Multimed Tools Appl (2011) 51:723–738
DOI 10.1007/s11042-010-0647-3

Content-based tag processing for Internet social images

Dong Liu · Xian-Sheng Hua · Hong-Jiang Zhang

Published online: 19 November 2010

Abstract Online social media services such as Flickr and Zooomr allow users to share their images with the others for social interaction. An important feature of these services is that the users manually annotate their images with the freely-chosen tags, which can be used as indexing keywords for image search and other applications. However, since the tags are generally provided by grassroots Internet users, there is still a gap between these tags and the actual content of the images. This deficiency has significantly limited tag-based applications while, on the other hand, poses a new challenge to the multimedia research community. It calls for a series of research efforts for processing these unqualified tags, especially in making use of content analysis techniques to improve the descriptive power of the tags with respect to the image contents. This paper provides a comprehensive survey of the technical achievements in the research area of content-based tag processing for social images, covering the research aspects on tag ranking, tag refinement and tag-to-region assignment. We review the research advances for each topic and present a brief suggestion for future promising directions.

Keywords Social images · Tag processing · Tag ranking · Tag refinement · Tag-to-region assignment

D. Liu (✉)
Harbin Institute of Technology, Harbin, China
e-mail: dongliu.hit@gmail.com

X.-S. Hua
Microsoft Research Asia, Beijing, China
e-mail: xshua@microsoft.com

H.-J. Zhang
Microsoft Advanced Technology Center, Beijing, China
e-mail: hjzhang@microsoft.com

1 Introduction

With the advent of Web 2.0 technology, there is an explosion of social media sharing system available online such as Flickr,[1] Youtube[2] and Zooomr.[3] Rather than simply searching for and passively consuming media content, these media repositories allow users to create and exchange their own media data for social interaction, which brings in a new revolution to our social lives and underscores a transformation of the Web as fundamental as its birth [1]. As one of the emerging Web 2.0 activities, tagging, the action of manually annotating the content with a set of freely-chosen tags, has become a more and more frequently-applied means to organize, index and search media content for general users, and it provides a potential way to realize real large-scale content-based multimedia retrieval [6, 13].

Despite the high popularity of tagging social images manually, the tags provided by the grassroot Internet users are actually far from satisfactory as qualified descriptive indexing keywords of the image contents. Specifically, the main issues associated with the social image tags lie in the following aspects:

- The relevance levels of the tags associate with a social image cannot be distinguished from the tag list, where the orders of different tags in the tag list are just based on the manual input and carry little information about the importance or relevance information, and this further limits the effectiveness of tags in search and other applications.
- The user-provided tags are often biased towards personal perspectives and context cues, and thus there is a gap between these tags and the content of the images that common users are interested in. Moreover, as it is impractical for the general users to annotate the images comprehensively, many potentially useful tags may be missed. Therefore, the user-provided tags are imprecise, biased, and incomplete for describing the content of the images.
- The current tags are typically annotated at the image level, while the correspondence between each semantic region within an image and its descriptive tag remains ambiguous, which hampers the development of reliable and visible content-based image retrieval systems.

In consideration of the facts stated above, we argue that the raw tags associated with the Internet social images need to be pre-processed before they can be applied as reliable content descriptors of the images. This opens a promising research direction which attracts a variety of research efforts from multimedia research community in recent years. Starting from the problems on social image tags, the main focus of the existing tag processing works has been put on the following three aspects:

- Tag ranking. This research dimension aims to differentiate the tags associated with the images with various degrees of relevance level. The tags with different relevance levels will benefit the visual search performance and in turn improve the relevance of tag-based applications.

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

[2] http://www.youtube.com/.

[3] http://www.zooomr.com/.

– Tag refinement. The purpose of tag refinement is to refine the unreliable user-provided tags associated with those social images. The refined tags better describe the contents of the social images and bring remarkable performance improvements for tag-based applications.
– Tag-to-region assignment. This research topic attempts to develop an effective mechanism to automatically assign tags annotated at the image level to the individual regions within an image, which is an interesting and practical valuable direction worth investigating.

In this survey paper, we present a comprehensive detailed study of the research topics above and review the recent advances on social image tag processing. Different from those existing works that solely rely on the tag statistic information for social image tag processing [22, 24, 26, 28], all the works involved in this paper adopt visual content analysis techniques to discover the relationship between social images and their associated tags. Therefore, we term these works as *content-based tag processing*. The remainder of the paper is organized as follows. Section 2 introduces various representative works on tag ranking. Section 3 discusses the state-of-the-art works on image tagging refinement with different statistical modeling algorithms. Section 4 presents the research topic of tag-to-region assignment along with its recent representative works. Section 5 presents a list of open challenges we are still facing and discuss the possible way-outs. Finally, Section 6 gives the conclusive remarks.

2 Tag ranking

As discussed in Section 1, the importance or relevance levels of the tags cannot be distinguished from the tag list associated with a social image. As an example, Fig. 1 is an image from Flickr, from which we can see that the most relevant (or descriptive) tag is actually "dog", but this cannot be discovered from the tag list directly.

This phenomenon is frequently observed in the social image sharing webistes. To justify this statement, we randomly select 1,200 Flickr images with at least ten tags. For each image, its most relevant tag from the list is labeled based on the majority voting of five volunteers. Figure 2 shows the position (in terms of the tag list) distribution of the most important tags. As can be seen, only less than 10% of the images have their most relevant tag at the top position in their attached tag list.

Fig. 1 An exemplary image from Flickr and its associated tag list. There are many imprecise and meaningless tags in the list and the most relevant tag "dog" is not at the top position

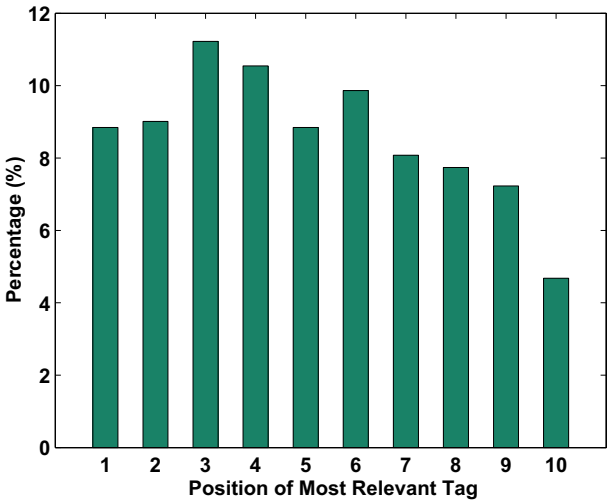

Fig. 2 Percentage of images that have their most relevant tag at the n-th position in the associated tag list, where $n = 1, 2, 3 \ldots, 10$

The lack of relevance information in the tag list has significantly limited the application of tags. For example, in Flickr's tag-based image search service,[4] currently it cannot provide the option of ranking the tagged images according to relevance level to the query.[5] However, relevance ranking is important for image search [8, 9], and all of the popular image search engines, like Google and Bing, rank the search results by relevance.

Clearly, research is required to improve this situation, and some recent efforts present a number of promising solutions towards this difficulty. Currently, the existing methods on tag ranking can be divided into two categories. The first category is based on the statistical modeling techniques [7, 17, 25]. As a pioneering work, Liu et al. [17] propose to estimate tag relevance scores using kernel density estimation, and then employ random walk to boost this primary estimation. Starting from this initial effort, Wang et al. [25] further propose a semi-supervised learning model to rank image tags, which learns a ranking projection from visual words distribution to the relevant tags distribution, and then uses it for ranking new image tags. In [7], Feng et al. investigate the tag ranking problem by combining both visual attention model and multi-instance learning model, and obtain encouraging results on some benchmark datasets. The second category is generally based on the data-driven techniques [11, 14–16]. The initial effort is the work performed by Li et al. [14], which scalably and reliably learns tag relevance by accumulating votes from visually similar neighbors. Recently, Kennedy et al. [11] find that the tags entered by separate people on visually similar images are likely to be highly related to the image content, and therefore are reliable and useful for visual applications. In the following, we

[4] http://www.flickr.com/search/?q=cat&m=tags.

[5] Currently Flickr offers two options in the ranking for tag-based image search. One is "most recent", which ranks the most recently uploaded images on the top and the other is "most interesting", which ranks the images by "interestingness", a measure that takes click-through, comments, etc, into account, as stated in http://www.flickr.com/explore/interesting.

will discuss the works in [17] and [14] in details, each of which is essentially the representative work of one tag ranking strategy.

2.1 Tag ranking by kernel density estimation initialization and random walk refinement

The tag ranking method proposed by Liu et al. [17] attempts to assign the relevance scores to the individual tags associated with the social images based on Kernel Density Estimation (KDE) initialization and random walk refinement. Given an image and its associated tags, the initial tag relevance estimation step aims to estimate the relevance score of each tag through a probabilistic approach, where it simultaneously considers the probability of the tag given the image and the descriptive ability of the tag in the entire image collection through the statistical model of KDE. Although the probabilistic scores obtained in this way reflect the tag relevance, the relationships among tags have not been taken into account. Therefore, a random walk-based refinement is further performed to boost tag ranking performance by exploring the relationship of tags. Finally, the tags of the image can be ranked according to their refined relevance scores. The overall framework of the proposed tag ranking method can be illustrated as in Fig. 3.

After the tag ranking process, the tags associated with the social images are differentiated with various degrees of relevance, which will further benefit the performance of tag-based applications. For example, the authors in [17] develop specific methodologies to apply the ranked tag list into the tasks of tag-based image search, neighbor voting based automatic tagging and tag-based group recommendation, and achieve remarkable performance improvement on each task.

Pros and cons Obviously, the tag ranking method based on KDE initialization and random walk refinement has a number of favorable characteristics. First, the KDE initialization procedure does not impose structure on the data in the way that some other statistical models often do, and thus is quite suitable for dealing with the Internet social images with significantly diversified visual contents. Second, the random walk procedure exploits the tag-to-tag correlations to reinforce relevant tags of an image, which has proved to be critical in the multi-label setting. On the other hand, the key issue in the above tag ranking method is the measurement of the visual

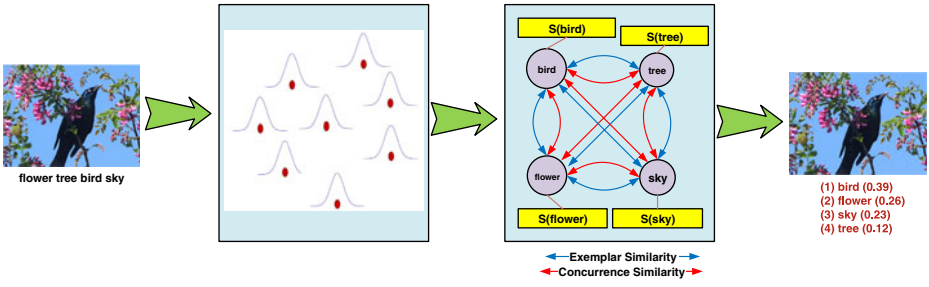

Fig. 3 The illustrative scheme of the tag ranking approach. A probabilistic method is first adopted to estimate tag relevance score. Then a random walk-based refinement is performed along the tag graph to further boost tag ranking performance

similarities, where the proposed method simply infers the pairwise image similarity based on the low level features such as color, texture and shape. However, whether two images are similar actually depends on what the semantic tags we are caring about. Using a holistic image feature representation to measure the image similarity is unable to take the tags into account, thus may not be able to well capture the desired semantic relationship among the images.

2.2 Tag ranking by neighbor voting

A representative work on the data-driven tag ranking methods is performed by Li et al. [14], which can be illustrated in Fig. 4. The basic assumption is intuitive: if different users label visually similar images using the same tags, these tags are likely to truly reflect the actual visual content. Starting from this assumption, the authors propose an algorithm which accurately and efficiently learns tag relevance by accumulating votes from visual neighbors of the target seed image. Different from the KDE process in [17], which uses both the tags and the visual features for the neighbor images seeking, the method in [14] determines the neighbor images based on visual features solely. To speed up this process, *K*-means based feature indexing strategy is employed to reduce the search space. Extensive experiments over real world Flickr image collection demonstrate the general effectiveness of the proposed method in both social image retrieval and image tag suggestion [15].

The deficiency in the above tag relevance learning method is that it only uses a single feature to estimate visual similarity between images. Unfortunately, as

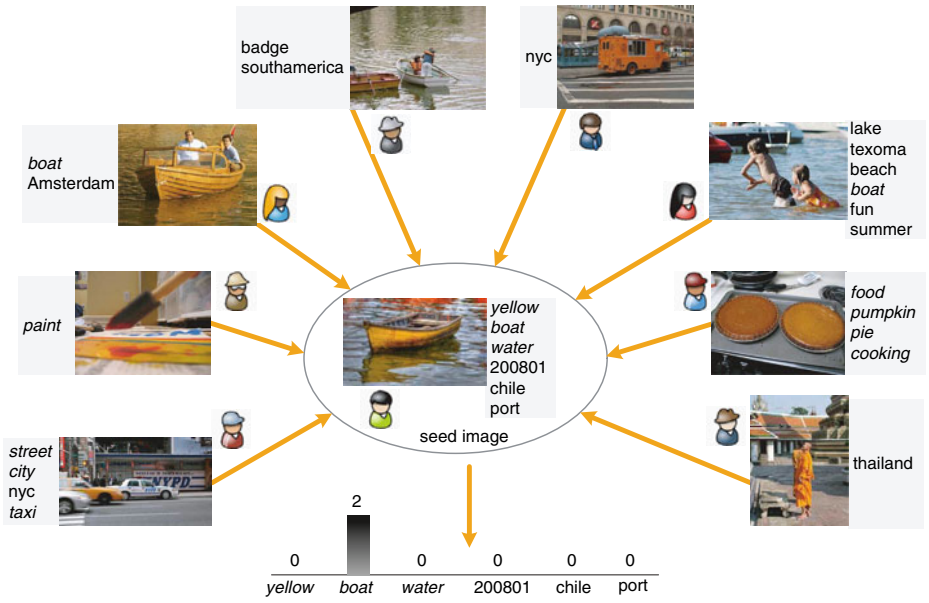

Fig. 4 Learning tag relevance by neighbor voting. The tag relevance value of each tag is estimated by accumulating the neighbor votes it receives from visually similar images of the seed image. In this example, since two neighbor images are labeled with boat, the tag relevance value of *boat* with respect to the seed image is 2

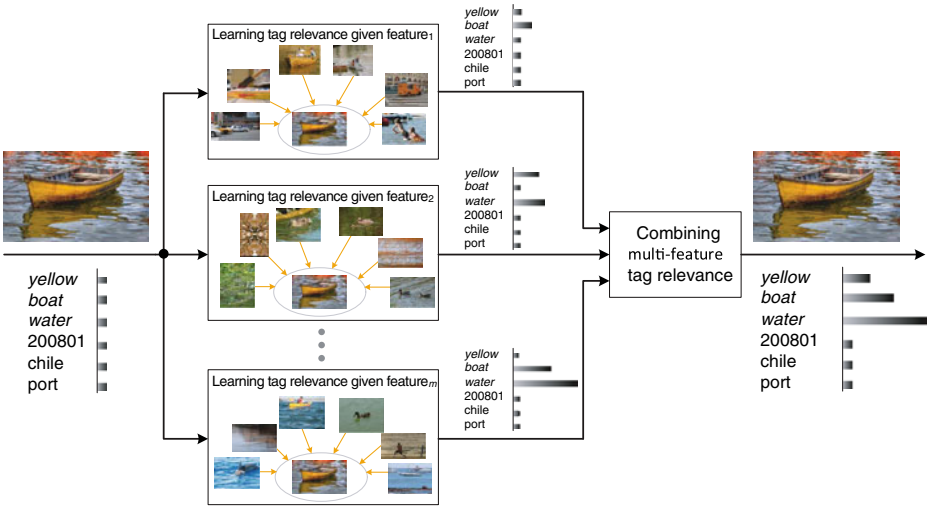

Fig. 5 Multi-feature tag relevance learning. Using a neighbor voting algorithm as a single-feature base learner, the authors propose to improve tag relevance learning by combining the output of many base learners obtained with different features and model parameters

stated by the authors [16], no single feature is able to represent the visual content completely, e.g., global features are suitable for capturing the gist of scenes, while local features are better for depicting objects. Therefore, Li et al. [16] further propose a multi-feature tag relevance learning method. Using the neighbor voting algorithm as a single-feature base learner, the proposed method is able to further improve tag relevance learning by combining the output of many base learners obtained with different features and model parameters. The schematic illustration of multi-feature based tag relevance learning can be shown in Fig. 5.

Pros and cons The neighbor voting based tag relevance learning method inherits the simplicity of the data-driven techniques. Moreover, involving the K-means clustering as a feature indexing mechanism further enhances its scalability in the large scale applications. However, there are still two limitations in the above method. First, the correlations between different tags are not exploited, which makes the learning process rely on visual clues solely, and thus the algorithmic performance is limited. Second, method also suffers from the ignorance of the underlying semantic concepts in the estimation of pairwise image similarities, which is similar to the deficiency of the tag ranking in Section 2.1.

3 Tag refinement

As aforementioned, the tags associated with the social images are frequently imprecise and incomplete, and many of them are almost only meaningful for the image owners. Recent studies reported in [5, 10] reveal that the user-provided tags associated with those social images are rather imprecise, with only about 50% precision rate. Moreover, the average number of tags for each social image is rather

small, which is far below the number required to fully describe the content of an image. Take Fig. 1 again as an example, we can observe that only "dog" and "leaves" truly describe the visual content of the image, and the other tags are imprecise or subjective. Meanwhile, some other tags that should be used to describe the visual content are missed, such as "grass" and "tree". Moreover, if we further consider the concepts' lexical variability and the hierarchy of semantic expressions, the tags such as "puppy", "pooch" and "canine" also need to be added.

The *imprecise*, *biased* and *incomplete* characteristics of the tags have significantly limited the performance of social image search and organization. For example, they will degrade precision and recall rates in tag-based image search. This deficiency demands a series of research on image tag refinement which aims at improving the quality of the tags (in terms of describing the real content of images), thus the performance of tag-based applications can be improved. There exist some recent efforts towards this dimension [3, 12, 18, 29]. In this section, we will review two representative works on this topic and discuss their advantages and shortcomings.

3.1 Image retagging based on visual and semantic consistency

As an initial effort, Liu et al. [18] propose a social image retagging scheme that aims at improving the quality of the tags. The basic assumption in this work is the consistency between visual similarity and semantic similarity in social images, that is, visually similar images tend to be assigned with similar tags, and vice verse. Based on this assumption, the tag refinement task is formulated as an optimization framework which tries to maximize the consistency while minimize the deviation from initially user-provided tags, which explicitly mines the information from different information channels in a collective way. However, the consistency assumption is mainly applicable for the "content related" tags, i.e., those tags that have high probability to describe the visual content of the images [21, 27]. If involving the "content unrelated" tags into the optimization process, the performance of the algorithm will be degraded. To solve this difficulty, the authors propose a tag filtering procedure to filter out those content unrelated tags by taking advantage of Wordnet taxonomy and domain knowledge in vision field. Specifically, five categories including "organism", "artifact", "thing", "color" and "natural phenomenon" are selected as high-level abstract concepts related to visual content. Then the decision of the visual properties of the tags is transformed into a word match problem where each tag is traversed along one path in Wordnet lexicon until one of the pre-defined visual categories is matched. If the match succeeds, the tag is decided as content-related, and otherwise it is decided as content-unrelated. Another favorable property of the proposed image retagging scheme is an effective tag enrichment component that expands each tag with appropriate synonyms and hypernyms by mining the Wordnet lexical knowledge base as well as the statistic information on social image website. The whole framework of the image retagging scheme is illustrated as in Fig. 6.

Pros and cons The encouraging characteristics in the above image retagging scheme can be summarized as follows: (1) Differentiating tags into the category of content-related or content-unrelated can be used to improve the performance of the content analysis algorithm. (2) The proposed tag enrichment component shows a good example on how to invent new tags, i.e., those tags that are not initially provided by

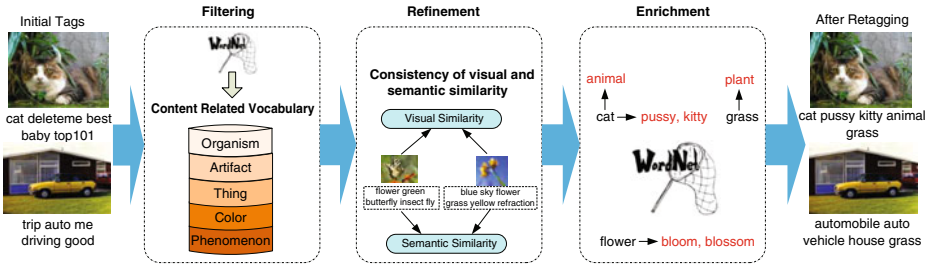

Fig. 6 The schematic illustration of the image retagging approach. Tag filtering is first adopted to filter out content-unrelated tags. Then an optimization algorithm is performed to refine the tags. Finally, we augment each tag with its synonyms and hypernyms

the users, to the social image collection. Meanwhile, the main issue with the proposed image retagging scheme is that it only works on a closed social image collection with fixed number of images and tags, and thus lacks an efficient strategy to handel the out-of-sample images, which limits its applicability in the dynamic social image collections on the Web.

3.2 Image tag refinement towards low rank, content tag prior and error sparsity

Inspired by the efforts in [18], Zhu et al. [29] further proposes an image tag refinement method motivated by the following four observations on large volume social image collections. (1) The semantic space of text information is typically approximated by a small subset of salient words derived from the original space. As a special kind of text information, image tags are also subject to such low-rank property. (2) Visually similar images are typically annotated with similar tags, which shows up the property of content consistency. (3) Semantic tags do not appear in isolation, instead they often appear correlatively and interact with each other at the semantic level. (4) The tagging results for each image are reasonably accurate to certain level, thus lead to error sparsity for the entire image tag matrix. By employing the nuclear norm, ℓ_1 norm and trace norm to model the low-rank, error sparsity, content consistency and tag correlation respectively, the tag refinement task is cast into a convex optimization problem, which simultaneously minimizes the matrix rank, priors and error sparsity. To obtained the tag refinement results, an efficient convergence provable iterative procedure is proposed to accomplish the optimization. Figure 7 illustrates the framework of image tag refinement towards low-rank, content consistency, tag correlation and error sparsity.

Pros and cons In the proposed tag refinement method, the incorporation of the low-rank and error sparsity properties on the refined image tag matrix well captures the desired relationship between the visual contents and the semantic tags, which leads to improved tag refinement performance than the method in [18]. However, it still suffers from the deficiency in handling the out-of-sample images, which hampers its effectiveness in real world applications. Besides, the optimization of nuclear norm is tackled with singular value decomposition (SVD), which is computational intensive and consequently limits its scalability in large-scale problems.

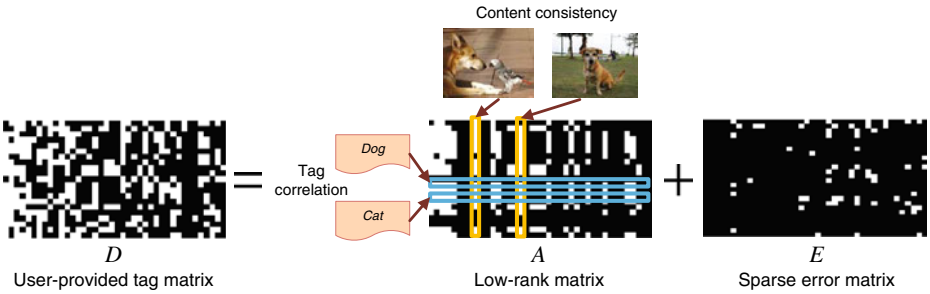

Fig. 7 The framework of image tag refinement towards low-rank, content consistency, tag correlation and error sparsity. The column-wise user-provided tag matrix D, where white grid represents the association of a tag with image and black one represents non-association, is decomposed into a low-rank matrix A (the refined tag matrix and here $rank(A) = 13$) and a sparse matrix E (tagging error in user-provided tags and sparse error is $\|E\|_0 = 72$ in this illustration) by considering the properties of content consistency and tag correlation

4 Tag-to-region assignment

To achieve reliable and visible content-based image retrieval systems, it is critical to obtain the exact correspondence between the tags and the individual regions within an image. However, in practice, it is a labor-intensive task to manually assign each tag to its corresponding region, and most users are willing to annotate the tags at the image level. This inspires an interesting and practically valuable research problem which automatically reassign the tags annotated at the image level to those derived semantic regions. i,e., the so called Tag-to-Region Assignment (TRA) problem. Figure 8 illustrates the problem inputs, i.e., images annotated with tags at the image-level, and the problem outputs, i.e., semantic regions with assigned tags, for the tag-to-region assignment task.

There exist some related works [2, 4, 23] in computer vision community, known as simultaneous object recognition and image segmentation, which aims to learn explicit detection model for each class/tag, and thus inapplicable for the TRA task due to the difficulties in collecting precisely labeled training regions for each tag. Besides, all the spatially connected objects within an image need to be assigned with proper tags through TRA, which may also challenge those conventional unsupervised learning

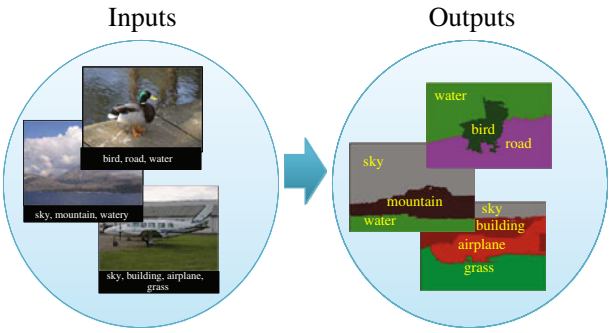

Fig. 8 Illustration of the tag-to-region assignment task. Note that no data with ground truth label-to-region relations are provided as priors for this task

algorithms. In the following, we will discuss two initial efforts towards this direction from multimedia research community.

4.1 Tag-to-region assignment by bi-layer sparsity priors

As the first effort toward tag-to-region assignment in multimedia research community, Liu et al. [20] propose a bi-layer sparse coding formulation for uncovering how an image or semantic region could be reconstructed from the over-segmented image patches of the entire image repository. The basic idea is that an image or semantic region can be sparsely reconstructed via the patches belonging to the images with common tags, and an additional constraint which enforces the reconstructing patches to be selected from as few images as possible imposes a second layer of sparsity. Each layer of the sparse coding above assigns the image-level tags to those selected reconstructing atomic patches and merged candidate regions according to the shared image tags. The results from all assignment results over all the candidate regions are then fused to obtain the entire result of tag-to-region assignment. Besides, the authors also apply the bi-layer sparse coding framework to perform multi-label image annotation on the new test images.

Pros and cons The above tag-to-region assignment method has the following advantages: (1) The process does not require the ideal image segmentation, which is still beyond the capabilities of those existing algorithms. (2) No statistical model is constructed for each tag, and thus is scalable to applications with large tag set. (3) The usage of the merged atomic patches within an image to reconstruct the reference image or semantic region guarantees the reliability of tag propagation. However, the method suffers from the possible ambiguities among the over-segmented atomic patches which are not descriptive enough, and thus the algorithmic performance is limited.

4.2 Tag-to-region assignment by multi-edge graph

Recently, Liu et al. [19] propose a new concept of multi-edge graph, and further apply it into the task of tag-to-region assignment. Figure 9 illustrates the definition of multi-edge graph along with its application in the TRA task. In the multi-edge graph model, each vertex is characterized as a unique image, which is encoded as a "bag-of-regions" representation with multiple segmentations, and the thresholding of the pairwise similarities between the individual image regions naturally constructs the multiple edges between each vertex pair. Based on the graph structure, the tag-to-region assignment is described as a cross-level tag propagation which propagates and adapts the tags between a vertex and its connected edges, as well as between all edges in the graph. A core vertex-vs-edge tag relation equation is derived to bridge the image/vertex tags and the region-pair/edge tags. That is, the maximum confidence scores over all the edges between two vertices indicate the shared tags of these two vertices. Based on this core equation, the tag-to-region assignment is formulated as a constrained optimization problem, where the objective characterizing the cross-region tag consistency is constrained by the core equations for all vertex pairs, and the cutting plane method is utilized for efficient optimization. The multi-edge graph model well captures the relationship between image tags and the regions tags through

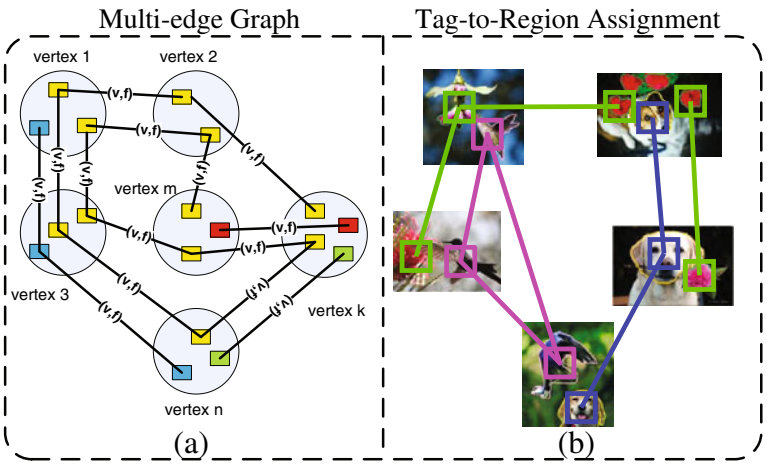

Fig. 9 **a** An illustration of multi-edge graph, where each vertex pair is connected with multiple edges. **b** For tag-to-region assignment, each edge connects two segmented image regions from two unique images/vertices, and thus each image/vertex pair is generally connected with multiple edges, where the number of edges for each image/vertex pair may be different

an appropriate core equation, and shows better performance than bi-layer sparse coding. Besides, the proposed multi-edge graph is a unified formulation and solution that can be utilized in other tag analysis tasks such as tag refinement and automatic tagging.

Pros and cons The following advantages can be considered when performing tag-to-region assignment with multi-edge graph. (1) The bag-of-regions image representation uses multiple image segmentation algorithms to simultaneously segment an image, which, on the one hand, relives the limitations of image segmenting, and on the other hand, avoids the ambiguities among the smaller size patches in the bi-layer sparse coding method. (2) The vertex-vs-edge core equation provides an simple and effective method to bridge the image tags and the region tags, which opens a promising solution for region-based content analyzing. The main shortcoming with respect to the proposed method is that the multi-edge graph still works under the transductive setting, making it difficult for handling the out-of-sample images.

5 Open issues

The exciting developments in tag processing for Internet social images are actually accompanied with some challenging open issues. Below we list a few important aspects which can be a part of future research.

5.1 Cross-modality content analysis

Multimodality analysis techniques have been shown effective for multimedia search, annotation and detection, where the visual contents and textual words are independently applied to build learning models, followed by a model fusion step with

a certain strategy. However, it is still difficult to effectively and automatically fuse different models obtained from diverse information channels for different applications. Meanwhile, social images and their associated tags are directly correlated, which provides a valuable clue to simplify the multimodality analysis. Therefore, a promising future direction is to learn an intermediate representation that maximizes the correlation between the visual content and semantic tags. In this way, both the tags and the image contents can be represented as consistent feature vectors in the same space, and the simple *k*-nearest neighbor classifier will be sufficient for the followup content analysis tasks.

5.2 Visual understanding using tag cues

Thus far, the focus of existing tag analysis works has been on improving the quality of the tags as image content descriptors, while little efforts have been devoted to the opposite direction, namely, using the tags as implicit contextual cues to boost visual understanding. Actually, the tags associated with a social image implies fruitful contextual information about the visual content. For example, an image's tag list is sufficient to reveal the underlying semantic theme of the image, which can be used to assist the decision of the automatic concept detection models. Another example is the relative order of a tag in the tag list of an image, where the more ahead a tag lies, the more prominent the corresponding object will be within the image. If we incorporate this cue into the process of object recognition or localization, their performance will be further improved. By adding more such contextual cues from the tags, we can expect that a more reliable visual understanding mechanism could be established.

5.3 Efficient manual tagging system design

Most online image sharing systems adopt two kinds of manual tagging approaches to facilitate the users in photo tagging. The first one is exhaustively tagging, in which the users provide tags for each individual image. This approach tends to result in relative high tagging accuracy, but the drawback is its high labor cost. The second approach is batch tagging, in which the users can assign tags to a suite of images. However, directly applying this approach to the whole image collection will introduce significant imprecise tags for many images. Based on the above observations, we argue that there is a dilemma between manual efforts and tagging accuracy. How can we precisely annotate an image collection with moderate manual efforts? Apart from accuracy and efficiency, a controllable tagging procedure which allows the users to adjust the tagging accuracy according to their preferences is another important factor that should be taken into account.

5.4 Scalable automatic tagging

As manual tagging is not scalable and very expensive when the volume of image repository becomes large, many of the vast quantity of the images that have been uploaded onto the Internet remain unlabeled with indexing tags. Therefore, an automatic process to predict the tags for the images in a huge image collection is highly desirable. How can we automatically annotate a large-scale image collection

with hundreds of thousands of tags? Here we propose some promising directions to exploit. (1) Develop scalable statistical learning algorithms to handle large scale training data with huge number of tags. (2) Leverage hashing techniques to realize the search-based automatic tagging.

6 Conclusion

We have presented a comprehensive survey highlighting the current progress and emerging directions to the exciting research topic of content-based tag processing for Internet social images. A number of representative works on tag ranking, tag refinement and tag-to-region assignment are discussed in details, and the specific future directions are conjectured alongside. We believe that the research topic will experience a booming in the future, with the focus being on utilizing tags as an intermediate vehicle to realize large-scale content-based image retrieval.

Acknowledgements The authors would like to thank Xirong Li, Xiaobai Liu, and Dr. Guangyu Zhu who contribute the figures illustrating their works introduced in this paper.

References

1. Anderson P (2007) What is web 2.0? Ideas, technologies and implications for education. JISC technical report
2. Cao L, Fei-Fei L (2007) Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: IEEE ICCV
3. Chen L, Xu D, Tsang I (2010) Tag-based web photo retrieval improved by batch mode re-tagging. In: IEEE CVPR
4. Chen Y, Zhu L, Yuille A, Zhang H-J (2009) Unsupervised learning of probabilistic object models (poms) for object classification, segmentation, and recognition using knowledge propagation. In: TPAMI
5. Chua T, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from National University of Singapore. In: ACM CIVR
6. Datta R, Joshi D, Li J, Wang J (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
7. Feng S, Lang C, Xu D (2010) Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking. In: ACM CIVR
8. Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR
9. Jing S, Baluja S (2008) VisualRank: applying pageRank to large-scale image search. TPAMI 30(11):1877–1890
10. Kennedy L, Chang S-F, Kozintsev I (2006) To search or to label?: predicting the performance of search-based automatic image classifiers. In: ACM MIR
11. Kennedy L, Slaney M, Weinberger K (2009) Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In: ACM WSMC
12. Lee S, Neve W, Ro Y (2010) Image tag refinement along the 'what' dimension using tag categorization and neighbor voting. In: IEEE ICME
13. Lew M, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. TOMCCAP 2(1):1–19
14. Li X, Snoek C, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: ACM MIR
15. Li X, Snoek C, Worring M (2009) Learning social tag relevance by neighbor voting. TMM 11(7):1310–1322
16. Li X, Snoek C, Worring M (2010) Unsupervised multi-feature tag relevance learning for social image retrieval. In: ACM CIVR

17. Liu D, Hua X-S, Yang L, Wang M, Zhang H-J (2009) Tag ranking. In: ACM WWW
18. Liu D, Hua X-S, Wang M, Zhang H-J (2010) Image retagging. In: ACM MM
19. Liu D, Yan S, Rui Y, Zhang H-J (2010) Unified tag analysis with multi-edge graph. In: ACM MM
20. Liu X, Cheng B, Yan S, Tang J, Chua T, Jin H (2009) Label to region by bi-layer sparsity priors. In: ACM MM
21. Lu Y, Zhang L, Tian Q, Ma W (2008) What are the high-level concepts with small semantic gaps? In: IEEE CVPR
22. Rattenbury T, Good N, Naaman M (2007) Towards extracting flickr tag semantics. In: ACM WWW
23. Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV
24. Sigurbjörnsson B, Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: ACM WWW, pp 327–336
25. Wang Z, Feng H, Yan S, Zhang C (2010) Learning to rank tags. In: ACM CIVR
26. Weinberger K, Slaney M, Zwol R (2008) Resolving tag ambiguity. In: ACM MM, pp 111–120
27. Yanai K, Barnard K (2005) Image region entropy: a measure of "visualness" of web images associated with one concept? In: ACM MM
28. Zha J, Yang L, Mei T, Wang M, Wang Z (2009) Visual query suggestion. In: ACM MM
29. Zhu G, Yan S, Ma Y (2010) Image tag refinement towards low-rank, content-tag prior and error sparsity. In: ACM MM

Dong Liu is currently pursuing the Ph.D. degree from School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. From January 2009 to November 2010, he worked as a research engineer in Department of Electrical and Computer Engineering, National University of Singapore. Prior to this, he worked as a research intern in the Internet Media Group at Microsoft Research Asia for two years. His research interests include multimedia information retrieval and analysis, machine learning, and computer vision. He has published over ten technical papers in the above areas. He won the Microsoft Research Asia Fellowship in 2009–2010.

Xian-Sheng Hua received the BS and PhD degrees in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a lead researcher with the Internet Media Group. His current research interests include video content analysis, multimedia search, management, authoring, sharing, and advertising. He has authored more than 130 publications in these areas and has more than 30 filed patents or pending applications. He is an adjunct professor at the University of Science and Technology of China, and serves as an associate editor of the IEEE Transactions on Multimedia and an editorial board member of Multimedia Tools and Applications. He won the Best Paper Award and the Best Demonstration Award at ACM Multimedia 2007 and also won the TR35 2008 Young Innovator Award from the MIT Technology Review. He is a member of the ACM and the IEEE.

Hong-Jiang Zhang received the BS and PhD degrees in electrical engineering from Zhengzhou University, Henan, China, in 1982, and the Technical University of Denmark, Lyngby, in 1991, respectively. From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a research manager at Hewlett-Packard Labs, Palo Alto, California, where he was responsible for research and development in the areas of multimedia management and intelligent image processing. In 1999, he joined Microsoft Research, where he is currently the managing director of the Advanced Technology Center in Beijing. He has coauthored/coedited four books, more than 350 papers and book chapters, numerous special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as more than 60 granted patents. Currently he is on the editorial board of the Proceedings of IEEE. He is a fellow of the IEEE and the ACM.