TAG REFINEMENT BASED ON MULTILINGUAL TAG HIERARCHIES EXTRACTED FROM IMAGE FOLKSONOMY

Shota Hamano, Takahiro Ogawa and Miki Haseyama

Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: {hamano, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

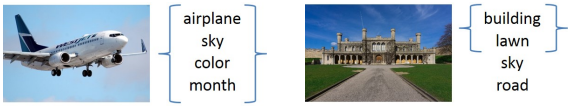
ABSTRACT

This paper presents a novel method for tag refinement using multilingual sources of tagged images in an image folksonomy. The proposed method enables accurate tag refinement by effectively leveraging multilingual sources of tags and considering the hierarchical structure of tags in the following way. First, synonymous tags across different languages are detected based on similarities between tagged images. In this stage, the proposed method utilizes visual similarities to effectively detect synonymous tags since the visual features extracted from images should be similar if they are assigned tags with the same meaning in different languages. Then hierarchical structure of the tags are extracted based on the similarity between the detected synonymous tags. The hierarchical structure provides hypernymous and hyponymous tags of the target tags, which are important for considering the relevance between tags and images. Consulting the hierarchical structure enables removal of irrelevant tags from the images and assignment of relevant tags to the images. The proposed method effectively utilizes tags in various languages in an image folksonomy. Experimental results show the effectiveness of introducing multilingual sources of tagged images for accuracy improvement in tag refinement.

Index Terms— Tagged images, multilingual applications, image folksonomies, hierachical structure extraction, tag refinement.

1. INTRODUCTION

Image sharing services have provided easier ways to upload and access images on the Web and led to a dramatic increase in the number of images [1]. For example, Flickr[1], one of the largest image folksonomies, stores ten billion images and attracts over 100 million users in 63 countries around the world as of September 2016[2]. Retrieval technology is essential for obtaining desired images from a huge number of images in an efficient manner. Typical image retrieval methods utilize tags provided by users for describing the semantic contents of images. These methods present lists of images with tags relevant to the input keywords. However, as shown in Fig. 1, user-provided tags in an image folksonomy are usually noisy and incomplete due to subjectivity of users [2]. They tend to assign tags to images based on their knowledge and experience, which often fall short of high consistency for describing the semantic contents of the images. These tags cause degradation in the tag-based retrieval accuracy.



(a) "color" and "month" are assigned to the image while they are irrelevant tags.

(b) "sky" and "road" are not assigned to the image while they are relevant tags.

Fig. 1: Examples of Flickr images.

Solutions to this problem include tag refinement techniques [3–15]. Tag refinement is technology that removes noisy tags and complements 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 [3, 4] or between images [5, 6]. Other methods [7–15] also take into consideration the relationships between tags and images simultaneously. These methods formulate tag refinement as problems of matrix completion [7–10], matrix factorization [11–13] and graph analysis [14, 15]. 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 while the number of tags in non-English languages is not significantly small [16]. Tags in these languages are discarded despite their potential for contribution to improvement of tag-based applications [17, 18]. Also, non-English speaking users, accounting for certain propotions in the Flickr folksonomy, are unable to be benefitted from the previous tag refinement methods. From this background, multilingual frameworks for tag refinement are absolutely necessary.

In this paper, we propose a novel method that leverages tags in different languages for improvement in tag refinement. We consider the language as a factor representing users' information and utilize multilingual sources of tagged images. The proposed method first calculates pairwise similarities between tags in one language and tags in another language. We detect synonymous tags in different languages based on the similarities to utilize otherwise discarded tags for performance improvement in tag refinement. Since user-provided tags are imperfect, we utilize visual similarities for accurate synonymous tag detection based on observation that visual features extracted from images with synonymous tags are similar [19, 20]. This is the biggest contribution of this paper. In this paper, we con-

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP15K12023.

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

[2] http://expandedramblings.com/index.php/flickr-stats/

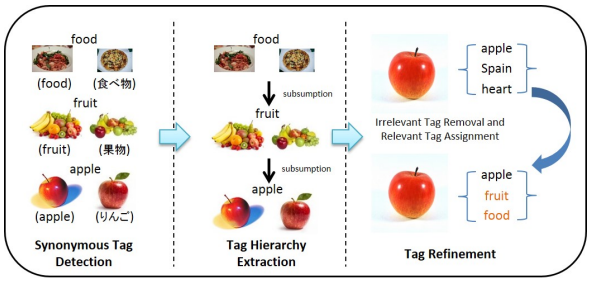


Fig. 2: Overview of the proposed method. In this figure, the proposed method is applied to English and Japanese tags.

sider the detected synonymous tag pairs as new tags. Then we collaboratively use visual and textual features to extract the hierarchical structure of tags to consider hypernyms and hyponyms of each tag. Consulting the extracted tag hierarchies, the proposed method removes irrelevant tags and complements hypernymous tags of the tags already assigned to the images. Collaborative use of tags in different languages leads to high accuracy in tag refinement. The proposed method works effectively in the current situation, where more and more languages are used for tag assignment.

2. TAG REFINEMENT BASED ON MULTILINGUAL TAG HIERARCHIES

In this section, we propose a method for tag refinement that effectively utilizes multilingual sources of tagged images in an image folksonomy. The proposed method consists of the following three stages as shown in Fig. 2.

Synonymous Tag Detection
: We detect synonymous tags across different languages based on similarities between tagged images. We utilize visual similarities for accurate synonymous tag detection based on observation that images with synonymous tags are represented by similar visual features. This stage is effective in that we can avoid discarding the relevant tags in different languages and leverage the potential of these tags in an image folksonomy.

Tag Relationship Extraction
: We extract the relationships between tags using visual and textual similarities between tagged images. Visual and textual features complementarily contribute to accurate extraction of tag relationships.

Tag Removal and Completion
: Based on the tag relationships extracted in the previous stage, we remove irrelevant tags from the tagged images and complement hypernymous tags to the images. Consulting the tag hierarchies improves performance in tag refinement.

The main contribution of the proposed method is that we effectively leverage otherwise discarded tags in various languages for improvement in tag refinement performance.

2.1. Synonymous Tag Detection

To leverage multilingual sources of tagged images, we first detect synonymous tags across different languages based on similarities between tagged images. We use visual similarities for accurate synonymous tag detection since images with synonymous tags are considered to be represented by similar visual features. We extract visual features $\boldsymbol{V}^{(i)} = [\boldsymbol{v}_1^{(i)}\boldsymbol{v}_2^{(i)}\cdots\boldsymbol{v}_{N_i}^{(i)}]$ from N_i images with tag $w_i \in W$ and $\boldsymbol{V}^{(i')} = [\boldsymbol{v}_1^{(i')}\boldsymbol{v}_2^{(i')}\cdots\boldsymbol{v}_{N_{i'}}^{(i')}]$ for tag $w_{i'} \in W'$ in the same way, where W and W' denote a set of tags in one language and a set of tags in another language, respectively. In this paper, we used 4,096-dimensional features extracted from the fc6 layer in the AlexNet [21].

Next, we represent tags w_i and $w_{i'}$ by feature vectors $\boldsymbol{v}^{(i)} = (\sum_{n=1}^{N_i}\boldsymbol{v}_n^{(i)})/N_i$ and $\boldsymbol{v}^{(i')} = (\sum_{n=1}^{N_{i'}}\boldsymbol{v}_n^{(i')})/N_{i'}$, respectively. Using these vectors, we calculate the similarity $S(w_i, w_{i'})$ as follows:

$$S(w_i, w_{i'}) = \frac{\boldsymbol{v}^{(i)\top}\boldsymbol{v}^{(i')}}{\|\boldsymbol{v}^{(i)}\|\|\boldsymbol{v}^{(i')}\|}, \tag{1}$$

where $\|\cdot\|$ and $^\top$ denote the Euclidean norm and the transposition of the vector, respectively.

Finally, we match tag w_i with $\hat{w}_i$ that satisfies

$$\hat{w}_i = \arg\max_{w_{i'}} S(w_i, w_{i'}) \tag{2}$$

to detect semantically identical tag pairs. In the following subsections, we regard a tag $\hat{w}_i$ as a tag with the same meaning as the tag w_i. We denote tag $c_i \in C$ to represent both of the tags w_i and $\hat{w}_i$, where C is the set of the tags.

2.2. Tag Hierarchy Extraction

Synonymous tag detection is followed by extraction of tag hierarchies to consider hypernymous and hyponymous tags. For each tag c_i, we extract visual features $\boldsymbol{f}_v^{(i)} = [\boldsymbol{v}^{(i)\top}\boldsymbol{v}^{(i')\top}]^\top$ and textual features $\boldsymbol{f}_t^{(i)} = [\boldsymbol{t}^{(i)\top}\boldsymbol{t}^{(i')\top}]^\top$ from the corresponding tagged images, where $\boldsymbol{t}^{(i)}$ and $\boldsymbol{t}^{(i')}$ are textual features for w_i and $\hat{w}_i$, respectively. In this paper, we adopt the word2vec [22] model in extraction of textual features. We then calculate similarities between tags c_i and c_j based on each modality $m \in \{v, t\}$ as follows:

$$s_m(c_i, c_j) = \frac{\boldsymbol{f}_m^{(i)\top}\boldsymbol{f}_m^{(j)}}{\|\boldsymbol{f}_m^{(i)}\|\|\boldsymbol{f}_m^{(j)}\|}. \tag{3}$$

We leverage the more effective modality to determine similarities between tags c_i and c_j. Here we sort $s_m(c_i, c_j)$ in ascending order and denote them as $s_m(l)\,(l = 1, 2, \ldots, L)$. L is equal to $|C|^2$, where $|\cdot|$ denotes the cardinality of the set. Next, we build the empirical distribution function (EDF) [23] $F_m(x)$ defined as

$$F_m(x) = \frac{1}{L}\sum_{l=1}^{L}\mathbb{I}[s_m(l) \le x], \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function, which returns 1 if the condition is satisfied, and 0 otherwise. We adaptively use the similarity on the basis of the more effective modality for each tag as follows:

$$s(c_i, c_j) = \max_{m\in\{v,t\}} F_m\left(s_m(c_i, c_j)\right). \tag{5}$$

In this way, we can select an optimal modality for tag relationship extraction since EDF enables comparison of the effectiveness of the modalities by considering the statistical distribution of similarities [24].

Table 1: The number of images collected from Flickr and training corpora for word2vec.

Language	Images	Vocabulary	Word tokens
EN	30,143	1,778,575	1,777,497,491
DE	29,684	1,505,149	618,569,591
FR	29,694	754,100	468,980,630
JA	28,854	367,607	256,158,273
ZH	25,874	137,567	229,237,678

Then we quantify the extent that a tag subsumes another tag to extract hierarchical structure of tags. We define a subsumption score utilizing the similarities obtained by Eq. (5) as follows:

$$p(c_i|c_j) = \frac{s(c_i, c_j)}{\sum_{c' \in C} s(c_i, c')}. \tag{6}$$

This score represents the extent that tag c_i subsumes c_j. When $p(c_i|c_j)$ is larger than $p(c_j|c_i)$, we regard c_i as a hypernymous tag of tag c_j.

2.3. Tag Removal and Completion

Finally, we remove irrelevant tags and complement missing relevant tags based on the tag hierarchies extracted in the previous stages. Given a tagged image, we calculate a confidence score of each tag defined as follows:

$$p(\boldsymbol{v}_n^{(i)}|c_i) = \frac{1}{N_i} \sum_{\hat{n}=1}^{N_i} K_\sigma \left(\boldsymbol{v}_n^{(i)}, \boldsymbol{v}_{\hat{n}}^{(i)} \right), \tag{7}$$

where

$$K_\sigma \left(\boldsymbol{v}_n^{(i)}, \boldsymbol{v}_{\hat{n}}^{(i)} \right) = \exp \left(-\frac{\|\boldsymbol{v}_n^{(i)} - \boldsymbol{v}_{\hat{n}}^{(i)}\|^2}{2\sigma^2} \right), \tag{8}$$

and σ is the median of $\|\boldsymbol{v}_n^{(i)} - \boldsymbol{v}_{\hat{n}}^{(i)}\|$.

We remove already assigned tags with lower confidence scores than the average. Also, we complement hypernymous tags of the already assigned tags by consulting the derived hierarchies extracted in Eq. (6). Utilizing the derived tag hierarchies effectively captures the relevant hypernymous tags to achieve high accuracy in tag refinement.

3. EXPERIMENTAL RESULTS

This section reports the experimental results to verify the effectiveness of the proposed method (**Ours**). For the dataset, we collected images with tags in English (EN), German (DE), French (FR), Japanese (JA) and Chinese (ZH) languages by Flickr API[3]. In this experiment, visual features were extracted from the fc6 layer in the AlexNet by a deep learning framework Caffe [25]. We trained word2vec using Wikipedia corpora provided by Polyglot[4] to extract 200-dimensional textual features. The detail of the dataset is shown in Table 1. We used 150 non-abstract tags such as "airplane", "food" and "road" for evaluations. We separately evaluated the quality of synonymous tag detection, tag relationship extraction and tag refinement in the following subsections to properly assess the effectiveness of each procedure in the proposed method.

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

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

Table 2: Accuracy of synonymous tag detection. Due to space limitations, average values over languages are presented.

Language	**Ours**	**KDE**	**BOF**
EN	**0.582**	0.012	0.493
DE	**0.579**	0.008	0.482
FR	**0.558**	0.012	0.474
JA	**0.562**	0.010	0.490
ZH	**0.547**	0.015	0.488

3.1. Synonymous Tag Detection

We first evaluated the performance of **Ours** in synonymous tag detection. For comparative methods, we adopted the following similarity metrics.

Comparative method 1 (KDE):
This method uses the similarity based on kernel density estimation (KDE) [26] as follows:

$$S(w_i, w_{i'}) = \frac{1}{N_i N_{i'}} \sum_{n=1}^{N_i} \sum_{n'=1}^{N_{i'}} K_\sigma \left(\boldsymbol{v}_n^{(i)}, \boldsymbol{v}_{n'}^{(i')} \right), \tag{9}$$

where $K_\sigma(\cdot, \cdot)$ is defined in Eq. (8). Here σ is the median of $\|\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(i')}\|$.

Comparative method 2 (BOF):
This method uses the similarity defined as follows:

$$S(w_i, w_{i'}) = \frac{\boldsymbol{v}^{(i)\top} \boldsymbol{v}^{(i')}}{\|\boldsymbol{v}^{(i)}\| \|\boldsymbol{v}^{(i')}\|}, \tag{10}$$

where $\boldsymbol{v}^{(i)}$ and $\boldsymbol{v}^{(i')}$ are 1,000-dimensional vectors represented by bag-of-features (BOF) [27] obtained by applying the clustering algorithm in [28] to the 4,096-dimensional AlexNet fc6 features.

For the evaluation criterion, we used accuracy defined as

$$\text{Accuracy} = \frac{1}{|C|} \sum_{i=1}^{|C|} \mathbb{I}[\hat{w}_i = w_{\text{GT},i}], \tag{11}$$

where $w_{\text{GT},i}$ is the ground truth tag for w_i obtained via Google Translate[5]. $\mathbb{I}[\cdot]$ is the indicator function already defined in Eq. (4).

The results are shown in Table 2. For all the language pairs, our method achieved higher accuracy than those of the comparative methods. This demonstrates that our method is not only the simplest but also more effective than KDE and BOF representations.

3.2. Tag Hierarchy Extraction

We also evaluated the performance of tag hierarchy extraction. For comparison, we adopted the state-of-the-art relationship extraction method [29] (**FBVO**). We evaluated these methods by average precision (AP) and average recall (AR) defined as follows:

$$\text{AP@}k = \frac{1}{|C|} \sum_{c \in C} \frac{|U_k(c) \cap U_{\text{GT}}(c)|}{k}, \tag{12}$$

$$\text{AR@}k = \frac{1}{|C|} \sum_{c \in C} \frac{|U_k(c) \cap U_{\text{GT}}(c)|}{|U_{\text{GT}}(c)|}, \tag{13}$$

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

Table 3: Average precision and average recall of tag hierarchy extraction at different position k. Due to space limitations, average values over languages are presented.

(a) Average Precision

	Ours		**FBVO**	
Language	AP@10	AP@20	AP@10	AP@20
EN	**0.208**	**0.249**	0.188	0.236
DE	**0.289**	**0.284**	0.262	0.238
FR	**0.268**	**0.255**	0.175	0.200
JA	**0.235**	**0.223**	0.201	0.213
ZH	**0.269**	**0.267**	0.261	0.166

(b) Average Recall

	Ours		**FBVO**	
Language	AR@10	AR@20	AR@10	AR@20
EN	**0.459**	**0.511**	0.448	0.506
DE	**0.463**	**0.516**	0.459	0.491
FR	**0.488**	**0.508**	0.477	0.488
JA	**0.493**	**0.502**	0.482	0.500
ZH	0.488	**0.502**	**0.492**	0.499

Table 4: The number of test images for each language pair.

Language pair	Images	Language pair	Images
EN & DE	1,031	DE & FR	1,349
EN & FR	1,172	DE & JA	727
EN & JA	862	DE & ZH	671
EN & ZH	913	JA & ZH	1,288

where $U_k(c)$ and $U_{\text{GT}}(c)$ denote the set of top k extracted subsumption relationships of tag c and the ground truth relationships generated by WordNet [30], respectively. In this experiment, we regarded synsets in WordNet as the corresponding tags in different languages.

The results are shown in Table 3. This table shows that our method is more effective for tag relationship extraction than the state-of-the-art method in [29]. The results also indicate that the performance was improved by integrating synonymous tags in advance. This verifies that the proposed method effectively leverages multilingual sources of tagged images for tag relationship extraction.

3.3. Tag Refinement

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 4, the number of test images is different according to the language pair. The test sets are relatively small to the total number of images since we selected the images with tags in multiple languages for proper evaluations. In this experiment, we take 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 AP@k, AR@k and coverage (C) defined as follows:

$$\text{C@}k = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \mathbb{I}[R(n) > 0], \tag{14}$$

where N_{test} and $R(n)$ denote the number of test images and correctly recovered tags for nth image, respectively. In this subsection, AP@k and AR@k in Eqs. (12) and (13) are redefined as follows:

$$\text{AP@}k = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \frac{R(n)}{k}, \tag{15}$$

$$\text{AR@}k = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \frac{R(n)}{R_{\text{GT}}(n)}, \tag{16}$$

where $R_{\text{GT}}(n)$ is the number of ground truth tags for nth image. We adopted the method in [9] (**LSR**) for a comparative method. Besides, we applied the proposed framework to **LSR**, *i.e.*, tag refinement was conducted by the method [9] after synonymous tag detection in Sec. 2.1.

Table 5: AP@2, AR@2 and C@2 of tag refinement. The postfix **S** indicates that the method does not use multilingual sources of tagged images. Due to space limitations, average values over languages are presented.

(a) AP@2

Language	**Ours**	**Ours-S**	**LSR**	**LSR-S**
EN	**0.28**	0.27	0.27	0.25
DE	**0.27**	**0.27**	0.26	0.25
FR	**0.29**	0.28	0.27	0.26
JA	0.27	**0.28**	0.27	0.26
ZH	**0.29**	0.26	0.28	0.27

(b) AR@2

Language	**Ours**	**Ours-S**	**LSR**	**LSR-S**
EN	**0.71**	0.69	0.70	**0.71**
DE	**0.72**	0.71	0.70	0.69
FR	0.69	**0.70**	0.67	0.67
JA	**0.70**	0.69	0.67	0.67
ZH	**0.71**	0.70	0.69	0.70

(c) C@2

Language	**Ours**	**Ours-S**	**LSR**	**LSR-S**
EN	**0.64**	0.52	0.60	0.55
DE	**0.51**	0.48	**0.51**	0.46
FR	**0.56**	0.52	0.51	0.48
JA	**0.53**	0.50	0.50	0.42
ZH	**0.62**	0.58	0.58	0.45

We show the tag refinement performance of each method in Table 5. From the results, we can confirm that the proposed method outperforms the comparative methods. We can also confirm that the improvement is attributed to the effective use of multilingual sources of tagged images. It is indicated that synonymous tag detection is effective not only in the proposed method but also in other tag refinement methods.

4. CONCLUSIONS

In this paper, we have proposed a novel tag refinement method that effectively leverages multilingual sources of tagged images for improving accuracy. In the proposed method, we detect synonymous tags across different languages using visual similarities to make use of otherwise discarded tags in tag refinement framework. We also extract hierarchical structure of tags to consider hypernyms and hyponyms of each tag. Based on the hierarchical structure, we can improve the accuracy in tag refinement. The experimental results have shown the effectiveness of the proposed method that collaboratively uses multilingual sources of tagged images.

5. REFERENCES

[1] 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 Computing Surveys*, vol. 49, 2015.

[2] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daumé, III, A. C. Berg, and T. L. Berg, "Detecting visual text," in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 762–772.

[3] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & WordNet," in *Proceedings of ACM International Conference on Multimedia*, 2005, pp. 706–715.

[4] D. Liu, X. Hua, M. Wang, and H. Zhang, "Image retagging," in *Proceedings of ACM International Conference on Multimedia*, 2010, pp. 491–500.

[5] D. Liu, X. Hua, and H. Zhang, "Content-based tag processing for Internet social images," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 723–738, 2011.

[6] M. Joseph and M. S. G. Premi, "Contextual feature discovery and image ranking for image object retrieval and tag refinement," in *International Conference on Communication and Signal Processing*, 2014, pp. 190–194.

[7] X. Li, Y. Zhang, B. Shen, and B. Liu, "Low-rank image tag completion with dual reconstruction structure preserved," *Neurocomputing*, vol. 173, no. P2, pp. 425–433, 2016.

[8] S. Lee, W. D. Neve, and Y. M. Ro, "Tag refinement in an image folksonomy using visual similarity and tag co-occurrence statistics," *Signal Processing: Image Communication*, vol. 25, no. 10, pp. 761–773, 2010.

[9] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1618–1625.

[10] Z. Lin, G. Ding, M. Hu, Y. Lin, and S. S. Ge, "Image tag completion via dual-view linear sparse reconstructions," *Computer Vision and Image Understanding*, vol. 124, pp. 42–60, 2014.

[11] Z. Li and J. Tang, "Deep matrix factorization for social image tag refinement and assignment," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, 2015, pp. 1–6.

[12] X. Li, B. Shen, B. Liu, and Y. Zhang, "A locality sensitive low-rank model for image tag completion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 474–483, 2016.

[13] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 276–288, 2017.

[14] D. Liu, S. Yan, X. Hua, and H. Zhang, "Image retagging using collaborative tag propagation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 702–712, 2011.

[15] M. Liu, L. Chen, W. Ye, and M. Xu, "A sparsity constrained low-rank matrix completion approach for image tag refinement," in *IEEE International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2016, pp. 1288–1295.

[16] 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 *Proceedings of ACM Workshop on Multimedia COMMONS*, 2016, pp. 35–42.

[17] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 976–983.

[18] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *Proceedings of ACM International Conference on Multimedia*, 2015, pp. 159–168.

[19] R. Li, Y. Zhang, Z. Lu, J. Lu, and Y. Tian, "Technique of image retrieval based on multi-label image annotation," in *Proceedings of IEEE International Conference on Multimedia and Information Technology*, 2010, pp. 10–13.

[20] H. Ha, S. Chen, and M. Shyu, "Utilizing indirect associations in multimedia semantic retrieval," in *Proceedings of IEEE International Conference on Multimedia Big Data*, 2015, pp. 72–79.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[23] A. W. van der Vaart, *Asymptotic statistics*. Cambridge University Press, 1998.

[24] R. Harakawa, T. Ogawa, and M. Haseyama, "Extraction of hierarchical structure of Web communities including salient keyword estimation for Web video retrieval," in *Proceedings of IEEE International Conference on Image Processing*, 2015, pp. 1021–1025.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of ACM International Conference on Multimedia*, 2014, pp. 675–678.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification and scene analysis*. Wiley, 1995.

[27] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of European Conference on Computer Vision*, 2004, pp. 1–22.

[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, 1967, pp. 281–297.

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

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