve, and Y. M. Ro, "Tag refinement in an image folksonomy using visual similarity and tag co-occurrence statistics," *Signal Process., Image Commun.*, vol. 25, no. 10, pp. 761–773, 2010.
[72] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1618–1625.
[73] Z. Lin, G. Ding, M. Hu, Y. Lin, and S. S. Ge, "Image tag completion via dual-view linear sparse reconstructions," *Comput. Vis. Image Understand.*, vol. 124, pp. 42–60, Jul. 2014.
[74] Z. Li and J. Tang, "Deep matrix factorization for social image tag refinement and assignment," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2015, pp. 1–6.
[75] X. Li, B. Shen, B. Di Liu, and Y.-J. Zhang, "A locality sensitive low-rank model for image tag completion," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 474–483, Mar. 2016.
[76] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 276–288, Jan. 2017.
[77] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug. 2011.
[78] M. Liu, L. Chen, W. Ye, and M. Xu, "A sparsity constrained low-rank matrix completion approach for image tag refinement," in *Proc. IEEE Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery*, Aug. 2016, pp. 1288–1295.
[79] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Comput. Surv.*, vol. 49, no. 1, 2015, Art. no. 14.
[80] M. Haseyama, T. Ogawa, and N. Yagi, "A review of video retrieval based on image and video semantic understanding," *ITE Trans. Media Technol. Appl.*, vol. 1, no. 1, pp. 2–9, 2013.
[81] R. White and R. Roth, *Exploratory Search: Beyond the Query-Response Paradigm*. San Rafael, CA, USA: Morgan & Claypool, 2009.
[82] J. Y. Park, N. O'Hare, R. Schifanella, A. Jaimes, and C.-W. Chung, "A large-scale study of user image search behavior on the Web," in *Proc. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 985–994.
[83] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

SHOTA HAMANO (S'16) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2016, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interests include multimedia analysis and natural language processing. He is a Student Member of the IEICE.



TAKAHIRO OGAWA (S'03–M'08) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 2003, 2005, and 2007, respectively, all in electronics and information engineering. He is currently an Associate Professor with the Graduate School of Information Science and Technology, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the ITE TRANSACTIONS ON MEDIA TECHNOLOGY AND APPLICATIONS. He is a member of the EURASIP, IEICE, and Institute of Image Information and Television Engineers.

MIKI HASEYAMA (S'88–M'91–SM'06) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively, all in electronics. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Graduate School of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Institute of Image Information and Television Engineers (ITE), Japan, and the Information Processing Society of Japan. She has been a Vice-President of ITE, the Editor-in-Chief of the ITE TRANSACTIONS ON MEDIA TECHNOLOGY AND APPLICATIONS, the Director of the International Coordination and Publicity of IEICE.