Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration

MENG WANG, National University of Singapore
BINGBING NI, Advanced Digital Science Center
XIAN-SHENG HUA, Microsoft
TAT-SENG CHUA, National University of Singapore

Along with the explosive growth of multimedia data, automatic multimedia tagging has attracted great interest of various research communities, such as computer vision, multimedia, and information retrieval. However, despite the great progress achieved in the past two decades, automatic tagging technologies still can hardly achieve satisfactory performance on real-world multimedia data that vary widely in genre, quality, and content. Meanwhile, the power of human intelligence has been fully demonstrated in the Web 2.0 era. If well motivated, Internet users are able to tag a large amount of multimedia data. Therefore, a set of new techniques has been developed by combining humans and computers for more accurate and efficient multimedia tagging, such as batch tagging, active tagging, tag recommendation, and tag refinement. These techniques are able to accomplish multimedia tagging by jointly exploring humans and computers in different ways. This article refers to them collectively as *assistive tagging* and conducts a comprehensive survey of existing research efforts on this theme. We first introduce the status of automatic tagging and manual tagging and then state why assistive tagging can be a good solution. We categorize existing assistive tagging techniques into three paradigms: (1) tagging with data selection & organization; (2) tag recommendation; and (3) tag processing. We introduce the research efforts on each paradigm and summarize the methodologies. We also provide a discussion on several future trends in this research direction.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: Annotation, tagging, interactive tagging, tag recommendation, tag location, tag refinement

ACM Reference Format:
Wang, M., Ni, B., Hua, X.-S., and Chua, T.-S. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. ACM Comput. Surv. 44, 4, Article 25 (August 2012), 24 pages.
DOI = 10.1145/2333112.2333120 http://doi.acm.org/10.1145/2333112.2333120

1. INTRODUCTION

Multimedia tagging[1] is referred to as the process of assigning a set of keywords to multimedia data to describe their content on semantic or syntactic levels. With such metadata, the management, summarization, and retrieval of multimedia content can be easily accomplished. For example, the tags can be directly used to index multimedia

[1] It is also widely known as *multimedia annotation* or *labeling*. For automatic tagging, it is also usually named *multimedia concept detection* or *high-level feature extraction*.

M. Wang is currently affiliated with Hefei University of Technology.
Corresponding author's email: M. Wang eric.mengwang@gmail.com.

DOI 10.1145/2333112.2333120 http://doi.acm.org/10.1145/2333112.2333120

data, and the desired data can thus be efficiently retrieved with a text query. For example, Flickr allows users to search images based on only their tags. TRECVID experience [Snoek and Worring 2009, Hauptmann et al. 2007] demonstrates that by predicting the relevant tags of video clips, video search performance can be greatly boosted.

Since generally manual tagging is a time-consuming task and requires intensive labor costs, automatic multimedia tagging has attracted great research interest, and years of efforts have been dedicated to this research field. Several competitions and challenges have also been organized towards solving the automatic multimedia tagging problem, such as the object classification task of PASCAL VOC that starts from 2005, the annotation task of imageCLEF that starts from 2005, and the high-level feature extraction task of TRECVID that starts from 2002. Here, automatic tagging is defined as the process that automatically assigns a set of tags to multimedia data without any interference from humans. Typically, machine learning is employed to accomplish the tagging by mapping low-level features of multimedia content to semantic tags in a process as follows. First, a training set is constructed that contains a set of multimedia data associated with tags. Low-level features are extracted from the data to describe their content, and models are learned with these features and tags. Second, these models are applied to predict the tags of new data. External or context knowledge may also be explored in this process. Note that manual efforts may be required before the model learning step, such as training data collection and labeling, but in automatic tagging, there is no interference from human in the tag prediction process. Several recent research efforts have been dedicated to search-based tagging technology, which assigns tags to a new sample by finding its neighbors in a training set. But it also belongs to the learning-based automatic tagging paradigm, and the difference is that it uses lazy learning instead of building models for tags.

It has long been recognized that the main challenge of automatic multimedia tagging lies in the *semantic gap*, namely, the gap between the low-level features and tags. Although significant progress has been made on multimedia feature extraction and modeling techniques [Snoek and Smeulders 2010; Makadia et al. 2008; Sande et al. 2010; Wang et al. 2010, 2009], the performance of existing methods is still far from satisfactory: even for many simple concepts, the tagging precision and recall measurements are fairly low (several details will be introduced in the next section).

As there is little evidence to show that the semantic gap problem can be solved in the near future via automated machine learning approaches, exploring human intelligence for multimedia tagging attracts more and more interests. Actually, the power of collective human efforts has been fully demonstrated in the Web 2.0 era. For example, many photos and videos on Flickr and Youtube are tagged by their owners. Several games are designed to motivate Internet users to tag multimedia data [von Ahn and Dabbish 2004; von Ahn et al. 2006; Seneviratne and Izquierdo 2006; Jesus et al. 2008; Ho et al. 2009; Steggink and Snoek 2011]. Large-scale tagging platforms are also available, such as Amazon Mechanical Turk [Sorokin and Forsyth 2008]. However, although manual tagging of large-scale multimedia data becomes feasible, it still encounters the following problems.

(1) *Intensive labor and time costs.* Even with good motivation, users may be frustrated by the intensive labor cost of naive manual tagging. Heavy labor and time costs are obstacles for the tagging of large-scale data.
(2) *Tagging quality.* The tags provided by nonprofessional users are usually noisy, incomplete, and personalized. For example, existing studies on Flickr reveal that only 50% of the tags are really related to the images [Kennedy et al. 2006]. Low-quality

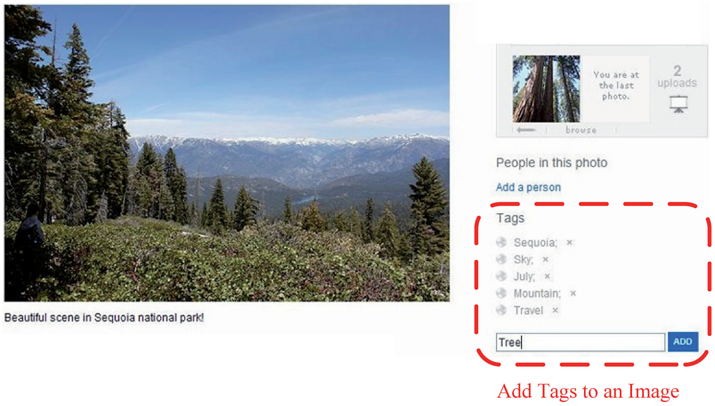
tags will produce difficulties in many tag-based applications, such as recommendation and search.

Since tagging relying on purely computers or humans have different difficulties, emerging research efforts are aiming at simultaneously exploring humans and computers for multimedia tagging. They can provide tags with higher quality than automatic tagging, as there is guidance from human's intelligence. In comparison with pure manual tagging, they are also able to provide better results as computers can correct several mistakes and need less labor cost, as users may not need to label all the data. Here we categorize the techniques into the following three paradigms.

(1) *Tagging with data selection and organization.* This paradigm aims to reduce tagging costs by only manually labeling several representative samples or improving tagging efficiency via intelligently organizing the to-be-tagged data.
(2) *Tag recommendation.* This paradigm suggests a set of candidates to labelers in the tagging process such that users can directly select the correct ones from the set of candidates. It can improve both the tagging efficiency and the quality of tags.
(3) *Tag processing.* This is defined as the process of refining human-provided tags or adding more information to them. For example, many tag refinement methods are proposed for improving the accuracy and completeness of tags, and some other methods analyze the relevance, saliency, and other characteristics of tags.

We collectively refer these techniques as *assistive tagging*, as their common principle is letting computers assist humans in tagging by reducing tagging cost or improving tagging quality. Thus, assistive tagging can also be defined as the process that assigns tags to multimedia data by jointly exploring humans and computers (automatically selecting and organizing data for manual labeling, automatically recommending tags for manual selection, and automatically processing human-provided tags are different kinds of *joint exploration*). In this article, we provide a survey on the existing research efforts along this direction. We start by introducing several industrial solutions to the multimedia tagging problem and the state-of-the-art methods of automatic tagging in the research community. We then introduce the research efforts on tagging with data selection and organization, tag recommendation, and tag processing separately. We do not only review existing algorithms but also analyze their implicit principles or cast them into unified schemes. We will focus on image and video tagging in the article, although many methods and principles can also be applied to other media types, such as music and speech. We do not take game-based tagging methods [von Ahn and Dabbish 2004; von Ahn et al. 2006; Seneviratne and Izquierdo 2006; Jesus et al. 2008; Ho et al. 2009; Steggink and Snoek 2011] into account, as they can be categorized to the tagging interface design problem. Actually, the assistive tagging methods can be combined with game-based tagging. For example, we can use data selection algorithms to select several representative samples and then tag them via the Extra Sensory Perception (ESP) game [von Ahn and Dabbish 2004].

As tagging plays an important role in multimedia information retrieval, it has been introduced in many survey papers on multimedia information retrieval [Smeulders et al. 2000; Lew et al. 2006; Rui et al. 1999]. However, most of them only introduce the research on automatic tagging. Liu et al. [2011] conduct a survey on the image tag processing, but it simply introduces several tag refinement and tag localization methods. To our knowledge, there is no dedicated review or survey on comprehensively introducing different assistive multimedia tagging paradigms, and this motivates our work.


Beautiful scene in Sequoia national park!

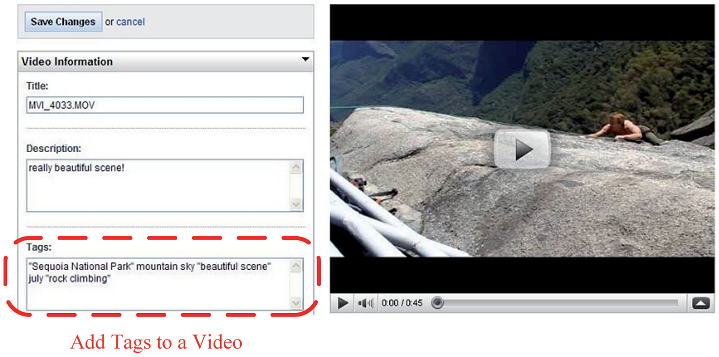
Tags

Sequoia; ×
Sky; ×
July; ×
Mountain; ×
Travel ×

Tree ADD

Add Tags to an Image

Fig. 1. The interface of image tagging on Flickr.


Video Information

Title: MVI_4033.MOV

Description: really beautiful scene!

Tags: "Sequoia National Park" mountain sky "beautiful scene" july "rock climbing"

Add Tags to a Video

Fig. 2. The interface of video tagging on Youtube.

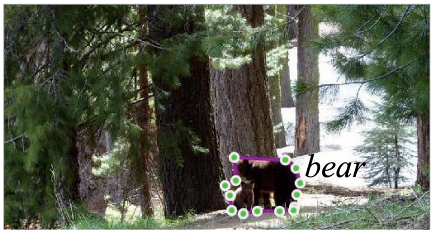
The rest of the article is organized as follows. In Section 2, we introduce the state-of-the-art approaches of manual tagging and automatic tagging and then introduce why assistive tagging attracts research interests. We then introduce the work on tagging with data selection & organization, tag recommendation, and tag processing in Sections 3, 4, and 5, respectively. In Section 6, we discuss the benchmark and database status for the research of assistive tagging. Finally, we conclude the paper and introduce future research directions in Section 7.

2. MANUAL, AUTOMATIC, AND ASSISTIVE MULTIMEDIA TAGGING

In this section, we briefly introduce the state-of-the-art approaches of manual tagging and automatic tagging and then explain why assistive tagging technology can be a solution for large-scale multimedia tagging.

2.1. Manual Multimedia Tagging

Manual tagging is the most direct approach, and it is the widely applied industrial solution. Actually, manual tagging is part of the collaborative nature of Web 2.0 [O'Reilly 2007]. Many multimedia sharing websites not only allow users to upload media data but also encourage them to label the content with descriptive tags. As representatives,



Add tag to a located region

Fig. 3. The interface of Labelme, which allows users to provide tags to a located region.

Figures 1 and 2 illustrate the tagging interfaces of Flickr[2] and Youtube[3], which allow users to manually type in tags for an uploaded image and video, respectively.

In addition to tagging photos one by one, Flickr has a batch tagging function that allows users to assign tags to a set of uploaded images. This is useful, as many photos are continuously captured with the same scene, and frequently, they should share the same tags. Batch tagging can save the labor cost of users, but it requires the users to organize related photos into a batch in order to facilitate batch tagging.

There are also approaches that support image and video tagging at finer granularities. For images we can add tags for specified regions, and for videos we can assign tags to subclips instead of the whole video. The fine-grained tagging is able to enable more accurate search of video content and the training of object detectors. On many videosharing websites, such as Youtube and Metacafe[4], tags are assigned at the video level. But video tends to be a linear experience that unfolds over a duration of time, and the tags' time information will be missed if they are assigned at the video level. This limits the usefulness of tags, especially for videos with long durations. Therefore, several websites, such as Veotag[5], Viddler[6] and MotionBox[7], also support the "tagging into video" service that enables users to assign tags at a specific location of a video. It allows users to type in tags in the video playing process, and users can also efficiently browse the video content via the tags. Figures 3 and 4 illustrate the interfaces of image tagging at Labelme [Russell et al. 2007] and video tagging at Veotag, which permit users to label images and videos at specified locations, respectively.

Some studies have been conducted on humans' tagging behavior. Sigurbjörnsson and Zwol [2008] provide some insights on how users tag their photos and what type of tags they are providing. Volkmer et al. [2005] introduce a Web-based system that allows multiple labelers to tag a dataset in a collaborative way. It has been used to collect the training dataset for TRECVID 2005, and it analyzes many statistics about human tagging. The study in Lin et al. [2003] shows that typically annotating 1 hour of video with 100 concepts can take anywhere between 8 to 15 hours, and the study in Yan et al. [2009] shows that a user will need 6.8 seconds to tag an image on average. Liu et al. [2008] demonstrate that user-provided tags are orderless. These studies demonstrate that although many strategies and interfaces have been designed to facilitate manual tagging, human tagging is still labor intensive and time consuming, and it frequently produces results that are noisy and incomplete. Therefore, a pure manual approach encounters many difficulties for large-scale multimedia tagging.

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

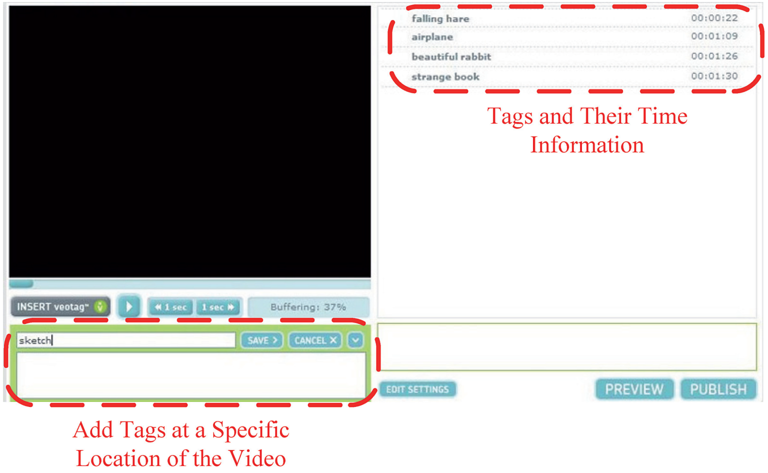
[3] http://www.youtube.com.

[4] http://www.metacafe.com.

[5] http://www.veotag.com.

[6] http://www.viddler.com.

[7] http://www.motionbox.com.



falling hare	00:00:22
airplane	00:01:09
beautiful rabbit	00:01:26
strange book	00:01:30

sketch

Tags and Their Time Information

Add Tags at a Specific Location of the Video

Fig. 4. The interface of video tagging on Veotag, which allows users to assign tags at a specific location of a video.

2.2. Automatic Multimedia Tagging

Automatic multimedia tagging usually employs machine learning, and the existing algorithms can mainly be classified into two categories, that is, model-based and search-based. The model-based methods build a model for each tag based on a training set and use these models to predict new data. Search-based methods predict the tags of a new sample via finding its neighbors in the training set.

In model-based methods, the prediction of each tag can be regarded as a binary classification problem, that is, each unit is predicted to be positive or negative according to whether it is associated with the tag. Therefore, almost all classification methods can be applied here. One approach is to treat tagging as a translation from image instances to tags, and it is usually accomplished based on some models that exploit the cooccurrences of image and text [Mori et al. 1999; Duygulu et al. 2002]. The translation approach of Duygulu et al. [2002] is extended to association modeling through latent topic/aspect/context spaces [Blei and Jordan 2003; Monay and Gatica-Perez 2003]. The Cross Media Relevance Model (CMRM) [Jeon et al. 2003], Continuous Relevance Model (CRM) [Lavrenko et al. 2004], and Multiple Bernoulli Relevance Model (MBRM) [Feng et al. 2004] adopt nonparametric density representations of the joint word-image space. Li and Wang [2008] developed a real-time ALIPR image annotation system that uses multi-resolution 2D hidden Markov models to model concepts based on a training set. Graph-based learning [Tang et al. 2011] and feature selection [Wu et al. 2010] have also been investigated. Naphade and Smith [2004] provide a survey on the video tagging algorithms applied to the TRECVID high-level feature extraction task, where a great deal of modeling methods can be found. For example, Amir et al. [2005] utilize a diverse set of learning methods for the TRECVID 2005 high-level feature extraction task, including support vector machine, Gaussian mixture models, maximum entropy methods, a modified nearest-neighbor classifier, and multiple instance learning. Jiang et al. [2010] conduct a study on SIFT-based representations for video tagging. There are also many works that investigate the relationship among tags for improving tagging performance. Qi et al. [2007] propose a multilabel learning algorithm that explores the semantic correlation of multiple tags in a kernel machine. Gao and Fan [2006] propose a hierarchical classification approach that explores the tree structure of ontology to accomplish image tagging.

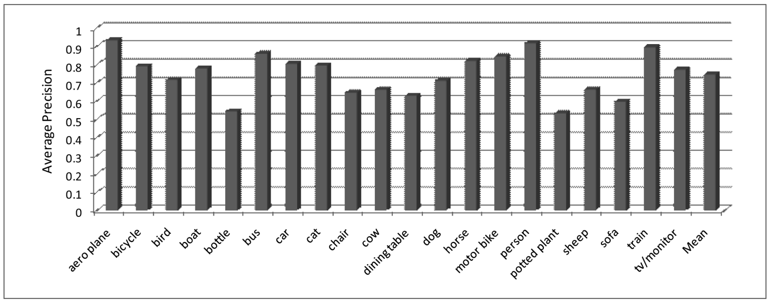


Fig. 5. The best tagging performance for each concept in the object classification competition of PASCAL VOC 2010.

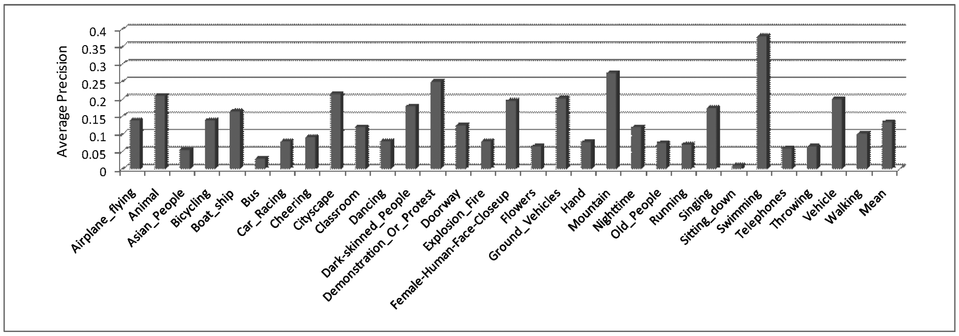


Fig. 6. The best tagging performance for each concept in the high-level feature extraction competition of TRECVID 2010.

Search-based methods can easily accomplish ontology-free tagging (i.e., the involved tags do not need to be constrained in a fixed dictionary), and via high-dimensional indexing technology, it is able to achieve good scalability with respect to the size of training data. Wang et al. [2006] collect 2.4 million high-quality Web images with abundant surrounding texts, and a new image is tagged by mining common phrases from the surrounding texts of its visually similar images. They further extend the method to a 2-billion Web image dataset in Wang et al. [2010]. Torralba et al. [2008] collect about 80 million tiny images of 32 × 32 pixels, and it is demonstrated that with simple nearest-neighbor methods, encouraging recognition performance can be achieved. Deng et al. [2009], on the other hand, structuralize 3.2 million Web images, and improved recognition performance assisted by image ontology is observed.

Despite the great progress made in the past years, the performance of the state-of-the-art image and video tagging algorithms is still far from satisfactory. Here, we analyze the performance by checking the results achieved in two benchmark challenges: the object classification task of PASCAL VOC and the high-level feature extraction task of TRECVID. Figures 5 and 6 illustrate the tagging performance for different concepts in PASCAL VOC 2010 and TRECVID 2010, respectively. We collect the results reported in Everingham [2010] and Over et al. [2010], and the best performance for each concept achieved in the evaluation tasks is shown. Average precision is adopted in both the two challenges (but it is worth mentioning that their average estimation methods are slightly different, and readers can refer to Everingham [2010] and Over et al. [2010] for more details). We can see that for many concepts that humans can easily recognize, such as "bottle" and "potted plant", the automatic image tagging performance is fairly low (average precision measurements are below 0.6). The problem is worse for video tagging. The average precision measurements for many concepts are even below 0.2. This indicates that the automatic tagging technologies may not be mature enough.[8]

2.3. Assistive Multimedia Tagging: Combining Humans and Computers

Up to now, we have discussed the strengths and weaknesses of manual tagging and automatic tagging. Assistive tagging, which can be regarded as their integration, lets the two approaches complement each other. Actually, it has long been realized that combining humans and machines can be a promising approach. Active learning is a good example of this approach [Huang et al. 2008; Goh et al. 2004; Ayache and Quénot 2007; Hauptmann et al. 2006]. It is a machine learning technique that learns models in an interactive way. When applied to the multimedia tagging problem, the active learning scheme will ask users to label samples that are selected by computers, and the concept models are learned based on the labeled samples. There are also other approaches that assist humans in tagging multimedia content, such as tag recommendation, as well as refining the human-labeled tags. In the following three sections, we introduce three assistive tagging paradigms in detail.

3. TAGGING WITH DATA SELECTION & ORGANIZATION

In this section, we introduce the approaches that make manual tagging more efficient by intelligently selecting training data or organizing the to-be-tagged data. The motivation of the approach comes from the fact that many multimedia contents in a database are usually closely related. For example, in a photo album, many photos are usually continuously captured and describe the same scene or object. Therefore, naively labeling the photos one by one is not an optimal approach, and the human labeling cost can be reduced by intelligently selecting or organizing data for tagging, as shown in Figure 7. We first introduce the existing works on tagging with data selection (including active learning-based multimedia tagging) and tagging with data organization, and we then summarize the underlying strategies employed in these methods. It is worth noting that data selection and organization are usually not separable. For example, several methods need to first cluster the data and then select samples accordingly. That is why we put them together in this section.

For multimedia tagging with data selection, active learning is the most widely adopted approach. A typical active learning system is composed of two parts, that is, model learning and sample selection. It works in an iterative way. In each round, the learning engine trains a model based on the current training set. The sample selection engine then selects the most informative unlabeled samples for manual annotation, and these samples are added to the training set. After that, the model is updated or retrained with the new training set. This process can iterate for many rounds. In this way, the obtained training set is more informative than that gathered by random sampling. Many active learning-based multimedia tagging methods have been proposed, and a comprehensive survey can be found in Huang et al. [2008] and Wang and Hua [2011].

Although active learning has shown its effectiveness in multimedia tagging, typically, it can only be applied to a fixed ontology of tags and can hardly generalize to

[8] Recent studies demonstrate that with a suitable query-tag mapping component, multimedia search performance can be greatly boosted even when tagging performance is not very good [Hauptmann et al. 2007; Snoek and Worring 2009]. But there is still little evidence to show that automatic tagging technology is sufficiently mature to enable a practical multimedia search engine with good performance (it is shown in Beitzel et al. [2005] that existing Web search engines are able to achieve an MAP measure of 0.65 for text search).

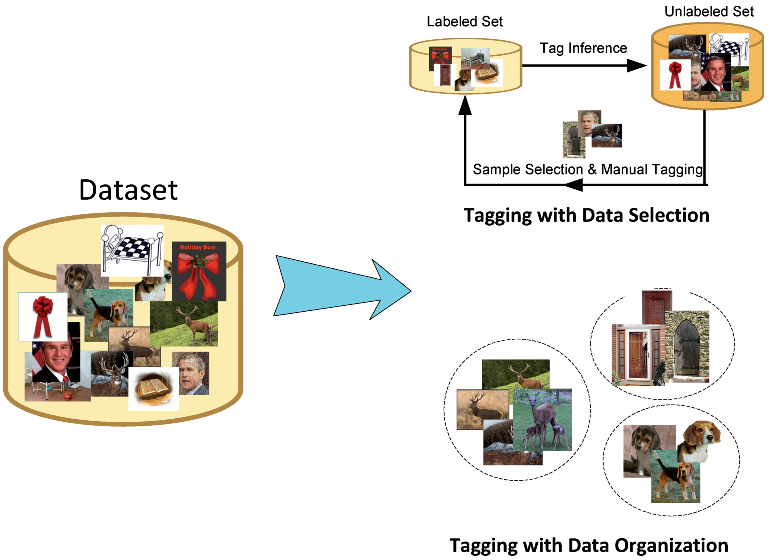

Labeled Set
Unlabeled Set
Tag Inference
Sample Selection & Manual Tagging
Tagging with Data Selection
Dataset
Tagging with Data Organization

Fig. 7. Tagging cost can be reduced by two approaches: (1) only tagging several representative samples and then the tags of the rest of the data are automatically inferred; and (2) improving tagging efficiency by intelligently organizing the data.

ontology-free tagging, as it needs to train models for the tags. To address this problem, Yang et al. [2009] propose a sample selection method for ontology-free photo tagging. Given a photo set with several samples that have already been assigned tags, a set of photos are selected based on three criteria, that is, *ambiguity* that estimates the semantic consistency of a sample's tagged neighbors, *citation* that estimates the representativeness of the selected photos, and *diversity* that enforces the selected photos to be diverse. Users manually tag the selected photos, and the tags of the rest of the photos are automatically inferred. Liu et al. [2010] propose an interactive photo album tagging approach based on sample selection via affinity propagation. With an affinity propagation method that simultaneously considers the visual content of photos and their temporal information, a set of exemplars that compactly represent the whole photo set are selected each time. After tagging the selected photos, the tag inference accuracy in the last round is estimated, and it is then decided whether the process should continue.

Besides tagging a set of informative data and inferring the remained data, intelligent data organization is also helpful in improving the tagging efficiency. For example, photo clustering technology has been widely explored in face tagging. By clustering photos based on the features extracted from face regions, it is expected that photos with the same faces can be grouped together, and thus, the tagging cost can be reduced as the photos in such clusters only need to be tagged once. Suh and Bederson [2004] adopt such an approach. But this method heavily depends on the clustering performance, as the tagging performance may severely degrade if the clustering performance is not good. Cui et al. [2007] introduce a reranking component for face tagging. When users click one photo/cluster, the algorithm re-ranks the photos/clusters and puts similar ones closer. It also allows users to label photos in the browsing and search process. Tian et al. [2007] adopt an interactive approach that works in an iterative way. It first groups faces with partial clustering. Similar faces will be grouped into a cluster, and the other faces are regarded as the background cluster. Information is learned in the tagging process of clusters, and several faces in the background cluster are clustered

Table I. The Categorization of the Works on Tagging with Data Selection & Organization According to the Three Strategies

Work	Strategy 1	Strategy 2	Strategy 3
Tagging based on Active Learning[9]	√		
Yang et al. [2009]		√	
Liu et al. [2010]		√	
Suh and Bederson [2004]		√	
Cui et al. [2007]		√	√
Tian et al. [2007]		√	
Tang et al. [2010]		√	
Yan et al. [2009]	√		√

Note: (1) Tagging a set of selected samples and predicting the rest data. (2) Using cluster as tagging unit such that more than one sample can be tagged each time. (3) Alternating sample order or tagging style according to human tagging behavior.

for tagging. Tang et al. [2010] propose an efficient image-tagging approach with region clustering. Each image is segmented into regions and then clustering is performed on the regions. Users then tag a cluster of regions each time, and cluster refinement is also supported. In this way, a tag can be assigned to hundreds or even thousands of images each time.

Yan et al. [2009] investigate the cost modeling of manual image tagging. They categorize manual tagging into two types, namely, image-oriented tagging and concept-oriented tagging. In image-oriented tagging, users label each image with a chosen set of keywords before proceeding to the next one, while in concept-orientated tagging, users browse images sequentially and judge each image's relevance to a concept. Linear models are adopted to predict the costs of the two approaches. For image-oriented tagging, its expected time cost is predicted as $t = K_l t_f + t_s$, where K_l is the number of concepts associated with the image, t_f is the average time of entering typing in a keyword, and t_s is the initial setup time for annotation. For concept-oriented tagging, its expected time cost is predicted as $t = L_k t_p + (L - L_k)t_n$, where L is the number of all images, L_k is the number of relevant images, and t_p and t_n are the time costs of labeling a relevant image and an irrelevant image, respectively. Based on the two models, Yan et al. [2009] propose two methods for organizing the tagging task via alternating between image-oriented and concept-oriented tagging, with the objective of minimizing human tagging time cost.

Overall, from the preceding discussion, we can see that these approaches generally adopt the following three strategies to achieve more efficient tagging: (1) tagging a set of selected samples and predicting the remained data; (2) using clusters as a tagging unit so that more than one sample can be tagged each time; and (3) alternating sample order or tagging style based on human tagging behavior. Table I summarizes the categorization of the preceding works according to the three strategies.

4. TAG RECOMMENDATION

Tag recommendation is an effective approach for improving both the tagging efficiency and accuracy by suggesting a set of relevant tags for a given image or video clip. First, tag recommendation can reduce the manual cost of tagging, as it will be faster for users to click on suggested candidates than typing the whole words. Therefore, if many relevant candidates are recommended, labelers just need to click them, and it will save much manual cost. On the other hand, a good tag recommendation scheme is also able

[9] See [Huang et al. 2008], [Wang and Hua 2011] and the references therein.

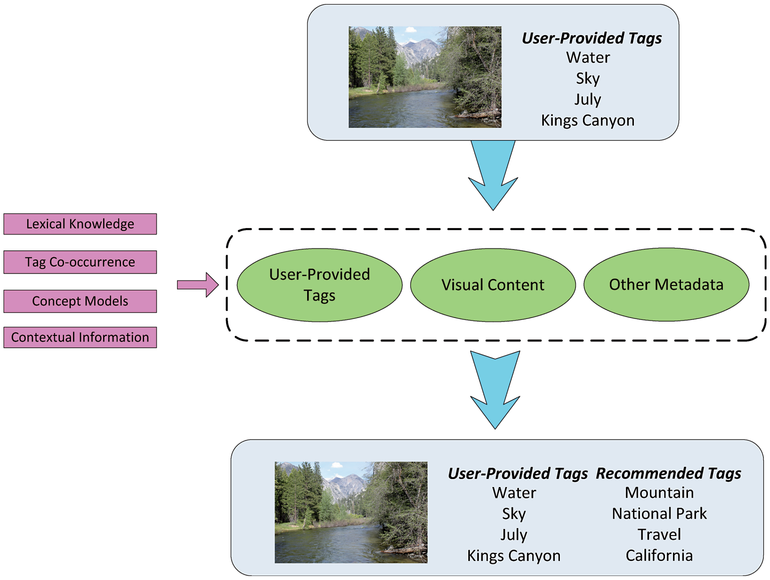


Fig. 8. A schematic illustration of multimedia tag recommendation. A set of tags are provided by exploring the multimedia content, user-provided, tags, and other metadata or contextual information.

to improve the tagging accuracy. Existing studies reveal that user-provided tags are usually limited in free tagging manner [Sigurbjörnsson and Zwol 2008]. One reason is that it is difficult and often requires high mental focus to figure out a lot of words to describe image or video content in a short moment. Tag recommendation is able to remind the labelers of alternative tags and also to help clarify the semantics of multimedia content. In addition, it also helps to reduce misspellings and meaningless words.

Figure 8 shows the schematic illustration of multimedia tag recommendation. Typically, there are three available information sources, that is, multimedia content, user-provided tags, and other metadata or contextual information, such as the titles and descriptions of multimedia data and users' social information. We now organize the existing works based on the information sources they leverage.

Tag recommendation by analyzing multimedia content is the most natural approach. It is actually similar to automatic tagging, as they both aim to find relevant tags for an image or a video clip. A difference can be that when applied to online applications, the computational efficiency is more critical for tag recommendation, as it is unacceptable for users to wait a long time for recommending tags. Shen and Fan [2010] employ a structured SVM algorithm to learn models from a large set of loosely labeled dataset, and these models are employed to predict the tags for new data. Abbasi et al. [2009] propose a clustering-based image tag recommendation approach. They group a large set of labeled images into several clusters, and for each cluster, several representative tags are generated via voting. Given a new image, they then find the closest cluster and recommend the representative tags of this cluster. Lee et al. [2010] propose a localized generative method. The tags are recommended based on their posterior probabilities, but these probabilities are estimated based on only the visually similar images of the target image. Chen et al. [2008] generate several representative tags for each Flickr group by voting. For a given image, they first perform concept detection for the image and then find Flickr groups with the detected concepts. Thereafter, the tags that are associated with these groups are collected and ranked for recommendation.

Toderici et al. [2010] propose a tag recommendation approach based on the content of user-uploaded videos. They train more than 20,000 models based on the audiovisual features with a large set of Youtube videos. These models are used to predict the tags of new videos for recommendation. In addition, they also mine rules for predicting the categories of videos based on their associated text information, whereby a set of tags that describe category information can also be recommended for each uploaded video.

Exploring user-provided tags is another approach. It usually consists of the following steps. Given an image or video clip, a user types in some initial tags. Then, for each user-provided tag, an ordered candidate list is generated according to certain correlation measures. Finally, the lists are aggregated to generate the final ranked recommendation. Sigurbjörnsson and Zwol [2008] propose such a system for Flickr photos. Their approach is based on the cooccurrence measurements of tags. For each user-provided tag, the highly correlated tags are selected as candidates. They conduct a study on different methods for integrating and ranking the candidates obtained with all user-provided tags. Statistics show that for about 70% of the photos, their method can provide a relevant recommendation at the first position; while for about 94% of the photos, it can provide a good recommendation for the top five ranks. On average, more than half of the recommended tags will be relevant if the top five tags are recommended for each photo. Wu et al. [2009] improve the method by exploring different correlation measurements of tags. Their method estimates the correlation of each tag pair via three different methods: (1) cooccurrence computation, that is, the method used in Sigurbjörnsson and Zwol [2008]; (2) visual content-based correlation estimation, which computes the similarity of image sets that are associated with the two tags; and (3) image-conditioned similarity estimation, which computes the similarity of the tags that considered the content of the specific image. Based on these three correlation measurements, it generates $3n$ features for each tag, where n is the number of user-provided tags. It then applies RankBoost [Freund et al. 1998] to rank the tags in folksonomy. Garg and Weber [2008] propose a personalized tag recommendation scheme that explores the tagging history profile of users. They empirically study different methods, and the conclusion is that combining the tag coconcurrence information and the personal tagging history achieves the best results. Weinberger et al. [2008] propose a tag recommendation scheme that not only aims to provide potentially reliable tags but also tries to resolve the ambiguity of the existing tags. Given an image with an existing set of tags, the algorithm finds two tags that frequently cooccur with the existing tag set but appear in different contexts. In this way, the ambiguity of the existing tag set can be reduced by choosing one of the recommended tags.

Several methods integrate both multimedia content and user-provided tags for tag recommendation. Anderson et al.'s [2008] method is close to that of Sigurbjörnsson and Zwol [2008]. For each user-provided tag, a ranking list of candidates is generated based on the cooccurrence relationship of tags. But a list of candidates is further generated based on the visual content of photos. The ranking lists are integrated by the Borda voting method. Sevil et al. [2010] propose a photo tag recommendation method that leverages both the content of photos and user-provided tags. Given a photo and several initial user-provided tags, photos that contain these tags in the database are retrieved. They then collect the tags that the retrieved photos are associated with, and these tags are weighted according to the similarity of the target tag to the retrieved photos.

Besides multimedia content and user-provided tags, several tag recommendation methods exploit the metadata associated with multimedia data and contextual information. Chen et al. [2010] propose a tag recommendation scheme for Web videos by using a search engine. For a given video, it constructs a query based on its title and existing tags, and the query is submitted to Google. They then collect the results and extract tags from the titles and abstracts of the retrieved results, and the tags are

Table II. The Categorization of the Works on Tag Recommendation Based on the Usage of Three Information Sources

Work	Content	Tags	Metadata and Contextual information
Abbasi et al. [2009]	√		
Lee et al. [2010]	√		
Chen et al. [2008]	√		
Toderici et al. [2010]	√		
Sigurbjörnsson and Zwol [2008]		√	
Wu et al. [2009]		√	
Garg and Weber [2008]		√	
Anderson et al. [2008]	√	√	
Sevil et al. [2010]	√	√	
Chen et al. [2010]		√	√
Shen and Fan [2010]	√		
Rae et al. [2010]		√	√
Ballan et al. [2010]	√		√
Weinberger et al. [2008]		√	
Naaman and Nair [2008]		√	√

ranked for recommendation. Ballan et al. [2010] propose a video recommendation scheme by mining a large Flickr image database. Given a video with several user-provided tags, they first retrieve Flickr images using these tags. The neighbors of the keyframes extracted from the video are then found, and their tags are collected for recommendation. This method is also able to supplement the tags that already exist in the video tag list with time information. Rae et al. [2010] investigate the usefulness of personalized and social information for photo tag recommendation. Four information sources are investigated, namely, (1) all the photos in the system, (2) the uploader's own photos, (3) the photos of the uploader's social contacts, and (4) the photos posted in the groups of which the uploader belongs to. The results of different combinations of the information sources are demonstrated. Naaman and Nair [2008] introduce a tag recommendation scheme for photos captured on mobile phones based on rich context information, including (1) the user's tagging history; (2) the user's location information; (3) the tagging history of the user's social contacts; and (4) the tags' temporal information.

Table II illustrates the categorization of these works according to the information sources they have used.

5. TAG PROCESSING

Tag processing is referred to as the approach that automatically processes user-provided tags such that more accurate tags or richer information can be obtained. Existing studies on the behavior of manual tagging, especially on the community-contributed tags, demonstrate that user-provided tags are usually noisy and incomplete. The study in Kennedy et al. [2006] shows that when a tag appears in a Flickr image, there is only about a 50% chance that the tag is really relevant. Sigurbjörnsson and Zwol's [2008] study shows that more than half of Flickr images are associated with less than four tags. Their study also demonstrates that many user-assigned tags do not describe images' visual content, and instead, they are related to the context of the images, such as location and time. Many works demonstrate that with more accurate tags or assigning tags richer information, better search experience could be obtained. For example, by assigning video tags temporal information, localized video search can be facilitated, that is, we can search subclips in videos [Ulges et al. 2008; Li et al. 2011]. By refining noisy tags, the performance of tag-based search can be improved

Liu et al. 2008, 2010]. In this section, we introduce two main schemes: tag information supplementation and tag refinement. Here, we note that most of the existing tag processing works focus on content-level tags, while little focus is put on context-level tags. This can be attributed to the fact that the gap between multimedia content and context-level is larger, and this brings difficulty to the processing of context-level tags. One possible approach to dealing with context-level tags is to keep them unchanged, such as the method in Liu et al. [2010].

5.1. Tag Information Supplementation

In the manual tagging process, generally human labelers will only assign appropriate tags to images without any additional information, such as the regions depicted by the corresponding tags. But by employing computer vision and machine learning technology, certain information of the tags, such as the descriptive regions and saliency, can be automatically obtained. Here, we refer to these as tag information supplementation. Most existing works employ such a process. First, tags are localized into regions of images or subclips of videos. Second, the characteristics of the regions or subclips are analyzed, and information is supplemented to the tags accordingly.

Liu et al. [2009] propose a method for locating image tags to corresponding regions. They first perform over-segmentation to decompose each image into patches and then discover the relationship between patches and tags via sparse coding. The over-segmented regions are then merged to accomplish the tag-to-region process. Liu et al. further extend the approach based on image search [Liu et al. 2010]. For a tag of the target image, they collect a set of images by using the tag as query with an image search engine. They then learn the relationship between the tag and the patches in this image set. The selected patches are used to reconstruct each candidate region, and the candidate regions are ranked based on the reconstruction error. Liu et al. [2010] accomplish the tag-to-region task by regarding an image as a bag of regions and then performing tag propagation on a graph, in which vertices are images and edges are constructed based on the visual link of regions. Ulges et al. [2008] propose an approach to localizing video-level tags to keyframes. Li et al. [2011] employ a multi-instance learning approach to accomplish the video tag localization, in which video and shot are regarded as bag and shot, respectively. Illustrative examples of tag localization for image data and video data can be found in Figure 1 of Liu et al. [2009] and Figure 1 of Li et al. [2011], respectively.

Feng et al. [2010] propose a tag saliency learning scheme which is able to rank tags according to their saliency levels to an image's content. They first locate tags to images' regions with a multi-instance learning approach and then analyze the saliency values of these regions. It can provide more comprehensive information when an image is relevant to multiple tags, such as those describing different objects in the image. Yang et al. [2010] propose a method for associating a tag with a set of properties, including location, color, texture, shape, size, and dominance. They employ a multi-instance learning method to establish the region that each tag is corresponding to, and the region is then analyzed to establish the properties. Sun and Bhowmick [2010] define a tag's visual representativeness based on a large image set and the subset that is associated with the tag. They employ two distance metrics, *cohension* and *separation*, to estimate the visual representativeness measure. Table III summarizes these works and the information that is supplemented to tags.

By supplementing information to tags, a lot of applications can be facilitated. First, more accurate search based on tags and many other applications can be facilitated. Yang et al. [2010] index images using the property tags that are learned to describe user-provided tags, and it is shown that much better search results can be obtained for the queries that contain some specific properties. Ulges et al. [2008] demonstrate

Table III. Summarization of the Works on Tag Information Supplementation

Work	Supplemented Information
Liu et al. [2009]	Region information
Liu et al. [2010]	Region information
Liu et al. [2010]	Region information
Ulges et al. [2008]	Time information
Li et al. [2011]	Time information
Feng et al. [2010]	Saliency information
Yang et al. [2010]	Location, color, texture, shape, size and dominance
Sun and Bhowmick [2010]	Visual representativeness

that by supplementing video tags with time information, they can collect much better training data for learning video concept detection models. Li et al. [2011] show that the localization of video tags can enable coarse-to-fine video search, as well as intelligent video browsing.

5.2. Tag Refinement

As previously mentioned, user-provided tags are often noisy and incomplete. Tag refinement algorithms are proposed aiming at obtaining more accurate tags for multimedia description. An illustrative example can be found in Figure 2 of Liu et al. [2010]. Here, we first briefly introduce the existing works and then cast most of them into a unified scheme by analyzing their underlying assumptions.

Chen et al. [2010] propose a tag refinement method. For each tag, they first train an SVM classifier with the loosely labeled positive and negative samples. The classifiers are used to estimate the initial relevance scores of tags. They further refine the scores with a graph-based method that simultaneously considers the similarity of photos and semantic correlation of tags. Liu et al. [2008] propose a tag ranking approach which is able to rank the tags that are associated with an image according to their relevance levels. It is motivated from the fact that the order of tags assigned by users usually carries little information about their relevance. It is shown in their study that only less than 10% of the images have their most relevant tags at the first position. Their method consists of two steps. First, initial relevance scores of tags are estimated using Kernel Density Estimation, and then the scores are propagated with a random walk process based on the semantic correlation of tags. Wang et al. [2010] propose a learning-based tag ranking approach. They learn a ranking projection from visual word distribution to the relevant tags distribution with a training set, and the tags for a new image can thus be ranked with the model. Li et al. [2009] introduce an approach that learns the relevance scores of tags by a neighborhood voting method. Given an image and one of its associated tags, the relevance score is learned by accumulating the votes from the visual neighbors of the image. They then further extend the work to multiple visual spaces [Li et al. 2010]. They learn the relevance scores of tags and rank them by neighborhood voting in different feature spaces, and the results are aggregated with a score fusion or rank fusion method. Different aggregation methods have been investigated, such as the average score fusion, Borda count, and RankBoost. The results show that a simple average fusion of scores is already able to perform close to supervised fusion methods, like RankBoost. Fan et al. [2010] group images with a target tag into clusters. They then regard a cluster as a unit, and the initial relevance scores of the clusters are first estimated, and then the scores are refined via a random walk process. Liu et al. [2010] adopt a three-step approach. The first step filters out tags that are intrinsically content-unrelated based on the ontology in WordNet. Second, it refines the tags based on the consistency of visual similarity and semantic similarity of images. More specifically, it is assumed that the visual similarity of images should be consistent with their semantic similarities that are estimated from their tags. Based on this assumption, a regularization scheme is formulated to accomplish the tag refinement. Finally, a tag enrichment step is performed, which expands the tags with their appropriate synonms and hypernyms. Xu et al. [2009] propose a tag refinement algorithm from a topic modeling point of view. A new graphical model named regularized latent Dirichlet allocation (rLDA) is presented to jointly model the tag similarity and tag relevance. Zhu et al. [2010] propose a matrix decomposition method. They use a matrix to represent the image-tag relationship: the (i, j)th elment is 1 if the ith image is associated with the jth tag. Then the matrix is decomposed into a refined matrix **A** plus an error matrix **E**. An optimization scheme is formulated based on four assumptions: (1) user-provided tags are reasonably good; (2) the tags should be highly correlated; (3) the tags of visually similar images are closely related; and (4) the tags of semantically close images should appear with high correlation. Thereby, they enforce the matrix **E** to be sparse and the matrix **A** to follow three principles: (1) let the matrix to be low-rank; (2) if two images are visually similar, the corresponding rows are with high correlation; and (3) if two tags are semantically close, the corresponding vectors are with high correlation.

By carefully observing these preceding works, we can see that most of the tag refinement methods are based on three assumptions.

(1) The refined tags should not change too much from those provided by users. This assumption is usually used to regularize the tag refinement.
(2) The tags of visually similar images should be closely related. This is a natural assumption (most automatic tagging methods are also built upon it).
(3) Semantically close or correlative tags should appear with high correlation. For example, when a tag “sea” exists for an image, the tags “beach” and “water” should be assigned higher confidence, and the tag “street” will tend to be irrelevant.

Now we provide a formulation to mathematically unify the aforementioned tag refinement methods. It can be helpful for us in analyzing how these methods are built based on the three assumptions, and it will also help to design a new algorithm according to the unified scheme.

We first introduce some notations. Denote by $\mathbf{T}_0$ the prior image-tag matrix that satisfies $T_{0,ij} = 1$ if the ith image is associated with the jth tag. Let $\mathbf{T}$ denote the image-tag matrix to be learned, where T_{ij} indicates the confidence score of assigning the jth tag to the ith image. We have two similarity matrices, $\mathbf{W}_I$ and $\mathbf{W}_T$, which record the pairwise similarities of images and tags, respectively. That means $W_{I,ij}$ denotes the similarity of the ith and jth images and $W_{T,ij}$ denotes the semantic similarity between the ith and jth tags. Denote by $T_{i.}$ and $T_{.i}$ the ith row and ith column of the matrix $\mathbf{T}$, respectively. Let $\|.\|$, $\|.\|_1$, and $\|.\|_*$ denote the 2-norm, 1-norm, and nuclear norm of a matrix, respectively.

Based on the aforementioned three assumptions, a unified optimization scheme can be written as

$$\min \Omega(\mathbf{T}) = R(\mathbf{T}, \mathbf{T}_0) + \mu f(\mathbf{T}, \mathbf{W}_I, \mathbf{W}_T). \tag{1}$$

In the preceding equation, the minimization of the term $R(\mathbf{T}, \mathbf{T}_0)$ will force the matrix $\mathbf{T}_0$ to be close to $\mathbf{T}$. The second and third assumptions are reflected in the term $f(\mathbf{T}, \mathbf{W}_I, \mathbf{W}_T)$. Many tag refinement methods can be cast into the formulation and only differ in the definition of these two terms. Table IV illustrates how these two terms are defined in different methods. Actually, we can easily derive new algorithms based on the unified scheme. For example, for the term $R(\mathbf{T}, \mathbf{T}_0)$, we can employ the LogDet

Table IV. Most Tag Refinement Methods Can Be Cast into a Unified Scheme of Eq. (1)

Method	$R(\mathbf{T}, \mathbf{T}_0)$	$f(\mathbf{T}, \mathbf{W}_I, \mathbf{W}_T)$
Chen et al. [2010][10]	$\\|\mathbf{T} - \mathbf{T}_0\\|^2$	$\sum_{i,j} W_{I,ij} \\|T_{i.} - T_{j.}\\|^2$ $+ \sum_{i,j} W_{T,ij} \\|T_{.i} - T_{.j}\\|^2$
Liu et al. [2008][11]	$\\|\mathbf{T} - \mathbf{T}_0\\|^2$	$\sum_{i,j} W_{T,ij} \\|T_{.i} - T_{.j}\\|^2$
Liu et al. [2010][12]	$\sum_{i=1}^{n} \sum_{j=1}^{m} (T_{ij} - \alpha T_{0,ij}) \exp(T_{0,ij})$	$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(W_{ij} - \sum_{k=1}^{m} \sum_{l=1}^{m} T_{ik} W_{T,kl} T_{jl} \right)$
Fan et al. [2010][13]	$\\|\mathbf{T} - \mathbf{T}_0\\|^2$	$\sum_{i,j} W_{I,ij} \\|T_{i.} - T_{j.}\\|^2$
Zhu et al. [2010][10]	$\\|\mathbf{T} - \mathbf{T}_0\\|_1 + \\|\mathbf{T}\\|_*$	$\sum_{i,j} W_{ij} \\|T_{I,i.} - T_{j.}\\|^2$ $+ \sum_{i,j} W_{T,ij} \\|T_{.i} - T_{.j}\\|^2$

Note: They only differ in the two terms $R(\mathbf{T}, \mathbf{T}_0)$ and $f(\mathbf{T}, \mathbf{W}_I, \mathbf{W}_T)$. This table shows the two terms defined in different methods.

divergence between $\mathbf{T}$ and $\mathbf{T}_0$, and for the term $f(\mathbf{T}, \mathbf{W}_I, \mathbf{W}_T)$, we can also use new formulations following the second and third assumptions.

Indeed, some works on the refinement of automatic tagging results [Wang et al. 2007, 2006] could also be cast into the preceding framework, and the main difference is that the matrix $\mathbf{T}$ needs to be replaced by the results of automatic tagging, that is, T_{ij} is the confidence score of assigning the jth tag to the ith image. Another difference is that the second assumption, that is, the tags of visually similar images should be close, is rarely used in these methods, because this assumption should have already been used in the automatic tagging process.

6. BENCHMARK AND DATABASES

As previously mentioned, for automatic tagging, there are many de facto benchmarks available, such as PASCAL VOC, TRECVID, and ImageCLEF. There are also many publicly available datasets. For image datasets, several widely known ones include Caltech-101 [Fei-Fei et al. 2006], Caltech-256 [Griffin et al. 2007], Labelme [Russell et al. 2007], PASCAL VOC [Everingham et al. 2010], 80million Tiny images [Torralba et al. 2008], ImageNet [Deng et al. 2009], etc. Video datasets are relatively fewer. TRECVID datasets are the most widely used. The Kodak dataset [Loui et al. 2007] is an important dataset for research on consumer videos, and MCG-WEBV [Cao et al. 2009] is a dataset for Web video analysis.

Although extensive research efforts have been dedicated to developing new algorithms, there is no well-acknowledged benchmark for assistive tagging. One main reason is that it is difficult to evaluate the performance for ontology-free tagging. For example, given an image or a video clip, you can hardly establish a comprehensive tag list as ground truth, as it depends on the lexicon considered. Alternatively, most methods actually employ a large lexicon that is constructed based on external ontology, such as WordNet [Liu et al. 2008, 2010; Yang et al. 2009] or the folksonomy of collaborative tagging [Wu et al. 2009; Sigurbjörnsson and Zwol 2008].

The previously mentioned datasets for automatic multimedia tagging can be employed in the research of tagging with data selection & organization. For example, in active learning-based multimedia tagging, experiments are usually conducted on these datasets to compare the developed active learning algorithms with automatic methods. For tag recommendation and tag processing, currently there are two publicly

[10] The second step of the algorithm.

[11] The second step of the algorithm.

[12] The second step of the algorithm.

[13] The second step of the algorithm, and here, image cluster is used as the unit instead of the image.

available datasets that have been used widely. Here, we provide a brief description of these two datasets.

(1) MIRFLICKR dataset [Huiskes and Lew 2008; Huiskes et al. 2010]. The MIRFLICKR dataset, first released in 2009, contains 25,000 photos collected from Flickr. By filtering out the associated tags that have appearance counts lower than 20, 1,386 unique tags are maintained in the dataset. The relevance levels of a set of topics, from general to specific, are manually labeled on the dataset. Five types of image features are provided, including HMMD color histogram, spatial color mode, MPEG-7 edge histogram, MPEG-7 homogeneous texture, and binary tag features. In addition, the EXIF metadata of the photos are also provided. Recently, the dataset has been extended to 1 million photos [Huiskes et al. 2010].

(2) NUS-WIDE dataset [Chua et al. 2009]. The NUS-WIDE dataset, released in 2009, contains 269,648 images and 5,018 unique tags collected from Flickr. Originally, the images are associated with about 425,000 tags. Chua et al. [2009] have employed a filtering process to remove the tags that are out of WordNet or have too low appearance frequencies. There are 81 tags that are of different types selected, including object, scene, people, graphic, event, and program, and their relevance levels are checked by human labelers with a semi-automatic process. Six types of image features are provided, including color histogram, color correlogram, edge direction histogram, wavelet texture, block-wise color moment features, and bag-of-visual words histogram.

To study the scalability of the existing assistive tagging works, we illustrate in Figure 9 the dataset sizes in their experiments, including the number of images/video clips and the number of tags.[14] From the figure, we can see an encouraging trend that some works, begin to investigate fairly large datasets. But in most works, the datasets are still not large enough. For example, for most works, the multimedia units are less than 100,000, and tags are less than 10,000. Especially for tag refinement, which usually involves complex matrix optimization, the scalability can be a problem.

7. CONCLUSIONS AND FUTURE DIRECTIONS

We have presented a survey of the state-of-the-art techniques for assistive multimedia tagging, which aims to accomplish multimedia tagging by combining the human's intelligence and, computer's computation power. We categorize the existing efforts into three paradigms, namely, multimedia tagging with data selection & organization, tag recommendation, and tag processing. We start from introducing the industrial solutions in manual tagging and the progress of automatic tagging in research community, and then we summarize the existing works on each assistive tagging paradigm. We also analyze the principles of the existing algorithms.

Although encouraging progress has been made, there are still many challenges that need further study in the future.

We first consider the investigation of humans. How to involve a large number of human labelers is a problem. When dealing with heavy tagging tasks, one choice is to involve more human labelers. For example, Amazon Mechanical Turk [Sorokin and Forsyth 2008] has provided tagging as a cloud service, and it has already been widely used in tagging multimedia data [Sorokin and Forsyth 2008; Deng et al. 2009; Settles et al. 2008]. Users can provide data with specific requirements, and then Amazon will distribute the tagging task to a large number of labelers to accomplish the task. These labelers are usually Internet grassroots and nonprofessionals in data labeling. Thus,

[14] Here we only consider those works that have clearly reported the sizes of datasets and the number of tags.

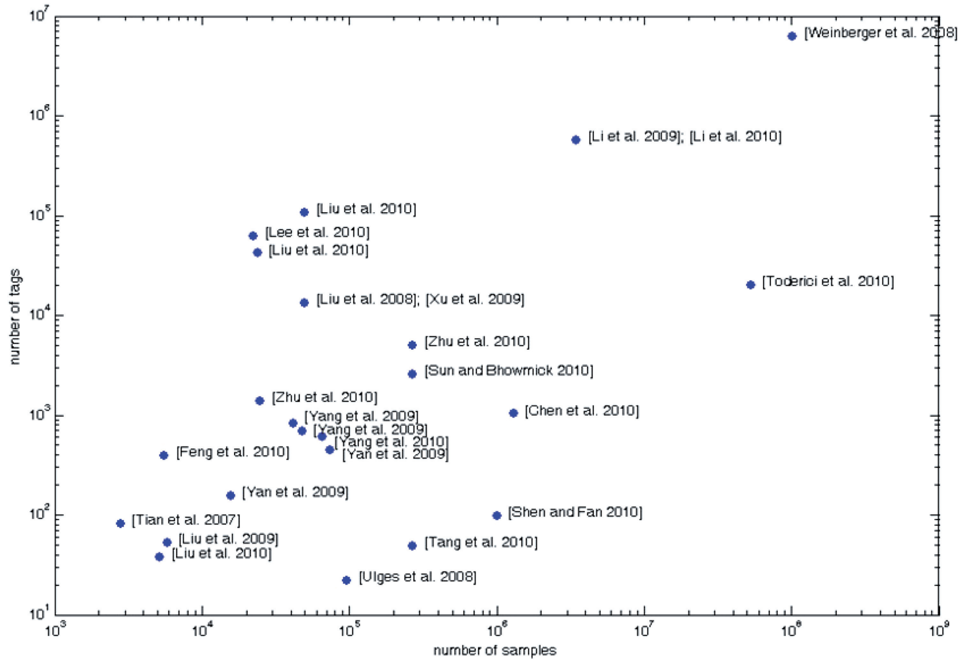


Fig. 9. A summarization of the scales of experiments in several introduced works.

how to reasonably leverage a nonprofessional labeling team becomes challenging, and it also involves many research problems, such as the distribution of tagging tasks and quality control. Besides performing tag information supplementation and tag refinement to enrich and refine the tags provided by nonprofessional labelers, we can design task distribution and tagging quality control components, and many strategies can be investigated. For example, we can insert some data with ground truths to judge the tagging quality of each labeler and avoid tagging with robots or using random algorithms. Then, we can intelligently divide and assign tagging tasks to labelers according to their tagging quality and other information.

For the investigation of computers, the tagging of large-scale data, is still a problem. Some tag recommendation algorithms are not difficult to be generalized on large-scale data, as they can be implemented with just the tag correlation measures, but it is a problem for tagging with data selection & organization, tag information supplementation, and tag refinement. One way is to adopt distributed computing platforms to perform large-scale computation and storage, such as the MapReduce framework. Another approach is to trade performance for speed, such as simplifying some matrix optimization algorithms. GPU can also be a choice for speeding up algorithms with intensive computation. There are some recent works on employing MapReduce and GPU in automatic multimedia content analysis [Yan et al. 2009; van de Sande et al. 2011]. They can be extended to deal with the large-scale assistive tagging tasks, but employing which algorithms and how to optimize them are still worth studying.

In addition, there are still several challenges for each of the assistive tagging paradigms. For tagging with data selection & organization, there are already many different strategies proposed, as introduced in Section 3. How to select and combine these strategies according to the specific task and data distribution is a problem. In tag recommendation, tracking and exploring users' behavior can be an interesting problem. Most of the existing tag recommendation methods only leverage the information of which tags have been selected by users, but we can also explore the unselected tags to improve the performance of tag recommendation. For tag information supplementation and tag refinement, content-level tags are easy to process, but they can hardly achieve good performance for context-level tags, since their gap with multimedia content is larger. Automatically refining or adding information for context-level tags remains a challenging task.

The integration of multiple assistive tagging techniques is also interesting. The introduced paradigms—tagging with data selection & organization, tag recommendation, and tag processing—can actually be put to different stages to complement each other. For example, we can design such an approach. It selects a set of informative samples for manual tagging, and in this process, tag recommendation is employed. The tags of the remaining data are automatically inferred, and after that, tag refinement and information supplementation are performed to further boost the tagging quality.

Finally, a benchmark with well-established ground truth or a public challenge will be helpful. The benchmark plays an important role in growing research directions, especially when it enters a stage in which many different approaches become available. It will be more interesting to set multiple tasks, one for each specific assistive tagging technique, like PASCAL VOC, ImageCLEF, and TRECVID.

REFERENCES

ABBASI, R., GRZEGORZEK, M., AND STAAB, S. 2009. Tagez: Flickr tag recommendation. In *Proceeding of the International Conference on Semantics and Digital Media Technologies*.

AMIR, A., ARGILLANDER, J., CAMPBELL, M., HAUBOLD, A., IYENGAR, G., EBADOLLAHI, S., KANG, F., NAPHADE, M. R., NATSEV, A., SMITH, J. R., TESIC, J., AND VOLKMER, T. 2005. IBM research TRECVID-2005 video retrieval system. In *Proceedings of the TRECVID Workshop*.

ANDERSON, A., RANGHUNATHAN, K., AND VOGEL, A. 2008. Tagez: Flickr tag recommendation. In *Proceeding of the National Conference on Artificial Intelligence (AAAI)*.

AYACHE, S. AND QUÉNOT, G. 2007. Evaluation of active learning strategies for video indexing. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*.

BALLAN, L., BERTINI, M., AND BIMBO, A. D. 2010. Tag suggestion and localization in user-generated videos based on social knowledge. In *Proceedings of the ACM Workshop on Social Media*.

BEITZEL, S. M., JENSEN, E. C., FRIEDER, O., CHOWDHURY, A., AND PASS, G. 2005. Surrogate scoring for improved metasearch precision. In *Proceedings of the International ACM SIGIR Conference on Research & Development on Information Retrieval*.

BLEI, D. AND JORDAN, M. 2003. Modeling annotated data. In *Proceedings of the International ACM SIGIR Conference on Research & Development on Information Retrieval*.

CAO, J., ZHANG, Y., SONG, Y., CHEN, Z., ZHANG, X., AND LI, J. 2009. Mcg-webv: A benchmark dataset for Web video analysis. Tech. rep. ICT-MCG-09-001, Institute of Computing Technology.

CHEN, H. C., CHANG, M. C., CHANG, P. C., TIEN, M. C., HSU, W. H., AND WU, J. C. 2008. Sheepdog c group and tag recommendation for flickr photos by automatic search-based learning. In *Proceeding of the ACM Multimedia Conference*.

CHEN, L., XU, D., TSANG, I. W., AND LUO, J. 2010. Tag-based Web photo retrieval improved by batch mode re-tagging. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

CHEN, Z., CAO, J., SONG, Y., GUO, J., ZHANG, Y., AND LI, J. 2010. Context-oriented Web video tag recommendation. In *Proceeding of the International World Wide Web Conference*.

CHUA, T. S., TANG, J., HONG, R., LI, H., LUO, Z., AND ZHENG, Y. T. 2009. Nus-wide: A real-world Web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

CUI, J., WEN, F., XIAO, R., TIAN, Y., AND TANG, X. 2007. Easyalbum: An interactive photo annotation system based on face clustering and re-ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Duygulu, P., Barnard, K., de Freitas, J., and Forsyth, D. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*.

Everingham, M. 2010. Overview and results of the classification challenge. In *Proceedings of the PASCAL Visual Object Classes Challenge Workshop*.

Everingham, M., Gool, L. V., Williams, C. K. I., Win, J., and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision 88,* 2.

Fan, J., Shen, Y., Zhou, N., and Gao, Y. 2010. Harvesting large-scale weakly-tagged image databases from the Web. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Fei-Fei, L., Fergus, R., and Perona, P. 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell. 28,* 4.

Feng, S., Lang, C., and Xu, D. 2010. Beyond tag relevance: Integrating visual attention model and multi-instance learning for tag saliency ranking. In *Proceedings of the International Conference on Image and Video Retrieval*.

Feng, S. L., Manmatha, R., and Lavrenko, V. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. 1998. An efficient boosting algorithm for combining preferences. In *Proceeding of the International Conference on Machine Learning*.

Gao, Y. and Fan, J. 2006. Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*.

Garg, N. and Weber, I. 2008. Personalized, interactive tag recommendation for Flickr. In *Proceeding of the ACM International Conference on Recommender Systems*.

Goh, K. S., Chang, E. Y., and Lai, W. C. 2004. Multimodal concept-dependent active learning for image retrieval. In *Proceedings of the ACM Multimedia Conference*.

Griffin, G., Holub, A., and Perona, P. 2007. Caltech-256 object category dataset. Tech. rep. 7694, California Institute of Technology.

Hauptmann, A., Lin, W. H., Yan, R., Yang, J., and Chen, M. 2006. Extreme video retrieval: Joint maximization of human and computer performance. In *Proceedings of the ACM Multimedia Conference*.

Hauptmann, A., Yan, R., and Lin, W. C. 2007. How many high-level concepts will fill the semantic gap in video retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

Ho, C., Chang, T., Lee, J., Hsu, J. Y., and Chen, K. 2009. Kisskissban: A competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*.

Huang, T. S., Dagli, C. K., Rajaram, S., Chang, E. Y., Mandel, M. I., Poliner, G. E., and Ellis, D. P. W. 2008. Active learning for interactive multimedia retrieval. *Proc. IEEE 96,* 4.

Huiskes, M. J. and Lew, M. S. 2008. The mir Flickr retrieval evaluation. In *Proceeding of the ACM International Conference on Multimedia Information Retrieval*.

Huiskes, M. J., Thomee, B., and Lew, M. S. 2010. New trends and ideas in visual concept detection. In *Proceeding of the ACM International Conference on Multimedia Information Retrieval*.

Jeon, J., Lavrenko, V., and Manmatha, R. 2003. Automatic image annotation and retrieval using cross media relevance models. In *Proceedings of the International ACM SIGIR Conference on Research & Development on Information Retrieval*.

Jesus, R., Goncalves, D., Abrantes, A., and Corriea, N. 2008. Playing games as a way to improve automatic image annotation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops*.

Jiang, Y. C., Yang, J., Ngo, C. C., Hauptmann, A. G., and Sukthankar, R. 2010. Representations of keypoint-based semantic concept detection: A comprehensive study,. *IEEE Trans. Multimedia 12,* 1.

Kennedy, L. S., Chang, S. F., and Kozintsev, I. V. 2006. To search or to label? predicting the performance of search-based automatic image classifiers. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*.

Lavrenko, V., Manmatha, R., and Jeon, J. 2004. A model for learning the semantics of pictures. In *Proceedings of the Advances in Neural Information Processing Systems*.

Lee, S., Neve, W. D., Plataniotis, K. N., and Ro, Y. M. 2010. MAP-based image tag recommendation using a visual folksonomy. *Pattern Recog. Lett. 31*.

Lew, M., Sebe, N., Djeraba, C., and Jain, R. 2006. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 2,* 1.

Li, G., Wang, M., Zheng, Y. T., Zha, Z. J., Li, H., and Chua, T. S. 2011. Shottagger: Tag location for internet videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*.

Li, J. and Wang, J. 2008. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell. 30,* 6.

Li, X., Snoek, C. G., and Worring, M. 2009. Learning social tag relevance by neighbor voting. *Pattern Recog. Lett. 11,* 7.

Li, X., Snoek, C. G., and Worring, M. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*.

Lin, C., Tseng, B., and Smith, J. R. 2003. VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning. In *Proceedings of the International Conference on Multimedia & Expo*.

Liu, D., Hua, X. C., Wang, M., and Zhang, H. C. 2010. Image retagging. In *Proceedings of the ACM Multimedia Conference*.

Liu, D., Hua, X. S., Yang, L., Wang, M., and Zhang, H. J. 2008. Tag ranking. In *Proceedings of the International World Wide Web Conference*.

Liu, D., Hua, X. S., and Zhang, H. J. 2011. Content-based tag processing for internet social images. *Multimedia Tools Appl*.

Liu, D., Wang, M., Hua, X. C., and Zhang, H. C. 2010. Semi-automatic tagging of photo albums via exemplar selection and tag inference. *IEEE Trans. Multimedia*.

Liu, D., Yan, S., Rui, Y., and Zhang, H. J. 2010. Unified tag analysis with multi-edge graph. In *Proceedings of the ACM Multimedia Conference*.

Liu, X., Cheng, B., Yan, S., Tang, J., Chua, T. C., and Jin, H. 2009. Label to region by bi-layer sparsify priors. In *Proceedings of the ACM Multimedia Conference*.

Liu, X., Yan, S., Luo, J., Tang, J., Huang, Z., and Jin, H. 2010. Nonparametric label-to-region by search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Loui, A. C., Luo, J., Chang, S.-F., Ellis, D., Jiang, W., Kennedy, L., Lee, K., and Yanagawa, A. 2007. Kodak's consumer video benchmark data set: Concept definition and annotation. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*.

Makadia, A., Pavlovic, V., and Kumar, S. 2008. A new baseline for image annotation. In *Proceedings of the European Conference on Computer Vision*.

Monay, F. and Gatica-Perez, D. 2003. On image auto-annotation with latent space models. In *Proceedings of the ACM Multimedia Conference*.

Mori, Y., Takahashi, H., and Oka, R. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*.

Naaman, M. and Nair, R. 2008. Zonetag's collaborative tag suggestions: What is this person doing in my phone? *IEEE Multimedia Mag. 15,* 3.

Naphade, M. R. and Smith, J. R. 2004. On the detection of semantic concepts at TRECVID. In *Proceedings of the ACM Multimedia Conference*.

O'Reilly, T. 2007. What is Web 2.0: Design patterns and business models for the next generation of software. *Commun. Strategies 1*, 17–37.

Over, P., Awad, G., Fiscus, J., and Michel, M. 2010. TRECVID 2010 c goals, tasks, data, evaluation, mechanisms and metrics. In *Proceedings of the TRECVID Workshop*.

Qi, G., Hua, X. S., Rui, Y., Tang, J., Mei, T., and Zhang, H. J. 2007. Correlative multi-label video annotation. In *Proceedings of the ACM Multimedia Conference*.

Rae, A., Sigurbjörnsson, B., and van Zwol, R. 2010. Improving tag recommendation using social networks. In *Proceedings of the International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*.

Rui, Y., Huang, T. S., and Chang, S. F. 1999. Image retrieval: Current techniques, promising directions, and open issues. *J. Visual Commun. Image Rep. 10,* 1.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. T. 2007. Labelme: A database and Web-based tool for image annotation. *Int. J. Comput. Vision*.

Sande, K. E., Gevers, T., and Snoek, C. G. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 32,* 9, 1582–1596.

Seneviratne, L. and Izquierdo, E. 2006. An interactive framework for image annotation through gaming. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*.

Settles, B., Craven, M., and Friedland, L. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*.

SEVIL, S. G., KUCUKTUNC, O., DUYGULU, P., AND CAN, F. 2010. Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools Appl. 49,* 1.

SHEN, Y. AND FAN, J. 2010. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *Proceedings of the ACM Multimedia Conference*.

SIGURBJÖRNSSON, B. AND ZWOL, R. V. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the International World Wide Web Conference*.

SMEULDERS, A. W., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell. 22,* 12.

SNOEK, C. G. AND SMEULDERS, A. W. 2010. Visual-concept search solved? *IEEE Comput. 43,* 6, 76–78.

SNOEK, C. G. AND WORRING, M. 2009. Concept-based video retrieval. *Found. Trends Inf. Retriev. 4,* 2.

SOROKIN, A. AND FORSYTH, D. 2008. Utility data annotation via Amazon mechanical turk. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops*.

STEGGINK, J. AND SNOEK, C. G. 2011. Adding semantics to image region annotations with the name-it-game. *Multimedia Syst*.

SUH, B. AND BEDERSON, B. B. 2004. Semi-automatic image annotation using event and torso identification. Tech. rep. HCIL-2004-15, Computer Science Department, University of Maryland.

SUN, A. AND BHOWMICK, S. S. 2010. Quantifying tag representativeness of visual content of social images. In *Proceedings of the ACM Multimedia Conference*.

TANG, J., CHEN, Q., YAN, S., CHUA, T. S., AND JAIN, R. 2010. One person labels one million images. In *Proceedings of the ACM Multimedia Conference*.

TANG, J., HONG, R., YAN, S., CHUA, T. C., QI, G. C., AND JAIN, R. 2011. Image annotation by knn-sparse graph-based label propagation over noisy-tagged Web images. *ACM Trans. Intell. Syst. Technol. 2,* 2.

TIAN, Y., LIU, W., XIAO, R., WEN, F., AND TANG, X. 2007. A face annotation framework with partial clustering and interactive labeling. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

TODERICI, G., ARADHYE, H., PASCA, M., SBAIZ, L., AND YAGNIK, J. 2010. Finding meaning on youtube: Tag recommendation and category discovery. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

TORRALBA, A., FERGUS, R., AND FREEMAN, W. 2008. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 30,* 11, 1958–1970.

ULGES, A., SCHULZE, C., KEYSERS, D., AND BREUEL, T. M. 2008. Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceeding of the International Conference on Image and Video Retrieval*.

VAN DE SANDE, K. E., GEVERS, T., AND SNOEK, C. G. 2011. Empowering visual categorization with the gpu. *IEEE Trans. Multimedia*.

VOLKMER, T., SMITH, J. R., AND NATSEV, A. 2005. A Web-based system for collaborative annotation of large image and video collections. In *Proceedings of the ACM Multimedia Conference*.

VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

VON AHN, L., LIU, R., AND BLUM, M. 2006. Peekaboom: A game for locating objects in images. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

WANG, C., JING, F., ZHANG, L., AND ZHANG, H.-J. 2006. Image annotation refinement using random walk with restarts. In *Proceedings of the ACM Multimedia Conference*.

WANG, C., JING, F., ZHANG, L., AND ZHANG, H.-J. 2007. Content-based image annotation refinement. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

WANG, M. AND HUA, X. C. 2011. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol. 2,* 2.

WANG, M., HUA, X. S., HONG, R., TANG, J., QI, G. J., AND SONG, Y. 2009. Unified video annotation via multi-graph learning. *IEEE Trans. Circuits Syst. Video Technol. 19,* 5.

WANG, M., HUA, X. S., TANG, J., AND HONG, R. 2009. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. Multimedia 11,* 3.

WANG, M., YANG, K., HUA, X. S., AND ZHANG, H. J. 2010. Towards relevant and diverse search of social images. *IEEE Trans. Multimedia 12,* 8.

WANG, X., ZHANG, L., LIU, M., LI, Y., AND MA, W. C. 2010. ARISTA c image search to annotation on billions of Web photos. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Wang, X.-J., Zhang, L., Jing, F., and Ma, W.-Y. 2006. Annosearch: Image auto-annotation by search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Wang, Z., Feng, J., Zhang, C., and Yan, S. 2010. Learning to rank tags. In *Proceedings of the International Conference on Image and Video Retrieval*.

Weinberger, K., M.Slaney, and R.V.Zowl. 2008. Resolving tag ambiguity. In *Proceeding of the ACM Multimedia Conference*.

Wu, F., Han, Y., Tian, Q., and Zhuang, Y. 2010. Multi-label boosting for image annotation by structural grouping sparsity. In *Proceedings of the ACM Multimedia Conference*.

Wu, F., Yuan, Y., and Zhuang, Y. 2010. Heterogeneous feature selection by group lasso with logistic regression. In *Proceedings of the ACM Multimedia Conference*.

Wu, L., Yang, L., Yu, N., and Hua, X. C. 2009. Learning to tag. In *Proceedings of the International World Wide Web Conference*.

Xu, H., Wang, J., Hua, X.-S., and Li, S. 2009. Tag refinement by regularized LDA. In *Proceedings of the ACM Multimedia Conference*.

Yan, R., Fleury, M. O., Merler, M., Natsev, A., and Smith, J. R. 2009. Large-scale multimedia semantic concept modeling using robust subspace bagging and mapreduce. In *Proceedings of the ACM Workshop on Large-Scale Multimedia Retrieval and Mining*.

Yan, R., Natsev, A., and Campbell, M. 2009. Hybrid tagging and browsing approaches for efficient manual image annotation. *IEEE Multimedia Mag. 16,* 2.

Yang, K., Hua, X.-S., Wang, M., and Zhang, H. C. 2010. Tagging tags. In *Proceedings of the ACM Multimedia Conference*.

Yang, K., Wang, M., Hua, X. S., and Zhang, H. J. 2009. Active tagging for image indexing. In *Proceedings of the International Workshop on Internet Multimedia Search and Mining*.

Zhu, G., Yan, S., and Ma, Y. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the ACM Multimedia Conference*.

Received January 2011; revised March 2011; accepted April 2011