



A survey and analysis on automatic image annotation

Qimin Cheng^{a,*}, Qian Zhang^a, Peng Fu^b, Conghuan Tu^a, Sen Li^a

^a School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

^b Department of Earth and Environmental Systems, Indiana State University, Terre Haute, IN 47809, USA

ARTICLE INFO

Article history:

Received 19 April 2017

Revised 31 January 2018

Accepted 11 February 2018

Available online 13 February 2018

Keywords:

Automatic image annotation

Generative model

Nearest-neighbor model

Discriminative model

Tag-completion

Deep learning

ABSTRACT

In recent years, image annotation has attracted extensive attention due to the explosive growth of image data. With the capability of describing images at the semantic level, image annotation has many applications not only in image analysis and understanding but also in some relative disciplines, such as urban management and biomedical engineering. Because of the inherent weaknesses of manual image annotation, Automatic Image Annotation (AIA) has been raised since the late 1990s. In this paper, a deep review of state-of-the-art AIA methods is presented by synthesizing 138 literatures published during the past two decades. We classify AIA methods into five categories: 1) Generative model-based image annotation, 2) Nearest neighbor-based image annotation, 3) Discriminative model-based image annotation, and 4) Tag completion-based image annotation, 5) Deep Learning-based image annotation. Comparisons of the five types of AIA methods are made on the basis of the underlying idea, main contribution, model framework, computational complexity, computation time, and annotation accuracy. We also give an overview of five publicly available image datasets and four standard evaluation metrics commonly used as benchmarks for evaluating AIA methods. Then the performance of some typical or well-behaved models is assessed based on benchmark dataset and standard evaluation metrics. Finally, we share our viewpoints on the open issues and challenges in AIA as well as research trends in the future.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

The big data era is characterized by the huge amount of image data available. Traditional image annotation techniques, labeling image contents at the semantic level manually, are not applicable in the big data era. The main disadvantages of manual image annotation are intuitionistic. First, it is impractical to annotate the mass image data totally through manual ways. Second, the subjectivity of manual annotation will lead to ambiguity over image contents. In other words, different persons may have totally different understandings of the very same image because of differences in the educational background, thinking mode, and even life experience.

Given deficiencies of traditional manual image annotation, research on Automatic Image Annotation (AIA) technology has become a tendency. Inspired by the word co-occurrence model proposed by Mori et al. [1] in 1999, more and more scholars have turned to conduct studies on annotating images by weak-supervision or totally automatic ways. These achievements have boosted the development of AIA to a great extent during the past

two decades. AIA methods are concerned with models/algorithms to label images by their semantic contents or to explore the similarity between image features and semantic contents with high efficiency and low subjectivity. Relevant labels are predicted for untagged images from a label vocabulary through the weak-supervision way or totally automatically. The key of the AIA is to narrow the semantic gap between low-level visual features and high-level semantic labels, i.e., to learn high-level semantic labels from low-level visual features by exploring the image-image correlation, image-label correlation, and label-label correlation. In addition to its applications in image understanding and analysis, such as image retrieval [2–5], scalable mobile image retrieval [6], face recognition [7], facial landmark annotation [8], and photo tourism [9], AIA is also used in urban management, biomedical engineering, social media services and tourism industry, to name a few. As an interdisciplinary discipline, AIA integrates achievements from data mining, semantic analysis, Natural Language Processing (NLP), Automatic Deep Understanding (ADU) of documents, document analysis and recognition, multimedia systems, machine learning, and even biology and statistics.

During the past two decades, considerable efforts have been made to develop various AIA methods [2–4,10–26]. The learning-based annotation techniques/algorithms include the TM model [10], CMRM model [2], CRM model [11], MBRM model

* Corresponding author.

E-mail address: chengqm@hust.edu.cn (Q. Cheng).

[3], Plsa-based model [27], Markov Random Fields(MRF)-based model [4,28], classification model [14,29–32], graph-based semi-supervised learning methods [33–41], ML-LOC [18], and deep learning-based methods [13,25,26,42–49]. The retrieval-based annotation techniques/algorithms include the baseline model [50], UDML model [51], PDML model [52], and 2PKNN model [16]. More recently, some researches perform AIA through automatically filling in the missing tags as well as correcting noisy tags for given images [20,24,53–55].

At present, a general classification and deep review of AIA methods is still lacking. Despite some surveys of AIA methods, the foci were placed on CBIR [56–59], feature extraction and semantic learning/annotation [60–62], statistical approaches [63], image segmentation [64], face recognition [65], Natural Language Processing (NLP) [66] and relevance feedback [67,68]. In this paper, a comprehensively comparative review of AIA methods is presented by synthesizing 138 literatures published during the past two decades. Specifically, this review covers papers published in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), International Journal of Computer Vision (IJCV), Pattern Recognition (PR), Journal of Machine Learning Research (JMLR), ACM Transactions on Graphics (TOG), IEEE Transaction on Multimedia (TMM), and IEEE Transaction on Image Processing (TIP), and papers published in conferences such as AAAI Conference on Artificial Intelligence (AAAI), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and International Conference on Computer Vision (ICCV).

We focus on a more general classification of the AIA methods by five categories: generative model-based AIA methods, nearest neighbor model-based AIA methods, discriminative model-based AIA methods, tag completion-based AIA methods, and deep learning-based AIA methods. Those AIA methods are analyzed and compared based on the underlying idea, main contribution, model framework, computational complexity, and annotation accuracy (in Section 2). Then this paper reviews five publicly available image datasets and four standard evaluation metrics adopted by AIA methods (in Section 3). This paper also assesses some typical or well-behaved models based on the benchmark dataset and the standard evaluation metrics (in Section 4). We also discuss some challenges, open issues, and promising directions in AIA (in Section 5).

2. Annotation methods

There are various classification schemes for AIA techniques, such as probability and non-probability methods, learning-based and retrieval-based methods, supervised, semi-supervised and unsupervised methods. In this paper we classify these methods into five categories: 1) generative model-based AIA methods, which are dedicated to maximizing generative likelihood of image features and labels; 2) nearest neighbor model-based AIA methods, which assume that images with similar features have a great probability to share similar labels; 3) discriminative model-based AIA methods, which view the annotation task as a multi-label classification problem; 4) tag completion-based AIA methods, which can not only predict labels by automatically filling in the missing labels but also can correct noisy tags for given images; 5) deep learning-based AIA methods, which use deep learning algorithms to derive robust visual features or exhaustive side information for AIA, especially for large-scale AIA.

The aforementioned five categories of AIA methods can be further classified into several sub-categories according to their underlying ideas. Fig. 1 provides a taxonomy, as well as some hot topics of AIA methods by covering 138 literatures.

In Fig. 1, generative model-based AIA methods can be mainly divided into three classes including the relevance model, topic

model and hidden Markov model (HMM). As for nearest neighbor model-based AIA methods, three key issues, i.e., distance metric learning (DML), class-imbalance, and weak-labeling, are receiving more attention. With regards to the discriminative model-based AIA, research efforts have been mainly devoted to developing the graph-based semi-supervised learning methods. The advantage of the graph-based methods is that the label correlation can be easily incorporated into the graph in the propagation process. As such, the way to describe the label correlation plays an important role in developing AIA methods. For the tag completion-based AIA methods, they can be further divided into the matrix completion, linear sparse reconstructions, subspace clustering, and low-rank matrix factorization. With respect to deep learning-based AIA methods, great progress has been made in two facets for annotation, i.e., derivation of robust visual features, and exhaustive utilization of side information.

2.1. Generative model-based AIA methods

The generative model-based AIA methods are quite popular, and great achievements have been made in the early 21st century. The generative models are dedicated to maximizing the generative likelihood of image features and labels. For an untagged image, the generative model-based AIA techniques provide the probability of an image label by computing a joint probabilistic model of image features and words from training datasets. The generative models used for AIA mainly consist of the relevance model, topic model, and Markov random field model.

2.1.1. The relevance model

The relevance model-based AIA methods are generally implemented in three steps: define the joint distributions over image features and labels; compute the posterior probability of each label for the unlabeled images (usually the visual feature); to annotate a new image by choosing a label of the highest probability. Various relevance models have been developed for image annotation, including the translation model (TM) [10], across media relevance model (CMRM) [2], continuous space relevance model (CRM) [11], and multiple Bernoulli relevance models (MBRM) [3].

The TM creates a one-to-one match between a blob and a word [10]. In this model, regions are firstly clustered from training images and represented by the index of the closest centroid of the cluster (blob). Next, each blob is associated with a word in the vocabulary, similar to the process of learning a lexicon, by maximizing the joint probability through the EM algorithm, which is computationally expensive and time-consuming.

The CMRM also uses the blob generated from image features to describe an image [2]. It computes the joint distribution between keywords and the entire image rather than specific blobs that are used in TM since the blob vocabulary may give rise to many errors. In the CMRM, an image I is represented by a set of blobs $\{b_1, b_2 \dots b_n\}$, and the conditional probability of image I belonging to a class w is approximated as (1):

$$p(w|I) = p(w|b_1, b_2 \dots b_n) \quad (1)$$

The training set derived from annotated images is used to estimate the joint probability for the word w and the blobs $\{b_1, b_2 \dots b_n\}$. The joint probability distribution can be computed over the image j in the training set T as (2):

$$p(w, b_1, b_2 \dots b_m) = \sum_{j \in T} p(j) p(w, b_1, b_2 \dots b_m | j) \quad (2)$$

Once the image j is known, the prior probability $p(j)$ is constant for the entire training set. By assuming words w and blobs $\{b_1, b_2 \dots b_m\}$ are independent, a word model and a blob model are

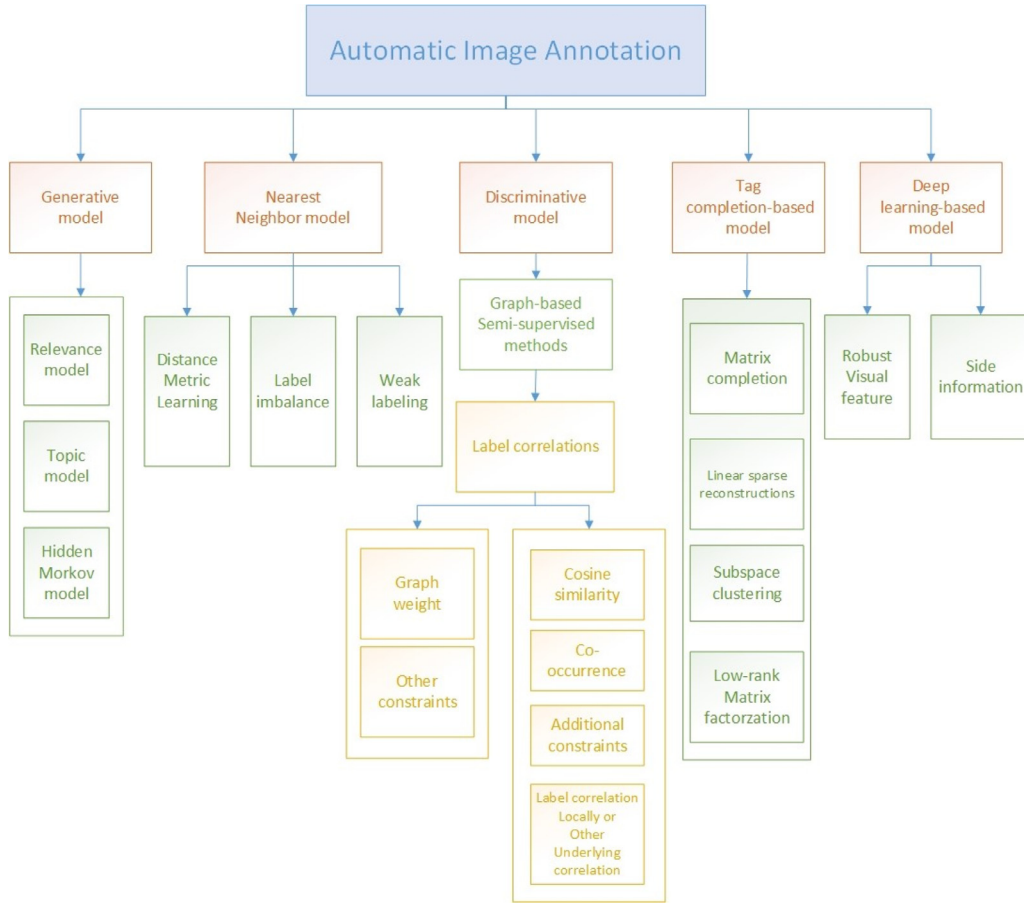


Fig. 1. Taxonomy of automatic image annotation techniques.

built for each individual training image j . Thus the Eq. (2) can be corrected as (3):

$$p(w, b_1, b_2 \dots b_m) = \sum_{j \in T} p(j) p(w|j) \prod_{i=1}^n p(b_i|j) \quad (3)$$

$$p(w|j) = (1 - \alpha_j) \frac{\#(w, j)}{|j|} + \alpha_j \frac{\#(w, T)}{|T|}$$

$$p(b_i|j) = (1 - \beta_j) \frac{\#(b_i, j)}{|j|} + \beta_j \frac{\#(b_i, T)}{|T|}$$

Where $\#(w, j)$ denotes the frequency that the word w occurs in the caption of image j , and $\#(w, T)$ denotes the frequency that occurs in all captions in the training set T . The meaning of $\#(b_i, j)$ and $\#(b_i, T)$ is similar to $\#(w, j)$ and $\#(w, T)$. Here $|j|$ stands for the count of all words and blobs occurring in the image J , and $|T|$ stands for the total size of the training set. Especially, α_j and β_j are adjustable parameters.

In the TM and CMRM, it is required to discretize continuous feature vectors. The CRM uses continuous-valued feature vectors directly to describe image since the quantification of continuous feature vectors into a discrete vocabulary will lose some necessary image information [11]. Furthermore, the CRM uses regions instead of blobs to describe the given image in that the annotation capability of the CMRM model is sensitive to clustering errors. It is assumed that each image contains several distinct regions $\{r_1, r_2 \dots r_n\}$, and each region is an element of R and contains the pixels of some prominent objects in the image. A function ϕ is modeled for mapping image region $r \in R$ to real-word vectors $g \in \mathbb{R}^k$, and the value $\phi(r)$ represents a set of features, or charac-

teristics of an image region. Then the joint distribution of image regions $\{r_1, r_2 \dots r_n\}$ and a set of words $\{w_1, w_2 \dots w_n\}$ is calculated in Eq. (4):

$$p(r_A, w_B) = \sum_{j \in T} p_T(j) \prod_{b=1}^{n_B} p_v(w_b|j) \prod_{a=1}^{n_A} \int_{\mathbb{R}^k} p_R(r_a|g_a) p_\phi(g_a|j) dg_a \quad (4)$$

Where $p_R(r_a|g_a)$ represents a global probability distribution responsible for mapping generator vectors $g \in \mathbb{R}^k$ to image regions $r \in R$. It is assumed that the feature vector $\phi(r)$ of all regions in each image follows the Gaussian distribution. $p_v(w|j)$ is computed by using the multinomial distribution, as the following Eq. (5):

$$p_R(r|g) = \begin{cases} 1/N_g & \text{if } \phi(r) = g, \\ 0 & \text{else.} \end{cases}$$

$$p_\phi(g|j) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2^k \pi^k |\Sigma|}} \exp\{(g - \phi(r_i))^T \Sigma^{-1} (g - \phi(r_i))\} \quad (5)$$

$$p_v(v|j) = \frac{\mu p_v + N_{v,j}}{\mu + \sum_{v'} N_{v',j}}$$

The MBRM [3] uses multiple Bernoulli models instead of the multinomial distribution to show the word probabilities, as illustrated by Eq. (6). The assumption is that a word itself rather than its frequency should be paid more attention. In other words, presence or absence of a word is concerned in image annotation rather

than the frequency of the very word used in the content.

$$p_v(v|j) = \frac{\mu\delta_{v,j} + N_v}{\mu + N} \quad (6)$$

2.1.2. The topic model

The topic model is another type of widely used generative models for AIA. The topic model-based AIA methods consider annotated images as samples from a specific mix of topics, where each topic is a probability distribution over image features and annotation words. The topic model may be superior to the relevance model since the relevance model builds latent space in which the text and visual modalities are equally important, and yields weak relationships between visual feature co-occurrence and semantic content. Typical topic models include the latent semantic analysis (LSA), probabilistic latent semantic analysis (pLSA), and latent Dirichlet allocation (LDA). Some prominent document analysis approaches, such as the pLSA and LDA, have been successfully adapted to handle the joint modeling of visual and content information [15,27,69–71], by utilizing the topic concept.

The pLSA model [27] assumes that a group of co-occurrence words is associated with a latent topic. In general, a topic is an intuitionistic concept and is characterized by a series of related words. For example, if “Microsoft” is regarded as a topic, then “Bill Gates” and “Microsoft Windows” probably appear frequently in this topic. Since the pLSA model assumes the existence of a latent topic z (aspect) in the generative process for each element x_i in a particular document d_i , the joint probability of element x and document d is calculated as:

$$p(x_j, d_i) = p(d_i) \sum_k p(z_k|d_i) p(x_j|z_k) \quad (7)$$

The pLSA model has been improved in many ways. For example, Lienhart and Romberg [71] proposed a multi-layer pLSA model. The multi-layered structure makes it convenient to extend the learning and inference rules to more layers and modalities. The MF-pLSA model [69] can be considered as an extension of pLSA methods in that it handles two kinds of visual feature domains by adding one more child node to the graphical structure of the original pLSA [27].

The tr-mmLDA model [70] presents a topic-regression multimodal Latent Dirichlet Allocation method to learn the joint distribution of texts and image features. The model provides an alternative method to learn two sets of hidden topics and incorporates a linear regression module to capture statistical associations between images and texts. This tr-mmLDA model is quite different from the former topic models which only share a set of latent topics between two data modalities. Furthermore, tr-mmLDA can handle differences in the number of topics in the two data modalities, an advantage over previous models [27] in which the sum of latent topics has to be decided manually.

2.1.3. The hidden Markov model

As one of the representative generative models, Hidden Markov model (HMM) has received much attention for the AIA task [72–74]. In [72], an image is characterized by a sequence of vectors of low-level visual features, such as color, texture, and oriented edges, and is deemed to be generated by a hidden Markov model. Let $I = \{i_1, i_2 \dots i_N\}$ be the set of training images with annotations, where N is the total number of images. For a given image i , let $i = \{r_1, r_2 \dots r_T\}$ illustrate its region set and $W = \{w_1, w_2 \dots w_M\}$ represent its keyword set, where T denotes the number of regions, and M is the number of keywords. The vocabulary v is generated by collecting the keywords of all training images. Thus, the HMM models the image-features and caption-text jointly as the realization

of a generative stochastic process:

$$f(r_1^T, Ww_0) = \sum_{w_1^T \in W} \prod_{t=1}^T f(r_t w_t) p(w_t w_{t-1}) \quad (8)$$

Where r_1^T represents the region sequence $\langle r_1, r_2 \dots r_T \rangle$ of image i and w_1^T illustrates the corresponding keyword sequence $\{w_1, w_2 \dots w_M\}$. The performance of the HMM based annotation scheme is directly affected by the emission density f and the transition probability p , which represent the visual feature distribution associated with each keyword and the keyword correlation, respectively.

In a nutshell, the HMM performs an abstraction of the information presented in the paired image and caption training data. Each concept $w \in v$ is modeled in the HMM by a mixture of Gaussians. Then for each concept, only the sufficient statistics (mean and covariance) of the image features are retained. By contrast, the relevance model retains the entire training corpus for a comparison with the test image. The HMM performs the annotations quite efficiently since each image-feature vector only needs to be compared with the Gaussian-mixture representation of each concept rather than with each training image. As a result, it requires less computation time compared to the relevance model.

The TSVM-HMM model [73] uses the transductive SVM (TSVM) to remarkably boost the reliability of the HMM by largely reducing users labeling efforts. The TSVM-HMM based annotation scheme takes advantages of both the discriminative classification and the generative model for AIA tasks. The SHMM model [74] structures a new spatial-HMM to describe the spatial relationships among objects and investigates the semantic structures of concepts in natural scene images.

The generative models have made remarkable contributions to the development of AIA, and many AIA methods are inspired by generative models. However, there are three main shortcomings in the generative models-based AIA methods. The first one is the generative models estimate the generative likelihood of image features and annotations but cannot guarantee optimization of the tag prediction. The second one is that generative models may not be able to capture the intricate relationship between image features and labels. The third one is the high computing demand caused by the complex algorithms, e.g., the EM algorithm, and the abundant parameters settings.

2.2. Nearest neighbor model-based AIA methods

The nearest neighbor model-based AIA methods assume that visually similar images are more likely to share common labels. For a given query image sample, the nearest neighbor model-based AIA methods usually search for a set of similar images firstly, and then the tags of the query image are derived based on the tags of the similar images.

The Joint Equal Contribution (JEC) model [50] is one of the most classical nearest neighbor models. It creates a family of very simple and intuitive baseline methods for image annotation. The JEC model utilizes global low-level image features and a simple combination of basic distance measures to find nearest neighbors of a given image. Keywords are then assigned using a greedy label transfer mechanism, which selects keywords from the nearest neighbor or neighbors based on co-occurrence and frequency factors.

The nearest neighbor models retrieve a set of top k similar images from candidate datasets through two key components: 1) A feature representation scheme to extract appropriate image features, and 2) A distance measure scheme to compute distance for extracted features. Studies pertinent to the extraction of appropriate visual features are legion, [60] and this paper focuses on some

open issues and challenges in the nearest neighbor models, including Distance Metric Learning (DML), and class-imbalance and weak-labeling.

2.2.1. Distance metric learning

DML plays an important role in many applications including machine learning and data mining. We will discuss DML based on the Mahalanobis distance.

In particular, given a dataset containing n images, image $x_i \in \mathbb{R}^d$ is represented in a d -dimensional space. The Mahalanobis distance between image x_i and image x_j is defined as:

$$d_M(x_i, x_j) = \|x_i - x_j\|_M^2 = (x_i - x_j)^T M (x_i - x_j) \quad (9)$$

Where M is a pre-defined matrix that satisfies the property of a valid metric, and the goal of DML is to learn an optimal Mahalanobis metric M from the training data (side information).

The side information derived from the tags and other rich content of images, referred to as uncertain side information, leads to a new challenge for DML. Conventionally, the DML methods require the learning task for explicit side information given in the form of either class labels for image classification, [75,76], or pairwise constraints for clustering and retrieval [77–79]. Here the pairwise constraints are used to measure whether two images are similar (“must-link”) or dissimilar (“cannot-similar”). Since most images in the image annotation task are labeled by a number of tags, it is difficult to determine if two images form the must-link constraints.

The incorporation of DML into AIA can contribute to searching for nearest neighbors and thus improve the annotation accuracy. Exemplified models include the Unified Distance Metric Learning (UDML) model and probabilistic distance metric learning (PDML) model.

The UDML model [51] utilizes both textual tags and visual content for metric learning and to combine inductive and transductive learning in a systematic framework. Different from DML, the UDML model aims to learn effective metrics from implicit side information. Extracting side information can be achieved in a “triplet” format, i.e., (x, x^+, x^-) , in which image x and image x^+ are similar, while image x and image x^- are dissimilar. The inductive learning formulation for optimizing distance metric from side information is shown in Eq. (10):

$$\min_{M>0} J_1(M) \triangleq \frac{1}{N_p} \sum_{i=1}^{N_Q} \sum_{\forall (x_{qi}, x_{ki}^+, x_{ki}^-) \in P_i} \ell(M; (x_{qi}, x_{ki}^+, x_{ki}^-)) \quad (10)$$

$$\ell(M; (x_{qi}, x_{ki}^+, x_{ki}^-)) = \max\{0, 1 - [d_M(x_{qi}, x_{ki}^-) - d_M(x_{qi}, x_{ki}^+)]\}$$

where N_p denotes the total number of triplets, and the loss function optimizes the metric by penalizing large distance between two similar images and small distance between two dissimilar images. In addition, Wu et al. [51] also developed a transductive approach to integrate textual tags and visual contents of social images as follows:

$$\min_{M>0} J_2(M) \triangleq \sum_{i,j} w_{ij} \|x_i - x_j\|_M^2 \quad (11)$$

where w_{ij} represents the cosine similarity between the two tag vectors of the two images. The equation indicates that if two images shared similar textual tags, a small visual distance between them is expected. Finally, the formulation of a unified distance metric learning is achieved by fusing the inductive formulation and the transductive formulation:

$$\min_{M>0} J(M) \triangleq \frac{1}{2} \text{tr}(M^T M) + cJ_1(M) + \lambda J_2(M) \quad (12)$$

The PDML scheme [52] can derive probabilistic side information from the data by using a graphical model, and then the probabilistic RCA algorithm is used to find an optimal metric from the probabilistic side information. It should be noted that the probabilistic side information derived by the PDML model is in the form of latent chunklets with probabilistic assignments. Specifically, the “chunklets” means that images in the same trunklets are similar to each other, while images in different chunklets can be similar or dissimilar dependent on the similarity of the two associated chunklets.

2.2.2. Class-imbalance and weak-labeling

The class-imbalance problem is quite common when the size of label vocabulary is large. It means there exists a high variance among the number of images corresponding to different labels. In most cases, class-imbalance leads to a poor-labeling since the nearest neighbor-based AIA methods label images with the help of adjacent images. The more frequent a candidate label is used to describe its neighbors, the higher probability this label will be used to the annotate adjacent unlabeled image. In another word, if a tag is only represented by a few instances, there will be a low probability to use the very same tag for the unlabeled image.

Another problem, weak-labeling, is caused by limitations of manual annotation to some degree. On one hand, it means that a significant number of available images are not annotated with all the relevant labels. On the other hand, it indicates that a number of images may be tagged by irrelevant labels. Undoubtedly, weak-labeling will also cause poor-labeling. Below we highlight some models that can overcome the weak-labeling problem in AIA.

The TagProp model [80] integrates a weighted nearest neighbor based method and metric learning capabilities into a discriminative framework. It transfers labels by taking a weighted combination of the label presence and absence of neighbors. Moreover, it introduces word-specific discriminant models, which boost the probability for rare tags and decrease the probabilities for frequent ones concurrently to overcome the class-imbalance problem.

The 2PKNN model [16] represents a classical solution to solve problems related to class-imbalance and weak-labeling. It identifies all related semantic neighbors for each label by selecting k similar images in the vocabulary. Thus, it can be guaranteed that each label appears at least k times in the training set. Suppose, for each w_i label, let $T_i \in \mathcal{T}$ be a subset of training dataset which contains all images labeled by w_i . For each unannotated image j , k images, which are similar to the target image, are selected to build the corresponding set $T_j, i \subseteq T_i$, where $T_j = \{T_{j,1}, T_{j,2} \dots T_{j,m}\}$ contain all semantic neighbors corresponding to the labels of the image j . The 2PKNN model also provides a metric learning framework that learns appropriate weights for combining multiple features in the distance space. In the 2PKNN model, image annotation is regarded as a problem of finding the posterior probability that a label is given to a new image A :

$$p(y_k|A) = \frac{p(A|y_i)p(y_i)}{p(A)} \quad (13)$$

Where $p(y_i)$ is the prior probability of the label y_i . Then, the conditional probability for j given a label $y_k \in Y$ is shown as:

$$p(j|y_k) = \sum \theta_{j, I_i} \cdot p(y_k|I_i) = \sum \exp(-D(j, I_i)) \cdot \delta(y_k \in Y_i) \quad (14)$$

Where $p(y_k|I_i) = \delta(y_k \in Y_i)$ indicates the presence/absence of label y_k in the label set of image I_i , and the distance between image A and B is expressed as:

$$D(A, B) = \sum_{i=1}^n w_i \sum_{j=1}^{N_i} v^i(j) \cdot d_{AB}^i(j) \quad (15)$$

Thus, each image is featured by a multi-dimensional (N^i) vector. The goal of metric learning is to learn weight w over multiple fea-

ture distances as well as the base distance that maximizes the annotation performance. As such, it is different from the methods to learn the optimal distance through forming the Mahalanobis metric M .

Inspired by the 2PKNN model [16], Bakliwal and Jawahar [81] also made each label to appear at least k times in the training data to handle the class-imbalance problem. The study uses a weighted nearest neighbor algorithm to assign importance to the labels based on image similarity and then computes the score for each label for a new image. Furthermore, the annotation accuracy is improved by a variable number of tags for images, which differ from previous studies [16,50,80] that have a fixed number of tags for unlabeled images.

In addition to the efforts made on the open issues such as DML, class-imbalance, and weak-labeling, studies on other aspects such as label set relevance [82], and alterable amount of nearest neighbors [19], are also aiming to improve the performance of nearest neighbor model-based AIA methods. Besides, Non-negative Matrix Factorization (NMF) [83] and convolutional neural network [42] are also adopted to improve the prediction performance in AIA. We will discuss these works as follows.

Tian and Shen [82] developed a model aiming at learning label set relevance. The meaning of the terminology “label set relevance” is twofold: 1) A label set is relevant to an image if the label set describes the content of the image accurately; 2) A label set is correlative if labels in the set are more correlative to each other. Instead of estimating the label relevance for an image by the labeled frequency derived from its nearest neighbor, Tian and Shen [82] derived label set relevance by formulating “label-set”-to-image relevance and relevance of the label to other labels into a joint framework.

Lin et al. [19] used a constrained range rather than an identical and fixed number of visual neighbors for label prediction. It is quite superior over many previous visual-neighbors-based methods which are sensitive to the number of visual neighbors [50]. Moreover, their model can not only identify the reliable tags but also enhance the robustness of annotation by performing a tag-based random search process over the graphical model from range-constrained visual neighbors.

Kalayeh et al. [83] conceived a hybrid model by integrating the nearest-neighbor scheme into the generative model. In this model, tags are treated as another view in addition to visual features. Then a joint factorization of all views into basis and coefficient matrices is performed so that the coefficients of each training image are similar across views and that each basis vector can capture the same latent concept in each view. The Non-negative Matrix Factorization (NMF) does not require any training processes to form the global mapping between features and tags or tag-specific discriminant models but previous studies usually do. Furthermore, in order to decrease the impacts of label-imbalance, the hybrid model introduces two weight matrices into the Multi-view NMF framework to increase the importance of both the rare tags and the images which contain rare tags.

The nearest neighbor model-based AIA methods are concept-clear and structure-intuitive, and many of them have been proved to be quite successful for tag prediction due to their high flexibility. However, improvements are still needed because of some inherent shortages. Firstly, the performance of nearest neighbor model-based AIA methods may be influenced by the size of training datasets. When the number of training examples is limited, the performance of the nearest neighbor-based models may be barely satisfactory because the similarity between training samples and query samples is mainly determined by visual features. Obviously, the larger the training data is, the better the model performance will be [81]. Secondly, the performance of nearest neighbor model-based AIA methods is highly sensitive to retrieval per-

formance. Therefore, an efficient way to identify appropriate neighbors for unlabeled images is highly sought.

2.3. Discriminative model-based AIA methods

Discriminative model-based AIA methods view image annotation as a multi-label classification problem. This problem is solved by learning an independent binary classifier for each label and then to utilize the binary classifiers to predict labels for unlabeled images. Such an idea yields study pertinent to image annotation using SVM [29], Bayes point machine [30] label ranking [31], and supervised multi-class labeling method [5]. The SML model [5] is one of the models to treat AIA as a multi-classification problem and learns class-specific distributions for each label. However, the multi-label classification approaches cannot extend to a large number of categories since a binary classifier has to be built for each category.

The SVM-DMBRM model [14] shows some improvements in classification based on previous studies and presents a hybrid model to take full advantages of the merits of both generative and discriminative models for AIA. While SVM tries to solve the weak-labeling issue, DMBRM strives to solve the class-imbalance issue. Another model HMC [32] exploits the annotation hierarchy by building a single predictive clustering tree (PCT) that predicts all annotations of an image simultaneously. The HMC model differs greatly from the prior frameworks which produce only one binary classifier for each given class. Image annotation is also formulated as a semi-supervised learning problem due to the lack of labeled data compared to the size of an image set in the real world. The correlation among different labels is taken into consideration to improve the classification performance. The semi-supervised learning task, given a data set with pairwise similarities, can be viewed as a label propagation from labeled data to unlabeled data and is quite suitable for AIA.

In recent years, research efforts on semi-supervised learning have been mainly devoted to the graph-based learning methods which model the whole data as a graph. Label correlation can be easily incorporated into the graph in this propagation based method. There are mainly two ways to use label correlations in the graph-based learning method. The label correlation is used as part of graph weights [33–37,84] or as an additional constraint [38,39].

2.3.1. Graph weight

The MLMC model [34] aims at employing label correlations to boost the accuracy of AIA via a graph-based learning method to maximize the label assignment consistency over the whole image similarity. In the MLMC model, nodes correspond to labeled or unlabeled data points, and edges reflect the similarities between data points. Both the pairwise data similarity and label similarity are used to measure the proximity of two data points. On a weighted graph, the weight w is used to represent similarity between nodes:

$$W = W_X + \gamma W_L \quad (16)$$

Where W_X indicates the similarity between pairwise image data and W_L represents the similarity between pairwise labels. And γ is a parameter used to balance the influences of the two similarity matrices. The Gaussian function is used to compute the similarity between two images:

$$W_X(i, j) = \exp\left(-\frac{\|x^i - x^j\|^2}{\sigma^2}\right) \quad (17)$$

Let $Z = \{z_1, z_2 \dots z_n\}$, where z_i is a k -dimensional vector which represents the label assignment indication vector for a data point x_i . If $z_i(k) = 1$, it means that image x_i is annotated with label k . If $z_i(k) = -1$, it means that image x_i is not annotated with label k .

Different from Eq. (17), the label similarity between two images is calculated as follows:

$$W_l(i, j) = \frac{z_i^T c z_j}{\|z_i\| \|z_j\|} \quad (18)$$

Where c is a square matrix to represent label correlations, and r_k and r_l are the k th and l th rows of Z which represent images labeled by a corresponding keyword. The label correlation between two keywords is measured by a cosine similarity:

$$C_{kl} = \cos(r_k, r_l) = \frac{\langle r_k, r_l \rangle}{\|r_k, r_l\|} \quad (19)$$

Wang and Huang et al. developed a new multi-label correlated Greens function approach [33] to propagate image labels over a graph, in which the correlations among labels were integrated into the objective function. A novel Bi-relational graph model (BG), developed also by Wang et al. [35], assumes both data graph and label graph as sub-graphs. These sub-graphs are connected by an additional bipartite graph induced from label assignments. The Random walk algorithm is performed on both class-to-image relevance and class-to-class relevance by considering each class and its labeled images as a semantic group. Especially, class-to-class relevance is described in an asymmetric way to imitate the semantic relationship in the real world, and class-to-image relevance is used to predict labels for unannotated images directly.

Similar to [33,36,37] combine image-based graph and the word-based graph into an integrated framework. Liu et al. [36] used the image-based graph learning method to generate candidate annotations followed by refinement performed by the word-based graph learning method. Furthermore, a nearest spanning chain (NSC) method is proposed to construct the image-based graph, in which the edge-weights of the graph are calculated with chain-wise statistical information rather than the pairwise similarity. In [37], a quadratic energy function is defined for the image-based graph and the word-based graph, respectively, to minimize the combination of the two energy function that balance the two energy terms. Feng proposed an approach to discover the co-occurrence patterns in a network structure where nodes represent semantic concepts and edges represent co-occurrence [84]. In [84], edge weight is determined by three types of co-occurrence measure, i.e., global semantic co-occurrence measures, global visual co-occurrence measure, and local visual co-occurrence measure.

2.3.2. Additional constraints

The second type of graph-based learning methods uses the label correlation as an additional constraint. For example, Zha et al. [38] designed a graph-based learning framework to exploit both the inherent correlations among multiple labels and label consistency over the graph. The vector-valued function estimated on the graph has three constraints: 1) it should be close to the given labels, 2) it should be smooth on the whole graph, and 3) it should be consistent with the label correlations.

Let $\chi = \{x_1, x_2, \dots, x_N\}$ be a set of data points in R^d , the first k data points are labeled as $Y_k = \{y_1, y_2, \dots, y_k\}$, and the task is to label the remaining $N - k$ data points. The edges are weighted by a matrix W in which w_{ij} indicates the similarity between x_i and x_j , and $f = \{f_1, f_2, \dots, f_k, f_{k+1}, \dots, f_N\}$ represent the predicted labels for all images. Specifically, the three constraints mentioned above are formed as Eq. (20), where $E_l(f)$ is a loss function to penalize the deviation from the given label. $E_s(f)$ is a regularization term to prefer smoothness, and $E_c(f)$ is to address the consistency of label correlation.

$$E_l(f) = \infty \sum_{i \in k} (f_i - y_i)^2 = (f - y)^T \Lambda (f - y)$$

$$E_s(f) = \frac{1}{2} \sum w_{ij} (f_i - f_j)^2 = f^T \Delta f$$

$$E_c(f) = -\text{tr}(f c f^T) \quad (20)$$

Where Λ is a diagonal matrix, and $\Delta = D - W$ is the combinatorial graph Laplacian, D is a diagonal matrix with its (i, i) -element equal to the sum of the i -th row of W . C is defined as $C = C^* - D_c$, $c_{ij}^* = \exp(-\|f_i - f_j\|^2 / 2\sigma_c^2)$, where D_c is a diagonal matrix with the (i, i) -element equal to the sum of the i -th row of C^* . Then the proposed framework can be formulated to minimize:

$$E(f) = \text{tr}((f - y)^T \Lambda (f - y)) + \alpha \text{tr}(f^T \Delta f) - \beta \text{tr}(f c f^T) \quad (21)$$

Where α and β are nonnegative constants to balance $E_s(f)$ and $E_c(f)$. By solving the optimization problem, the labels for the unannotated images can be predicted directly.

HDIALR [39] uses a hidden-concept driven image annotation and label ranking algorithm. HDIALR conducts label propagation based on the similarity over a visually and semantically consistent hidden-concepts space. To improve the performance of AIA, the HDIALR model takes into considerations the label locality, inter-label similarity, and intra-label diversity among multiple label images. It is expected in the formulation that: 1) Each label is represented as a linear combination of the hidden concepts so that the inter-label similarity is revealed by the coefficients of the linear combination; 2) Each hidden concept is expressed by its respective subspace, in which different elements correspond to different expressions of the hidden concept, and the intra-label diversity is covered by those elements; 3) Label locality is revealed by the decomposition of image representation into label representations implicitly. As such, the semi-supervised multiple label learning task can be formulated as a regularized nonnegative data factorization problem.

The Markov Random Field (MRF) is a probabilistic undirected graph model. It has been used to explore semantic relationships among concepts and low-level features in AIA. The MRF-based discriminative models have been proved more efficient than the generative model that suffers from the weak learning ability due to the lack of appropriate learning strategy for characterizing the semantic context. In [4,12,28], the MRF is adopted to boost the potential of AIA models.

Feng and Manmatha [12] used the MRF model to perform the direct retrieval task. By maximizing average precision rather than the likelihood, the authors simplified their model while keeping high efficiency because normalizer did not need to be calculated. It should be noted that the model is built using a discrete vocabulary of vector quantized regions. A large vocabulary of a couple of million visterms are generated by the hierarchical k-means method since performance is dependent on the size of the vocabulary.

Xiang et al. [28] and Liorente and Manmatha [4] presented approaches based on the MRF to utilize the semantic dependencies of the images. Xiang et al. [28] adopted the MRF to model the context relationships among semantic concepts with keyword sub-graphs generated from training samples for each keyword. In [28], a site potential function and an edge potential function are defined to model the joint probability of an image feature and a word. Liorente and Manmatha [4] built an undirected graph where a node could be an image of the test datasets or a query. The study concentrated on exploring the dependencies between image features and words, the dependencies between two words, and the dependencies among image features and two words. The novelty of the method lies in the use of different kernels (such as the “square-root” or Laplacian kernel) in the non-parametric density estimation as well as the utilization of configurations to explore semantic relationships among concepts. Thus, it is easy to compare and analyze performances over several different configurations.

In addition to the classical graph-based learning methods, some researchers strive to solve the problem from other viewpoints. For

example, there exist studies concentrating on exploiting the local label correlations [18] or underlying correlations among labels [85] and on handling the missing tag issue [86]. The ML-LOC model [18] allows label correlations to be used locally by assuming that the instances can be separated into different groups and each group shares a subset of label correlations. As in the real world tasks, a label correlation may be shared by only a subset of instances rather than all the instances. The MLDL model [85] uses a multi-label dictionary learning algorithm to explore the underlying correlation among labels. A noteworthy point is that the MLDL model puts the label correlation in the input feature space rather than in the output label space. The MLRank model [87] formulates image annotation as a Multi-correlation Learning to the Rank problem where the visual similarity among images and the semantic relevance among tags are explored simultaneously. By assuming the ranking objects are independent, the study ranks relational data using the consistency between “visual similarity” and “label relevance”, which indicates similar images are usually annotated with relevance tags to reflect similar semantic themes.

Many AIA methods are based on a common assumption that a complete label assignment for each training image should be provided. However, in practice, it is difficult to obtain complete label assignment for each training image. Therefore, the MLML model [86] presents a multi-label learning model to handle this issue. The MLML model trains classifiers by enforcing the local smoothness among the label assignments and the consistency between the predicted labels and the provided labels. Ivacic-Kos et al. [88] presented a two-tier annotation model where the first tier corresponded to object-level annotation and the second tier to scene-level annotation. In the proposed model, objects refer to things that can be recognized in an image, such as the sea and building, while scene labels represent the context of the whole image, such as natural scene and wild animals. There are two assumptions used for the framework of this two-tier model. The first one is that there can be many objects in an image, but one image can only be classified into one scene. Since many object labels may be assigned to an image, the object-level annotation is treated as a multi-label classification problem for low-level features extracted from images. The second assumption is that there are typical objects of which scenes are composed. The scene-level annotation relies on the originally developed inference-based algorithm to support reasoning about scenes and objects.

Although discriminative model-based AIA methods are intuitionistic and efficient, there still exist some shortcomings. First, the correlation between image visual features and labels are often neglected by some discriminative model-based AIA methods. Second, the discriminative model-based AIA methods mainly use label correlation and thus are sensitive to the quality of training datasets. As such, appropriate label correlations may not be obtained since the frequency of each label in the training dataset is unequal. Finally, the graph-based image annotation methods are transductive and can only predict labels for specified unlabeled samples. In other words, to annotate a new test image, the test image should be first added to the unlabeled set and then the training phase will be repeated. However, this is unpractical for the mass image annotation tasks nowadays.

2.4. Tag completion-based AIA methods

The tag completion-based AIA methods are quite different from the other four types of image annotation methods, and many studies in this field have been performed in recent years [20,24,40,41,53–55,89]. The AIA methods often assume that images in the training dataset are completely annotated with appropriate tags. However, recent studies have shown that manual tags are often unreliable and incompatible. The novelty of the tag

completion-based AIA methods is that missing tags can be filled automatically without training processes and that noisy tags for given images can be corrected.

Although various methods for tag completion with different frame structures have been developed, they focus on content consistency and tag relationship. The whole dataset can be represented as an initial tagging matrix with each row as an image and each column as a tag. The tag completion is operated at a matrix level by recovering the initial matrix through identifying correct associations between images and labels. The tag completion-based AIA methods can be further divided into several subclasses: matrix completion-based methods, linear space reconstruction-based methods, subspace cluster-based methods, and low-rank matrix factorization-based methods.

2.4.1. Matrix completion-based methods

The TMC model [24] casts tag completion into a problem of matrix completion. The relationship between tags and images is described by a tag matrix, where each entry in the tag matrix represents the relevance of a tag to an image. Let n and m be the number of images and available tags, respectively. Let $\hat{T} \in \mathbb{R}^{n \times m}$ denote the tag matrix derived from manual annotation, where $T_{ij}^{\hat{}} = 1$ indicates that image i is labeled with tag j while $T_{ij}^{\hat{}} = 0$ indicates that image i is not labeled with tag j . Then visual features of the image are represented by matrix $V \in \mathbb{R}^{n \times d}$, where each image can be described with d kinds of features. In addition, the label correlation $R \in \mathbb{R}^{m \times m}$ is considered in this model, and R_{ij} represents the correlation between tag i and tag j . Finally, matrix $T \in \mathbb{R}^{n \times m}$ denotes the complete tag matrix that needs to be computed. Thus, the TMC model searches for an optimal tag matrix which preserves correlation structures for both images and labels directly by keeping the consistency with the given labels. The objective is to minimize the discrepancy between the correlation in visual content and the correlation in semantic tags. The following optimization equation is used to compute the complete tag matrix:

$$\min_{T \in \mathbb{R}^{n \times m}} \|TT^T - VV^T\|_F^2 + \lambda \|T^T T - R\|_F^2 + \eta \left\| T - \hat{T} \right\|_F^2 \quad (22)$$

Where $\lambda > 0$ and $\eta > 0$ are parameters determined by cross validations.

Qin et al. [53] formulated the image annotation as a constrained optimization problem. In [53], the TMC model is improved by introducing the tag constraints into the optimization and by solving the optimization via the efficient linearized alternating direction method.

2.4.2. Linear space reconstruction-based methods

Lin et al. [20] presented a scheme for image tag completion via image-specific and tag-specific Linear Sparse Reconstructions (LSR). The LSR model formulates the image-specific and tag-specific reconstructions as a convex optimization problem under constraints of sparsity. The image-specific reconstruction utilizes the visual and semantic similarities among images, while the tag-specific reconstruction mines the concurrence between tags. Finally, LSR normalizes and merges tag completion results derived from the two linear reconstructions by adopting a weighted linear combination in Eq. (23):

$$\Omega = \delta T + (1 - \delta)R \quad (23)$$

Where Ω is the expected final result, T and R are the normalized completion results from image-specific and tag-specific reconstructions, respectively, and δ is a weighting parameter range from 0 to 1.

Given the incomplete initial tagging matrix $D_{m \times n}$, where m and n denote the number of images and tags respectively, image-specific reconstruction is intended to perform tag completion from

the point of row. As mentioned above, the low-level features and high level tagging vectors are both taken into consideration. Assume the feature vector of a to-be-reconstructed image is f_{l*1} , where l is the dimension of the feature vector, then the image-specific reconstruction for low-level features can be formulated as follows:

$$\Theta_1 = \min_a \|f - Fa\|_2^2 + \lambda \|a\|_1 \quad (24)$$

$F_{l*(m-1)}$ is a dictionary matrix consisting of feature vectors of other images, $a_{(m-1)*1}$ is the objective weighting vector with each element representing the weight of the corresponding image in the linear sparse reconstruction of f , and λ is a tuning factor for penalizing the non-sparsity of a .

With respect to the linear sparse reconstruction for high level tagging vectors, a group sparse structure is introduced on the basis of the observation that images associated with an identical tag probably share more common semantic content and thus form a group. The objective function is formulated as follows:

$$\Theta_2 = \min_{\beta} \left\| W(t - \hat{T}\beta) \right\|_2^2 + w \sum_{i=1}^n \|g_i\|_2 \quad (25)$$

Where t_{n*1} is the tagging vector of a to-be-reconstructed image, $\hat{T}_{n*(m-1)}$ is the dictionary matrix containing tagging vectors of other images, $\beta_{(m-1)*1}$ is the objective weighting vector denoting the weights of other images in the linear sparse reconstruction of t , and w is a tuning factor for balancing the group sparsity. Here the group sparsity $\sum_{i=1}^n \|g_i\|_2$ separately uses $L2$ norm for smoothing intragroup weights and $L1$ norm for emphasizing inter-group sparsity. Additionally, W is a diagonal matrix for weighting the reconstruction residual of each entry in t , defined as $w_{i,i} = \exp(t_i)$.

Furthermore, for image-specific reconstruction, the LSR model integrates the two objective functions above into a unified optimization framework as Eq. (26):

$$\Theta = \min_{a,\beta} \|f - Fa\|_2^2 + \lambda \|a\|_1 + \mu \left(\left\| W(t - \hat{T}\beta) \right\|_2^2 + w \sum_{i=1}^n \|g_i\|_2 \right) + \nu \|a - \beta\|_2^2 \quad (26)$$

Where μ is a weighting parameter for balancing the reconstructions of low-level features and high-level tagging vectors, and ν is a tuning factor to penalizing the difference between a and β . Then the optimal a and β can be merged for obtaining a reconstructed tagging vector t' for the target image, as shown in formula (27):

$$t' = \hat{T}(\rho a + (1 - \rho)\beta) \quad (27)$$

By performing linear sparse reconstructions for all to-be-completed images, all the corresponding t' constitute the image-specific reconstructed tagging matrix T_{m*n} .

The tag-specific reconstruction is intended to perform tag completion for the incomplete initial tagging matrix D_{m*n} from the perspective of column. The tagging column vector in D of a to-be-completed tag is denoted as τ_{m*1} . The dictionary matrix consisting of other tagging column vectors is denoted as $\hat{R}_{m*(n-1)}$. The process of tag-specific reconstruction can be formulated as Eq. (28):

$$\psi = \min_{\gamma} \left\| W'(\tau - \hat{R}\gamma) \right\|_2^2 + \xi \|\gamma\|_1 \quad (28)$$

Where $\gamma_{(n-1)*1}$ is the objective weighting vector with each element representing the weight of the corresponding tag in the reconstruction, and ξ is a tuning factor for penalizing the non-sparsity of γ . Additionally, W' is a diagonal weighting matrix for

the reconstruction residuals of all entries in τ , which is defined in the same way as W in formula (25). The tag-specific objective function is convex, and thus there exists a global optimal γ , which can be utilized to obtain a reconstructed tagging column vector $r' = \hat{R}\gamma$ for the target tag. Finally, all the corresponding r' constitute the tag-specific reconstructed tagging matrix R_{m*n} .

Specifically, the LSR model improves the tag completion performance for each tag and each image but probably at the cost of high computational complexity in case of high-dimensional images or tags.

Lin et al. further extended and improved LSR, namely, the Dual-view Linear Sparse Reconstruction (DSLRS) model [89]. DSLRS performs tag completion via reconstructing each image and each tag, respectively. The goal of DSLRS is to make the reconstruction methods more effective and practical since the LSR method is computationally expensive. The study mainly concentrates on the utilization of the same reconstruction weights of feature and initial tags instead of different weights and the exploration of a better strategy for combining image-view and tag-view reconstruction tagging vectors. To combine reconstruction tagging vectors, the study treats image-view reconstructed tagging vector t_1 and tag-view reconstructed tagging vector t_2 as results of retrieving related tags for a given to-be-completed image and its initial labeled tags from two distinct “search engines”.

2.4.3. Subspace clustering-based methods

The Subspace Clustering and Matrix Completion (SCMC) model [55] is proposed to perform tag completion and refinement sequentially. It first treats tag completion task in a subspace clustering framework, by assuming that images are sampled from a union of multiple linear subspaces and that their corresponding tags form a compatible sub-matrix. The model then refines the tag matrix by using a matrix completion model to narrow the semantic gap as well as the sparsity of the tag matrix.

SCMC uses LRR to cluster the visual feature vectors into different subspace. The LRR algorithm outputs a block-diagonal affinity matrix, in which each sub-matrix corresponds to a subspace (cluster). The image can be clustered according to the affinity matrix, and tag completion can be performed by transferring tags in each cluster. The set of image feature vectors can be denoted as $X = [x_1, x_2, \dots, x_n]$, drawn from a union of k subspaces $\{s_i\}_{i=1}^k$. Each column of X is a feature vector in R^D and can be represented by a linear combination of the basis in a “dictionary”. The LRR model uses the matrix X itself as the dictionary and takes error into consideration:

$$\begin{aligned} \min_{Z,E} \|Z\|_* + \mu \|E\|_{2,1} \\ \text{s.t.}, X = XZ + E \end{aligned} \quad (29)$$

Where $Z = [z_1, z_2, \dots, z_n]$ is the coefficient matrix with each z_i being the representation of x_i and E is the sparse error matrix. Specifically, the SCMC adopts a tag transfer algorithm [50] to complete tags in each cluster using tag frequency, tag co-occurrence, and local frequency.

Tag refinement aims to correct noisy tags. The tag refinement issue can be treated as matrix completion, where the puzzle is to ‘delete’ the unbefitting tags in the user-item preference matrix and ‘complete’ the missing ones given a sample of observed preferences. It constructs a tag matrix $P \in R_{N_{im}*N_{tg}}$, where each row corresponds to an image (the number of images is N_{im}), and each column corresponds to a tag (the number of tags is N_{tg}), such that $p_{ij} = 1$ if image i is annotated with tag j and $p_{ij} = 0$ otherwise. Let $x_i \in R_{f_{im}}$ denote the feature vector of image i , and $y_j \in R_{f_{tg}}$ denote the word embedding of tag j , which is computed from pre-trained word-to-vector (word2vec). The matrix completion problem can be

viewed as a multi-label regression framework as Eq. (30):

$$\min_{Z \in R_{f_{im} \times f_{tg}}} (p_{i,j} - x_i^T Z y_j)^2 + \lambda \|Z\|_* \quad (30)$$

The loss function penalizes the deviation of estimated entries from the observations. The regularization parameter λ trades off losses on observed entries and the low-rankness constraint.

The subspace cluster-based methods are superior over traditional clustering methods because (1) the subspace clustering-based methods do not need to measure the similarity between features and (2) the subspace clustering-based methods can precisely model distributions of the image feature.

2.4.4. Low-rank matrix factorization-based methods

Li et al. [54] proposed a formulation with the features of low-rank and error sparsity and local reconstruction structure consistency. The study attempts to implement linear reconstruction in both the feature space and label space similar to [20]. In addition, [54] uses a low-rank and error sparsity method to decompose the initial tagging matrix into a sparse error matrix, a factorization of a basis matrix, and a sparse coefficient matrix. The merit of the low-rank and error sparsity method is in its ability to handle the noisy data [20].

The Locally Sensitive Low-rank model (LSLR) [40] performs image tag completion by estimating a global nonlinear model with a collection of local linear models. Compared with methods based on the linear structure, nonlinear models can explore the complex correlation between images and tags efficiently. Given the incomplete initial tagging matrix $D_{m \times n}$, where m and n denotes the number of images and tags, respectively, and the visual feature matrix is denoted as $X_{n \times d}$, and d represents the dimension of visual features. The objective of tag completion is to recover the complete tag matrix Y . The pre-processing is to learn suitable representation for the data partition. All images in the dataset are divided into several clusters according to the semantic content. Then a local model is estimated by factorizing the complete Y_i matrix into a basis matrix W_i and a sparse coefficient matrix H_i , shown in Eq. (31):

$$Y_i = W_i H_i, \forall i \in 1, 2, \dots, c \quad (31)$$

$$W_i \in R^{n_i \times k} \text{ and } H_i \in R^{k \times m}$$

Where n_i is the number of samples in the i -th cluster. Then the locality sensitive low-rank model is calculated as follows:

$$f = \sum_{i=1}^c (L_i + \lambda R_{g_i}) \quad (32)$$

Where R_{g_i} represents the global consensus regularizer to mitigate the risk of overfitting. L_i is the local model for the i -th cluster and the loss function which can be further broken down as the following equation:

$$L_i = \|D_i - W_i H_i\|_F^2 + \eta R_{W_i} + \gamma R_{H_i} + 2\beta \|H_i\|_1 \quad (33)$$

Where D_i is the initial tag matrix for cluster i , R_{W_i} and R_{H_i} are regularization for W_i and H_i . η , γ and β are input parameters. The final complete tag matrix Y is obtained by integrating all the sub-matrices Y_i .

The tag completion-based AIA methods have attracted much attention and made considerable achievements recently. Tag completion is robust in AIA since it does not require the training process to predict labels for a given image. Traditionally, a large dataset with reliable labels is required in the training process and is completed by manual ways. Therefore, missing or noisy tags can potentially lead to a biased estimation of predicted models. In addition, tag completion-based AIA can automatically fill in the missing tags and correct biased tags. The Tag completion-based AIA

methods are efficient and have scale independence. However, the tag completion-based AIA methods suffer some disadvantages. The most obvious weakness is the transformation of the tag completion process to an optimization problem. The process of optimizing the objective function may be time-consuming and computation-complex, and cannot guarantee global optimization.

2.5. Deep learning-based AIA methods

The most recent decade has witnessed the significant development of deep learning techniques, which enables deep learning-based feature representation to solve AIA task. The latest advances in deep learning allow a variety of deep models for large-scale image annotation. The deep-learning based AIA can be summarized in two aspects. Firstly, robust visual features are generated by using convolution neural network (CNN), for image annotation [13,25,42–45]. Secondly, side information (such as semantic label relationships) are fully extracted through deep learning techniques for AIA [26,46–49]. The deep learning-based AIA is a quite new but promising direction for AIA.

2.5.1. Robust visual features

Robust visual features are the most fundamental factors for image annotation. The traditional handcrafted features are stochastic and dissatisfactory. Inspired by the success of CNN in computer vision, researchers tend to use CNN to generate robust visual features for AIA.

The CNN+WARP (Weighted Approximate Ranking) model [25] uses ranking to train deep convolutional neural networks for multi-label image annotation problems. The study used five convolutional layers and three densely connected layers in the CNN architecture. The loss function is defined as a multi-label variant of the WARP loss function with the top- k annotation accuracy optimized by a stochastic sampling approach. For a set of images x , the convolutional network is denoted by $f(\cdot)$ where the convolutional layers and dense connected layers filter the images. The output of $f(\cdot)$ is a scoring function at the data point x containing a vector of activations. It is assumed that there are n images and c tags for training. The WARP loss function minimizes the following formula:

$$J = \sum_{i=1}^r \sum_{j=1}^{c_+} \sum_{k=1}^{c_-} L(r_j) \max(0, 1 - f_j(x_i) + f_k(x_i)) \quad (34)$$

where $L(\cdot)$ is a weighting function for different ranks, and r_j is the rank for the j th class for image i . The weighting function $L(\cdot)$ in (35) is defined as:

$$L(\cdot) = \sum_{j=1}^r a_j \quad (35)$$

Where a_j is defined as $1/j$, while the weights defined by $L(\cdot)$ control the top- k of the optimization. In particular, if a positive label is ranked top in the label list, then $L(\cdot)$ will assign a small weight to the loss and will not cost the loss too much. However, if a positive label is not ranked top, $L(\cdot)$ will assign a much larger weight to the loss, which pushes the positive label to the top. In addition, the rank r_j is estimated by the formulation (36) for c classes and s sampling trials.

$$r_j = \left\lfloor \frac{c-1}{s} \right\rfloor \quad (36)$$

Then the sub-gradient for this layer during optimization is computed.

For comparisons, Gong et al. [25] used a set of 9 different visual features (GIST, D-SIFT, D-CSIFT, D-RGBSIFT, H-SIFT, H-CSIFT, H-RGBSIFT, HOG and Color feature) and combined them to serve as

baseline features. Based on these features, two simple but powerful classifiers (kNN and SVM) were performed for image annotation. Through the comparison of CNN feature-based frameworks with baseline features-based classifiers, experimental results showed that the deep network had a better performance than existing visual-feature-based methods in image annotation.

Similarly, Mayhew et al. [42] trained two different image annotation algorithms (TagProp [80] and 2PKNN [16]) with features derived from the two CNN architectures (AlexNet and VGG-16). Experimental results clearly revealed that better, at least similar, annotation performance was achieved by using features derived from a deep convolutional neural network than by using larger handcrafted features. Moreover, their study proved the idea that complementary information in both the deep and handcrafted features could be jointly used to enhance predictive performance.

The CCA-KNN model [13] is based on the Canonical Correlation Analysis (CCA) framework that helps in modeling both visual features (CNN feature) and textual features (word embedding vectors) of the data. It was shown that CNN features were advantageous over 15 handcrafted features in the existing models, including JEC [50], 2PKNN [16] and SVM-DMBRM [14]. Furthermore, their study showed that word embedding vectors performed better than binary vectors as a representation of the tags associated with an image.

Given that the quality of the original label of the dataset has a great influence on the AIA performance, the Multitask Voting automatic image annotation CNN (MVAIACNN) model [43] adopts the Multitask Voting method via a multitask learning mechanism for the selection of training and test datasets. By combining the multitask learning method with the Bayesian probability model, the MV method achieves the adaptive label. Finally, the AIACNN model, which contains five convolutional layers to extract features hierarchically and four pooling layers, was proposed, followed by two fully-connected layers and the softmax output layer indicating identity classes. The MVAIACNN has shallow layers and regards each category as a label directly, using the raw images as inputs for large-scale image annotation. To a certain extent, fewer layers reduce the efficiency defects caused by more layers.

Johnson et al. [44] proposed a model to generate neighborhoods of related images with similar social-network metadata by using Jaccard similarities. The metadata carried by most images on the web, such as user-generated tags and community-curated groups, can be highly informative as to the semantic contents of images. The types of image metadata considered in [44] include user tags, image photo-sets, and image groups. Photo-sets are images commonly gathered by the same user. For example, pictures from a sports meeting are uploaded by the same social network user. Image groups are a set of images which belong to the same circumstance, concept, and event in the social network site, e.g., a set of images that all contain Satsuma in the social network.

The multi-view stacked auto-encoder (MVSAE) model [45] builds a novel SAE framework with the sigmoid predictor for image annotation. Image features are usually used as model inputs and keywords are used as model objects in most deep neural network-based AIA models. Meanwhile, several hidden layers are set for modeling the complex relationship between features and tags. Since the performance of the deep neural network is highly dependent on initial parameters, the MVSAE model adopts the pre-trained parameters to for model optimization. Specifically, at first, visual feature I as the model input x is used to train the SAE for generating the initial keyword probability distribution D_1 . Then I and D_1 are used as new model inputs x to retrain the SAE model for generating the final keyword probability distribution D_2 . Finally, image keywords \hat{T} are obtained from D_2 .

2.5.2. Side information

The CNN-RNN framework [26] utilizes recurrent neural networks (RNN) to capture high-order label relationships at a moderate level of computational complexity. In this framework, the CNN and RNN are jointly utilized to derive image representation and the correlation between the adjacent labels, based on which the final outputs, such as label probability, are computed. It is important to rank the labels for training multi-label CNN-RNN models. In the CNN-RNN framework, the orders of the labels are determined according to their occurrence frequencies in the training data.

The RIA model [46] also uses the CNN-RNN framework for image annotation. Inspired by the recent success of RNN in image captioning [90], RIA uses CNN to extract image visual features, and RNN to generate the tag sequence from the visual features one by one. The advantages of RNN in AIA lie in two aspects. On the one hand, RNN can generate output with different length. On the other hand, RNN is able to refer to previous inputs in predicting the current time step output. Image captioning [90] aims to generate sentences in a natural order for training the RNN model. It is noted that that the RNN model uses the frequent-first rule rather than the rare-first rule adopted by the RIA model.

The Deep Multiple Instance Learning (DMIL) model [47] presents a framework for learning correspondences between image regions and keywords. In DMIL, two sets of instances, object proposals, and keyword, are learned simultaneously by a joint deep multi-instance learning framework. Specifically, the DMIL uses a CNN that contains five convolutional layers, a pooling layer, and three fully connected layers for learning visual representation. Then, it uses another deep neural network framework that contains one input layer, one hidden layer, and one output layer with softmax for multi-instance learning. Finally, it combines both the image and text outputs in the fully connected layer.

The structured inference neural network (SINN) [48] is presented for layered label predictions. The main idea of SINN is that an image with various objects and abundant attributes can be assigned with fine-grained labels and rough-grained label to describe specific contents and abstract concepts, respectively. Firstly, CNN features were extracted as visual activations at each concept layer. Concept layers are stacked from fine-grained to coarser levels. Secondly, label relations between consecutive layers are generated as a layered graph, in which each concept layer represents a time-step of RNN. The inter-layer correlations and intra-layer relations are obtained according to the time-steps.

Niu et al. [49] mainly focused on two issues for large-scale image annotation. The first issue is how to learn stronger and robust features for various images. Unlike the conventional methods that only adopt CNN to obtain image features, their study extracted textual features from noisy tags by a multi-layer perception subnetwork to boost the visual features extracted by a multi-scale CNN subnetwork. The integrated features were connected in a fully connected layer for image annotation. The second issue is that how to annotate an image with an automatically-determined number of class labels. Different from the methods based on RNN [26,46], the annotation method in [49] views label quantity prediction as a regression problem. It has been proved that the quantity prediction has great effectiveness on image annotation.

In summary, the deep learning-based AIA methods have brought both challenges and opportunities to AIA. On the one hand, recent progress and breakthroughs in deep learning significantly improve the AIA performance on large-scale image datasets. On the other hand, there are still three main shortcomings in the deep learning-based AIA methods. The first issue is the decreased efficiency of deep-learning based AIA methods as the depth and breadth of deep neural networks increase. Although Deep Neural Networks can learn complicated relationships between inputs and outputs, they are prone to local optimum and difficult in converg-

Table 1

A survey of different types of AIA methods.

Annotation methods	Advantages	Disadvantages
Generative model-based AIA methods	Conditional probabilistic distribution, well-formed theory, alternative number of labels.	Require prior image segmentation, expensive training and computation, sensitive to noisy data, parametric, might not be optimal.
Nearest neighbor-based AIA methods	Conceptually simple, non-parametric, do not require prior image segmentation. large dataset.	Sensitive to small dataset, distance metric learning, sensitive to cluster result, fixed number of labels.
Discriminative model-based AIA methods	Multi-label, graph framework usual, computation-efficient.	Sensitive to label-imbalance, classes relevance, parametric.
Tag completion-based AIA methods	Robust to noisy data, more scalable in dataset size, non-train process, non-parametric.	Sparse matrix, optimization problem, various constraints on the relationship between text and visual components, little guarantees that the annotations are optimal.
Deep learning-based AIA methods	Deal with mass data, learn very complicated relationships, derive Robust features, no manual selection is required, Obtain sufficiently side information, alternative number of labels	Local optimum, Vast training images, Training process cannot be controlled.

ing by using the back-propagation algorithm. Second, the training process of Deep Neural Networks cannot be controlled. At present, there is a lack of unified and complete theoretical guidance on the choice of network structure. Last but not least, although RNN has been combined with CNN to solve issues related to label quantity prediction and label dependencies for large-scale image annotation, a better solution to rank label orders is still needed since RNN requires an ordered sequential list as input.

2.6. Summary

In the previous section, we discussed five types of AIA methods in terms of the ideas, models, algorithms, and open issues. We summarize the advantages and disadvantages of AIA methods in Table 1.

The generative model-based, the discriminative model-based, and the deep learning-based AIA methods are all learning-based methods. The generative model-based methods annotate images based on the conditional probability over images and labels. The discriminative model-based AIA methods treat image annotation as a multi-label classification problem. Thus, the correlation between classes is critical but cannot be solved directly by a binary classifier. The deep learning-based methods generally use CNN to obtain robust visual features or use different network frameworks, such as RNN, to exploit the side information, i.e., label correlation, for AIA. Comparatively, the nearest neighbor model-based AIA methods adopt a two-step framework to annotate images. Similar images for the query image are retrieved first and then used to predict for the query. The tag completion-based AIA methods do not require the training process. In addition, noisy tags can be corrected for the given image automatically. It should be noted that all the five types of AIA methods dictate appropriate ways to characterize the semantic context to overcome the semantic gap.

3. Dataset and evaluation measures

To compare and analyze the performance of different AIA methods, standard datasets and metrics are required. In this section, we summarize popular benchmark datasets publicly available and standard measure metrics used for evaluating AIA methods.

3.1. Dataset

Table 2 lists some datasets that are frequently used for assessment of AIA methods.

Corel 5K: Since its first utilization in [10], Corel 5K has become an important benchmark dataset. Corel 5K has 5000 images from 50 categories. Each category includes 100 images, and there are 260 keywords totally in the vocabulary. Each image is labeled by

Table 2

Descriptive statistics of the three benchmark datasets.

Dataset	Number of images	Vocabulary size	Training images	Testing images	Words per img	Img per word
Corel 5K	5000	260	4500	500	3.4	58.6
ESP Game	20,770	268	18,689	2081	4.7	362.7
IAPR TC-12	19,627	291	17,665	1962	5.7	347.7

1–5 keywords. Usually, Corel 5K is divided into 3 parts: a training set of 4000 images, a validation set of 500 images, and a test set of 500 images. In other words, the total number of images for training is 4500 and for validation is 500.

ESP Game : The ESP dataset consists of images collected from the ESP online labeling game [91]. In the ESP game, two players gain points by predicting the same keyword for an image without communication. This dataset is very challenging as it contains a wide variety of images including: logos, drawing, and scenery and personal photos. A list of colloquial words is also accumulated. This dataset contains 67796 images totally, but the part used for experiments usually consists of 20,770 images with 268 keywords, including 18,689 training images and 2081 test images [16].

IAPR TC-12 : Different from other similar datasets, IAPR TC-12 is described in three languages (English, German and Spanish) and is typically used for cross-lingual retrieval. IAPR TC-12 is first proposed in [92] and includes 20,000 images covering several different scenes such as sports, city pictures, landscape shots, animals, action shots, buildings, and plants that appear frequently in our daily life. The generally used dataset includes a total of 19,627 images with 291 keywords, i.e., 17665 images for training and 1962 images for validation [80].

NUS-WIDE : NUS-WIDE is created by NUS's Lab for Media Search at National University of Singapore [93]. The NUS-WIDE dataset contains 269,648 images and 5018 tags from Flickr. There are a total of 1000 tags after removing noisy and rare tags. These images are further manually annotated into 81 concepts by human annotators.

MS-COCO : The Microsoft COCO (MS-COCO) dataset is used for image recognition, segmentation, and caption. It contains 123 thousand images of 80 objects types with per-instance segmentation labels. It is worth noting that the number of images used according to MS-COCO dataset also varies from different studies. For example, the MS-COCO dataset used in [49] consists of 87,188 images, 80 class labels, and 1,000 most frequent noisy tags from the 2015 MS-COCO image captioning challenge. The training/test split of the MSCOCO dataset is 56,414/30,774. While 82,783 images are utilized as training data, and 40,504 images are employed as testing data in the experiment of [26]. The details of the NUS-WIDE and the MS-COCO datasets are not listed in Table 2 since the num-






models	Examples			Examples		
	Image	Ground Truth	Predictions	Image	Ground Truth	Predictions
CRM		bear	snow		smoke	train
		polar	bear		train	railroad
		snow	polar		locomotive	tracks
		tundra	tundra		railroad	locomotive
MBRM						
SML		bear	polar		sky	plane
		polar	tundra		jet	jet
		snow	bear		plane	smoke
		tundra	snow		smoke	flight
			ice			prop
JEC		bear	bear		sky	sky
		snow	snow		jet	jet
		grass	grass		plane	plane
		deer	deer		smoke	smoke
			white-tailed			formation
TMC		bear	snow		cars	formula
		polar	polar		formula	cars
		snow	sky		tracks	building
		tundra	ocean		wall	fly
			fox			athlete
MLDL		tree	tree		sky	sky
		horse	horse		jet	plane
		mare	mare		plane	flight
		foals	grass		smoke	smoke
			foals			jet
FastTag						
Tagprop		iguana	iguana			
		lizard	marine			
		marine	lizard			
		rocks	water			
			sky			

Fig. 2. Image tags identified by manual annotation (Ground Truth) and some well-behaved AIA models including CRM, MBRM, SML, JEC, TMC, MLDL, FastTag, Tagprop, 2pKNN, CCA-knn, WARP, RNN, and MLRP (Predictions).

ber of images /tags of such datasets used in different studies is discrepant.

Apart from the five image datasets, there are also 4 datasets used to assess the performance of AIA methods. (1) **The MSRC** dataset [33–35,39], provided by the computer vision group at Mi-

crosoft Research Cambridge, contains 591 images annotated by 23 classes. (2) The **Flickr30** dataset [20,40,54,89] is also a real-world dataset crawled from Flickr and is constructed by submitting 30 non-abstract concepts as queries to Flickr and then collecting the top 1000 of the retrieved images for each concept.








2pKNN		fields	fields		cars	wall
		horses	horses		formula	cars
		mare	mare		tracks	balcony
		foals	foals		wall	tracks
			tree			formula
CCA-knn						
WARP		animal	horse		boat	lake
		cloud	animal		cloud	ocean
		cow	cloud		ocean	cloud
		grass	grass		vehicle	sky
			sky		water	water
RNN		clouds	clouds			
		sun	sky			
		sunset	sun			
			sunset			
MLRP		animal	animal		airport	clouds
		tiger	tiger		clouds	plane
					military	airport
					plane	sky
					sky	military

Fig. 2. Continued

(3) Another dataset widely used in computer vision is the **LabelMe** dataset [69,84], which is a collection of 72,852 images containing more than 10000 concepts. Furthermore, there exist some methods [36,51,52] to retrieve images by crawling from the world web according to the user intent. (4) The **TRECVID dataset** [12,28,33,34,38] is also widely used for image annotation as well as video annotation. The **TRECVID 2005** dataset contains 137 broadcast videos from 13 different programs in English, Arabic, and Chinese, and are segmented into 74,523 sub-shot.

3.2. Evaluation metrics

There are several metrics to evaluate the performance of various kinds of AIA methods. Among them, the recall and precision, F1-score, and N+ are quite popular.

Recall and Precision: For a given keyword, let m_1 be the number of images in the test dataset annotated with the label, and m_2 be the number of images correctly annotated with the label. m_3 means the number of images annotated with the label using the ground-truth data. Then $recall = \frac{m_2}{m_3}$, and $precision = \frac{m_2}{m_1}$.

Recall measures the ability to retrieve the relevant information, while the precision measure the ability to refuse the uncorrelated information. Recall and precision are usually combined to evaluate the performance of the AIA models. However, it is difficult to evaluate AIA models only using recall and precision since the two metrics conflict with each other.

Note that test images are usually forced to be annotated with k (usually 5) labels by the AIA methods, even if images are labeled with fewer or more tags in the ground truth. Therefore, it may yield biased recall and precision values even if a model predicts all ground truth labels.

F1-score is computed as, $F_1 = \frac{2 * p * R}{p + R}$.

Since either the recall or precision is not adequate to comprehensively assess the performance of AIA models, they have been integrated into one evaluation index. In addition, the F1-score can

be used to measure the robustness of the AIA methods. The larger the F1-score is, the more robust the model will be.

N+: N+ is used to denote the number of keywords that are correctly assigned to at least one test image. The indicator also means the number of keywords with positive recall. A high value of N+ means the good performance of AIA methods.

4. Performance comparison

In this section, we assess the performance of some typical or well-behaved models, including: CRM [11], MBRM [3], SML [5], JEC [50], TMC [24], MLDL [85], FastTag [94], TagProp [80], 2PKNN [16], CCA-KNN model [13], CNN+WARP [25], CNN+RNN [26] and MS-CNN+MLP [49].

A comprehensive comparison over those models is provided based on the usual Corel 5k, ESP Game, IAPRTC-12, NUS-WIDE, and MS-COCO, using the evaluation metrics including Recall(R), Precision (P), F1-score and Number of keywords (N+). Table 3 shows the performances of some typical or well-behaved AIA models.

Fig. 2 presents some image annotation examples performed by the selected models. For each image, images tags identified from the manual annotation (left panel) and AIA (right panel) are provided for comparisons. Almost all generative model-based AIA methods are inferior to other kinds of AIA methods since the generative models cannot capture the intricate dependencies between image features and labels. Although tag completion-based AIA methods have many advantages (e.g., they are robust to noises and do not require the training phase), their performances for annotating images are not satisfactory. However, tag completion-based AIA still has great potential as a new research direction. The nearest neighbor model-based AIA methods show great annotation performance compared to other methods by incorporating DML and graph-learning. The deep learning-based AIA methods show the best performance compared to other methods since these methods can acquire robust features. Furthermore, the deep

Table 3

Performances of some typical or well-behaved AIA models including CRM, MBRM, SML, JEC, TMC, MLDL, FastTag, Tagprop, 2pKNN, CCA-KNN, CNN+WARP, CNN+RNN, and MS-CNN+MLP.

Models	Corel 5k(%)				ESP Game(%)				IAPRTC-12(%)				NUS-WIDE(%)				MSCOCO(%)			
	P	R	F1	N+	P	R	F1	N+	P	R	F1	N+	P	R	F1	N+	P	R	F1	N+
CRM	16	19	17	107	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MBRM	24	25	24	122	18	19	18	209	24	23	23	223	–	–	–	–	–	–	–	–
SML	23	29	26	137	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
JEC	27	32	29	139	24	19	21	222	29	19	23	211	–	–	–	–	–	–	–	–
TMC	16	23	19	124	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MLDL	45	49	47	198	56	31	40	259	56	40	47	282	–	–	–	–	–	–	–	–
FastTag	32	43	37	166	46	22	30	247	47	26	34	280	–	–	–	–	–	–	–	–
Tagprop	33	42	37	160	39	27	32	238	45	34	39	260	–	–	–	–	–	–	–	–
2pKNN	44	46	45	191	53	27	36	259	54	37	44	278	–	–	–	–	–	–	–	–
CCA-KNN	42	52	46	201	46	36	41	260	45	38	41	278	–	–	–	–	–	–	–	–
CNN+WARP	–	–	–	–	–	–	–	–	–	–	–	–	32	36	34	97	53	60	56	–
CNN+RNN	–	–	–	–	–	–	–	–	–	–	–	–	41	31	35	–	66	56	61	–
MS-CNN+MLP	–	–	–	–	–	–	–	–	–	–	–	–	80	61	69	–	75	65	70	–

Note: The test images in CRM, MBRM, SML, JEC, TMC, MLDL, FastTag, Tagprop, 2pKNN, and CCA-KNN are labeled by 5 keywords while the test images in CNN+WARP, CNN+RNN, and MS-CNN+MLP are labeled by 3 keywords.

learning-based AIA methods usually annotate an image with an automatically-determined number of class labels.

Overall, several conclusions can be made. First, manual annotation does not always provide all relevant tags because of its subjectivity and ambiguity, and AIA methods may help perfect the annotations (the words are highlighted in red). Also, test images may contain irrelevant tags. It means that the quality of manual annotation should be considered for AIA. Second, most AIA methods have a fixed annotation length k . However, we argue that this convention may be insufficient since it is not the normal way that we humans annotate images. For example, the number of labels for each image provided by manual annotation is usually less than 5. Since most AIA methods yield 5 keywords for each image, it turns out that the AIA methods will annotate some unnecessary labels. Third, there is no consensus on annotation standards. Some people may prefer professional annotation words, e.g., “formula” “polar” “iguana”, while others prefer more general annotation words, e.g. “car”, “bear” and “lizard” (these example words are highlighted in green in Fig. 2). It is not proper to consider that all the keywords in an image’s annotation are equally important. Keywords, especially those professional annotations, may play an important role to describe an image. Last, it can be observed that keywords such as sky, sea, animal, and buildings have a greater chance to be used for annotation. Thus, it is necessary to handle the problem of “label-imbalance”.

5. Discussions and conclusions

In this paper, we provide a review on state-of-the-art AIA methods which can be classified into five categories: generative model-based AIA methods, nearest neighbor model-based AIA methods, discriminative model-based AIA methods, tag-completion model-based AIA methods, and deep learning-based AIA methods. Comparisons of the five types of AIA methods are presented in terms of the underlying idea, main contribution, model framework, computational complexity, and annotation accuracy. We also review some benchmark image datasets and evaluation metrics used to evaluate AIA methods. The comprehensive comparison of some typical models is provided as well.

The challenge of the AIA technique is to reduce the semantic gap between low-level visual image features captured by machines and high-level semantic concepts perceived by a human. Many studies have been conducted on mining the image-image, image-label and label-label correlation. Open issues, such as class-imbalance and weak-labeling of the training dataset, are also dis-

cussed in this paper. Furthermore, as an important similarity measure for image and image, feature and feature, label and label, DML is also discussed in this paper.

Overall, AIA is still a very challenging research area with some promising future directions. The first one is how to annotate the mass web images efficiently and effectively. It is well known that more and more online social media portals like Flickr provide a popular way to share and embed personal photographs on websites for users. It is expected that more than millions of new images are uploaded daily in recent years. Tags generated by web users are usually imperfect, and only a few of them are related to the content of the image. Therefore, it is difficult for researchers to derive appropriate annotation models from the large-scale and weak-correlation web images, not to mention an expectation of well-prediction. The huge number of images online may bring unprecedented difficulties and pressure to train a flexible model for image annotation. Furthermore, images uploaded by different users are random and may have nothing in common. On the contrary, in the Corel dataset, each category includes 100 images on the same topic and thus provide great convenience for the training process of the annotation model. Recently, web image annotation and retrieval have received a lot of attention. Ma et al. [95] formulated an annotation model via Subspace-Sparsity Collaborated Feature Selection and Gong et al. [96] used the canonical correlation analysis (CCA) for mapping visual and textual features to the same latent space, and then incorporated the third view to capture high-level image semantics.

The second direction is on deep learning-based methods on AIA. Most recent progress in this field demonstrate that deep learning can bridge the ‘semantic gap’. Further studies are highly needed in providing guidance on the choice of network structure, on the training process of Deep Neural Networks, and on the improvement of computation efficiency.

Another research direction worth attention is to describe images with sentences [90,97]. Images are usually labeled with fixed keywords; however, sentences have richer content and more compact and subtle representations of information than discrete keywords. It is also expected that with a good intermediate embedding space linking images and tags, the subsequent step of sentence generation may become easy [98]. Furthermore, using a human-like methodology like stochastic Petri-nets (SPN) graphs for deeper understanding of the natural language text sentence is a challenging problem with many applications [99].

Acknowledgments

The work of this paper was supported by the National Key Research and Development Program of China (No. 2016YFB0502603) and National Natural Science Foundation of China (No. 41771452).

The authors are grateful for the valuable comments made by the anonymous reviewers.

References

- [1] Y. Mori, H. Takahashi, R. Oka, Image-to-word Transformation Based on Dividing and Vector Quantizing Images with Words, in: First International Workshop on Multimedia Intelligent Storage and Retrieval Management, Citeseer, 1999, pp. 405–409.
- [2] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 119–126.
- [3] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2004.
- [4] A. Llorente, R. Manmatha, Image retrieval using Markov random fields and global image features, in: ACM International Conference on Image and Video Retrieval, 2010, pp. 243–250.
- [5] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 394–410.
- [6] X. Yang, X. Qian, T. Mei, Learning salient visual word for scalable mobile image retrieval, Pattern Recognit. 48 (10) (2015) 3093–3101.
- [7] M. Davis, M. Smith, J. Canny, N. Good, S. King, R. Janakiraman, Towards context-aware face recognition, in: Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 483–486.
- [8] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, A semi-automatic methodology for facial landmark annotation, in: Computer Vision and Pattern Recognition Workshops, 2013, pp. 896–903.
- [9] N. Snavely, S.M. Seitz, R. Szeliski, Photo tourism: exploring photo collections in 3d. ACM trans graph, ACM Trans. Graph. 25 (3) (2006) 835–846.
- [10] P. Duygulu, K. Barnard, J.F.G.D. Freitas, D.A. Forsyth, Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, Springer Berlin Heidelberg, 2002.
- [11] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of pictures, Nips (2003) 553–560.
- [12] S. Feng, R. Manmatha, A discrete direct retrieval model for image and video retrieval, in: International Conference on Content-Based Image and Video Retrieval, 2008, pp. 427–436.
- [13] V.N. Murthy, S. Maji, R. Manmatha, Automatic image annotation using deep learning representations, in: ACM on International Conference on Multimedia Retrieval, 2015, pp. 603–606.
- [14] V.N. Murthy, E.F. Can, R. Manmatha, A hybrid model for automatic image annotation, in: International Conference on Multimedia Retrieval, 2014, pp. 369–376.
- [15] D.B.M.I. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2) (2009) 1107–1135.
- [16] Y. Verma, C.V. Jawahar, Image annotation using metric learning in semantic neighbourhoods, in: European Conference on Computer Vision, 2012, pp. 836–849.
- [17] Y.V. C. Jawahar, Image annotation by propagating labels from semantic neighbourhoods, Int. J. Comput. Vis. (2017) 1–23.
- [18] S.J. Huang, Z.H. Zhou, Multi-label learning by exploiting label correlations locally, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 949–955.
- [19] Z. Lin, G. Ding, M. Hu, J. Wang, J. Sun, Automatic image annotation using tag-related random search over visual neighbors, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012, pp. 1784–1788.
- [20] Z. Lin, G. Ding, M. Hu, J. Wang, X. Ye, Image tag completion via image-specific and tag-specific linear sparse reconstructions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1618–1625.
- [21] D.G. Lowe, D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
- [22] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
- [23] S.Z. Li, Markov Random Field Modeling in Image Analysis, Springer Science and Business Media, 2009.
- [24] L. Wu, R. Jin, A.K. Jain, Tag completion for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 716.
- [25] Y. Gong, Y. Jia, T. Leung, A. Toshev, S. Ioffe, Deep convolutional ranking for multilabel image annotation, arXiv:1312.4894, (2013).
- [26] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: a unified framework for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2285–2294.
- [27] F. Monay, D. Gatica-Perez, Plsa-based image auto-annotation: constraining the latent space, in: ACM International Conference on Multimedia, 2004, pp. 348–351.
- [28] Y. Xiang, X. Zhou, T.-S. Chua, C.-W. Ngo, A revisit of generative model for automatic image annotation using markov random fields, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1153–1160.
- [29] G. Ciocca, C. Cusano, S. Santini, R. Schettini, Halfway through the semantic gap: prosemantic features for image retrieval, Inf. Sci. 181 (22) (2011) 4943–4958.
- [30] E. Chang, K. Goh, G. Sychay, G. Wu, Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines, IEEE Trans. Circuits Syst. Video Technol. 13 (1) (2003) 26–38.
- [31] D. Grangier, S. Bengio, A discriminative kernel-based approach to rank images from text queries, IEEE Trans. Pattern Anal. Mach. Intell. 30 (8) (2008) 1371–1384.
- [32] I. Dimitrovski, D. Koccev, S. Loskovska, S. Deroski, Hierarchical annotation of medical images, Pattern Recognit. 44 (10–11) (2011) 2436–2449.
- [33] H. Wang, H. Huang, C. Ding, Image annotation using multi-label correlated Green's function, in: IEEE International Conference on Computer Vision, 2009, pp. 2029–2034.
- [34] H. Wang, J. Hu, Multi-label image annotation via maximum consistency, in: IEEE International Conference on Image Processing, 2010, pp. 2337–2340.
- [35] H. Wang, H. Huang, C. Ding, Image annotation using bi-relational graph of images and semantic labels, in: Computer Vision and Pattern Recognition, 2011, pp. 793–800.
- [36] J. Liu, M. Li, Q. Liu, H. Lu, S. Ma, Image annotation via graph learning, Pattern Recognit. 42 (2) (2009) 218–228.
- [37] G. Chen, Y. Song, F. Wang, C. Zhang, Semi-supervised multi-label learning by solving a Sylvester equation, in: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM, 2008, pp. 410–419.
- [38] Z.J. Zha, T. Mei, J. Wang, Z. Wang, X.S. Hua, Graph-based semi-supervised learning with multi-label, in: IEEE International Conference on Multimedia and Expo, 2008, pp. 1321–1324.
- [39] B.K. Bao, T. Li, S. Yan, Hidden-concept driven multilabel image annotation and label ranking, IEEE Trans. Multimedia 14 (1) (2012) 199–210.
- [40] X. Li, B. Shen, B.D. Liu, Y.J. Zhang, A locality sensitive low-rank model for image tag completion, IEEE Trans. Multimedia 18 (3) (2016) 474–483.
- [41] G. Zhu, S. Yan, Y. Ma, Image tag refinement towards low-rank, content-tag prior and error sparsity, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 461–470.
- [42] M.B. Mayhew, B. Chen, K.S. Ni, Assessing semantic information in convolutional neural network representations of images via image annotation, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 2266–2270.
- [43] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu, Q. Zhang, Large scale automatic image annotation based on convolutional neural network, J. Vis. Commun. Image Represent. 49 (2017) 213–224.
- [44] J. Johnson, L. Ballan, F.F. Li, Love thy neighbors: Image annotation by exploiting image metadata, in: IEEE International Conference on Computer Vision, 2015, pp. 4624–4632.
- [45] Y. Yang, W. Zhang, Y. Xie, Image automatic annotation via multi-view deep representation, J. Vis. Commun. Image Represent. 33 (2015) 368–377.
- [46] J. Jin, H. Nakayama, Annotation order matters: recurrent image annotator for arbitrary length image tagging, in: International Conference on Pattern Recognition, 2017, pp. 2452–2457.
- [47] J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3460–3469.
- [48] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, G. Mori, Learning structured inference neural networks with label relations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2960–2968.
- [49] Y. Niu, Z. Lu, J.-R. Wen, T. Xiang, S.-F. Chang, Multi-modal multi-scale deep learning for large-scale image annotation, arXiv:1709.01220 (2017).
- [50] A. Makadia, V. Pavlovic, S. Kumar, A new baseline for image annotation, in: European Conference on Computer Vision, 2008, pp. 316–329.
- [51] L. Wu, S.C.H. Hoi, R. Jin, J. Zhu, N. Yu, Distance metric learning from uncertain side information with application to automated photo tagging, in: International Conference on Multimedia, 2009, pp. 135–144.
- [52] P. Wu, S.C.-H. Hoi, P. Zhao, Y. He, Mining social images with distance metric learning for automated image tagging, in: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 197–206.
- [53] Z. Qin, C.-G. Li, H. Zhang, J. Guo, Improving tag matrix completion for image annotation and retrieval, Vis. Commun. Image Process. (VCIP) (2015) 1–4.
- [54] X. Li, Y.J. Zhang, B. Shen, B.D. Liu, Image tag completion by low-rank factorization with dual reconstruction structure preserved, in: IEEE International Conference on Image Processing, 2014, pp. 3062–3066.
- [55] Y. Hou, Z. Lin, Image tag completion and refinement by subspace clustering and matrix completion, 2015 Vis. Commun. Image Process. (VCIP) (2015) 1–4.
- [56] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2008) 5:1–5:60.
- [57] H.B. Kekre, D. Mishra, A. Kariwala, A survey of CBIR techniques and semantics, Int. J. Eng. Sci. Technol. (2011). 3(5)
- [58] F. Long, H. Zhang, D.D. Feng, Fundamentals of content-based image retrieval, Feng D Multimedia Inf. Retr. Manag. (1) (2003) 1–26.
- [59] R. Datta, J. Li, J.Z. Wang, Content-based image retrieval: approaches and trends

- of the new age, in: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, ACM, 2005, pp. 253–262.
- [60] D. Zhang, M.M. Islam, G. Lu, A Review on Automatic Image Annotation Techniques, Elsevier Science Inc., 2012.
- [61] Y. Liu, D. Zhang, G. Lu, W.Y. Ma, A survey of content-based image retrieval with high-level semantics, *Pattern Recognit.* 40 (1) (2007) 262–282.
- [62] A.-M. Tousch, S. Herbin, J.-Y. Audibert, Semantic hierarchies for image annotation: a survey, *Pattern Recognit.* 45 (1) (2012) 333–345.
- [63] S.A. Manaf, M.J. Nordin, Review on statistical approaches for automatic image annotation, in: International Conference on Electrical Engineering and Informatics, 2009, pp. 56–61.
- [64] V. Dey, Y. Zhang, M. Zhong, A Review on Image Segmentation Techniques with Remote Sensing Perspective, 2010.
- [65] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, *ACM Comput. Surv. (CSUR)* 35 (4) (2003) 399–458.
- [66] M.T. Mills, N.G. Bourbakis, A survey-analysis on natural language understanding methodologies, *IEEE Trans. Syst. Man Cybern.* 44 (1) (2014) 59–71.
- [67] I. RUTHVEN, M. LALMAS, A survey on the use of relevance feedback for information access systems, *Knowl. Eng. Rev.* 18 (2) (2003) 95–145.
- [68] M. Crucianu, M. Ferecatu, N. Boujemaa, Relevance Feedback for Image Retrieval: A Short Survey, Report of the DELOS2 European Network of Excellence (FP6), 2004.
- [69] R. Zhang, L. Zhang, X.J. Wang, L. Guan, Multi-feature pLSA for combining visual features in image annotation, in: International Conference on Multimedia 2011, Scottsdale, Az, Usa, November 28, - December, 2011, pp. 1513–1516.
- [70] D. Putthividhy, H.T. Attias, S.S. Nagarajan, Topic regression multi-modal latent Dirichlet allocation for image annotation, in: Computer Vision and Pattern Recognition, 2010, pp. 3408–3415.
- [71] R. Lienhart, S. Romberg, Multilayer pLSA for multimodal image retrieval, in: ACM International Conference on Image and Video Retrieval, Civr 2009, Santorini Island, Greece, July, 2009, pp. 1–8.
- [72] A. Ghoshal, P. Ircing, S. Khudanpur, Hidden markov models for automatic annotation and content-based retrieval of images and video, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 544–551.
- [73] Y. Zhao, Y. Zhao, Z. Zhu, Tsvm-hmm: transductive svm based hidden markov model for automatic image annotation, *Expert Syst. Appl.* 36 (6) (2009) 9813–9818.
- [74] F. Yu, H.S. Ip, Automatic semantic annotation of images using spatial hidden Markov model, in: IEEE International Conference on Multimedia and Expo, 2006, pp. 305–308.
- [75] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning a Mahalanobis metric from equivalence constraints, *J. Mach. Learn. Res.* 6 (6) (2005) 937–965.
- [76] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, *J. Mach. Learn. Res.* 10 (Feb) (2009) 207–244.
- [77] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighborhood components analysis, in: International Conference on Neural Information Processing Systems, 2004, pp. 513–520.
- [78] E.P. Xing, M.I. Jordan, S.J. Russell, A.Y. Ng, Distance metric learning with application to clustering with side-information, in: Advances in neural information processing systems, 2003, pp. 521–528.
- [79] S.C.H. Hoi, W. Liu, M.R. Lyu, W.Y. Ma, Learning distance metrics with contextual constraints for image retrieval, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 2072–2078.
- [80] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: IEEE International Conference on Computer Vision, 2010, pp. 309–316.
- [81] P. Bakliwal, C. Jawahar, Active Learning Based Image Annotation, in: Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2015 Fifth National Conference on, IEEE, 2015, pp. 1–4.
- [82] F. Tian, X. Shen, Learning label set relevance for search based image annotation, in: International Conference on Virtual Reality and Visualization, 2014, pp. 260–265.
- [83] M.M. Kalayeh, H. Idrees, M. Shah, Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 184–191.
- [84] L. Feng, B. Bhanu, Semantic concept co-occurrence patterns for image annotation and retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (4) (2016) 785.
- [85] X.Y. Jing, F. Wu, Z. Li, R. Hu, D. Zhang, Multi-label dictionary learning for image annotation, *IEEE Trans. Image Process. Publication IEEE Signal Process. Soc.* 25 (6) (2016) 2712–2725.
- [86] B. Wu, S. Lyu, B.G. Hu, Q. Ji, Multi-label learning with missing labels for image annotation and facial action unit recognition, *Pattern Recognit.* 48 (7) (2015) 2279–2289.
- [87] Z. Li, J. Liu, C. Xu, H. Lu, Mlrank: multi-correlation learning to rank for image annotation, *Pattern Recognit.* 46 (10) (2013) 2700–2710.
- [88] M. Ivasic-Kos, M. Pobar, S. Ribaric, Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme, *Pattern Recognit.* 52 (2016) 287–305.
- [89] Z. Lin, G. Ding, M. Hu, Y. Lin, S.S. Ge, Image tag completion via dual-view linear sparse reconstructions, *Comput. Vis. Image Understanding* 124 (2014) 42–60.
- [90] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: lessons learned from the 2015 mscoco image captioning challenge, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4) (2017) 652–663.
- [91] L. Von Ahn, L. Dabbish, Labeling images with a computer game, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2004, pp. 319–326.
- [92] M. Grubinger, Analysis and evaluation of visual information systems performance, Victoria University, 2007 Ph.D. thesis.
- [93] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: ACM International Conference on Image and Video Retrieval, 2009, p. 48.
- [94] M. Chen, A. Zheng, K.Q. Weinberger, Fast image tagging, in: International Conference on International Conference on Machine Learning, 2013, pp. III–1274.
- [95] Z. Ma, F. Nie, Y. Yang, J.R.R. Uijlings, N. Sebe, Web image annotation via subspace-sparsity collaborated feature selection, *IEEE Trans. Multimedia* 14 (4) (2012) 1021–1030.
- [96] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, *Int J Comput Vis* 106 (2) (2014) 210–233.
- [97] Q. Wu, C. Shen, L. Liu, A. Dick, A. van den Hengel, What value do explicit high level concepts have in vision to language problems, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 203–212.
- [98] N. BOURBAKIS, M. MILLS, Converting natural language text sentences into spn representations for associating events, *Int. J. Semant. Comput.* 6 (03) (2012) 353–370.
- [99] N. Bourbakis, Converting diagrams, symbols, formulas, tables and graphics into SPN and NL text sentences for automatic deep understanding of technical documents, in: Int. IEEE Conference on ICTAI, Boston, MA, USA, Nov. 6–8, 2017.

Qimin Cheng obtained her Ph.D. degree from Institute of Remote Sensing Applications, Chinese Academy of Sciences in 2004. She is currently an associate professor of Huazhong University of Science and Technology, Wuhan, China. Her research interests include image retrieval and annotation, Remote sensing images understanding and analysis.

Qian Zhang is currently a M.S degree candidate of Huazhong University of Science and Technology, Wuhan, China. Her research interests include automatic image annotation and image analysis.

Peng Fu is currently a PhD candidate at Indiana State University. He received the B.S degree from Huazhong Agricultural University, Wuhan, China in 2012 and the M.A degree from Indiana State University in 2014. His research focuses on image sequence analysis, data fusion, and environmental modeling.

Conghuan Tu is currently a M.S degree candidate of Huazhong University of Science and Technology, Wuhan, China. Her research interests include content-based image retrieval and salient region extraction.

Sen Li is currently a M.S degree candidate of Huazhong University of Science and Technology, Wuhan, China. His research interests includes image understanding and image analysis etc.