IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2017, accepted August 31, 2017, date of publication September 5, 2017,
date of current version September 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2749252

Automatic Image Annotation Based on Multi-Auxiliary Information

PENGYU ZHANG, ZHIHUA WEI, YUNYI LI, AND CAIRONG ZHAO
Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Corresponding author: Zhihua Wei (zhihua_wei@tongji.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61573259 and in part by the program of Further Accelerating the Development of Chinese Medicine Three Year Action of Shanghai under Grant ZY3-CCCX-3-6002.

ABSTRACT This paper introduces an automatic image annotation framework based on multi-auxiliary information which aims at improving the annotation performance. We propose three novel ideas in the framework of annotation: 1) multi-information extraction: besides various visual features, tag co-occurrence, and user interest vector are added to enrich the multi-auxiliary information; 2) initial labeling: based on the traditional term frequency—inverse document frequency model—we utilize the visibility of words and extended tag set to enhance the result of initial labeling and propose a more efficient model, TF-IDF, visibility and extended tag set model; and 3) tag refinement: by considering multi-auxiliary information, including multi-visual content, tag co-occurrence, and user interest similarity, we propose the multi-information all-labels model for tag refinement. The tag refinement process is formalized as an optimization problem by adjusting confidence score set by the initial labeling model. Experimental results demonstrate that, compared with the state-of-the-art methods, our method achieves the best performance on MIR-Flickr data sets, outperforming the second best by 2%.

INDEX TERMS Automatic image annotation, extended tag set multi-auxiliary information, tag refinement, visibility.

I. INTRODUCTION

We are in a new era of big data where social websites and personal digital devices are all filled with information. Image is one of the most familiar and popular types. Currently, user-provided tags are accessible on Flickr, Picasa and other photo sharing websites. These tags can be used to describe the content of the image and facilitate further applications, such as image retrieval, classification, genome research [33]–[35] and other management tasks. However, these tags provided by users through various kinds of websites are often noisy. These manually annotated tags are frequently irrelevant, imprecise and incomplete. Fig. 1 is an example from the dataset MIR-Flickr 25K that contains a mountain and a lake. User-supplied tags of ''mirror'', ''abigfave'' and ''flickrelite'' are irrelevant to the objects in the image or misspelled. Meanwhile, tags like 'lake' and 'mountain' which can describe the real visual content are missing. Such unsatisfying tag sets must be refined before using in other multimedia applications, or they will lower the performance.

FIGURE 1. An example of social tags.

An image from a photo sharing website consists of three basic elements: image, tags and user. Then six correlations among them can be defined: image-image, image-tag, tag-tag, tag-user, image-user and user-user. Analyzing the relevant score of image-tag correlation is the most important step in image tag refinement [1]. If the relevant score of image-tag relation is larger than the threshold, the tag can be assigned to the image. Otherwise, it will be viewed as a noisy one. The image-user and user-tag correlations define the ownership of an image or a tag. The image-image, tag-tag and user-user correlations are typically characterized by making use of the visual similarity, tag correlation and user interest similarity.

Many efforts have been made to mine the relation information among the three elements. But people often focus on one or two correlations, such as image-tag correlation and image-image correlation. Tag-tag and user-user correlations

are often ignored, but they are also very important during image tag refinement, because this auxiliary information may allow us to achieve better results.

As mentioned above, the image-image correlation is typically characterized by making use of visual similarity, which can be defined by the distance of image visual features. There are many low level visual features, such as wavelet texture [2], color histogram [3], edge direction histogram [4], color moment [5], MPEG-7 edge histogram and MPEG-7 homogeneous texture [6]. Many people simply combine some features as one or only use one feature. The former solution may suffer from the ''curse-of-dimensionality'' problem while the later will lose useful information for image description and discrimination.

In this paper, we propose a novel optimization framework to solve the problems mentioned above. The contribution of the work is summarized as follows.

1) In multi-information extraction, we use several features, instead of simply combining them into a long vector; we assign different features different weights, which can be calculated automatically to evaluate the importance of different features. Moreover, tag co-occurrence and user interest vector are considered so that multi-information can be more abundant.
2) In initial labeling, we propose a new model of initial labeling, TF-IDF, Visibility and Extended Tag Set model (TIVETS, for short), which takes human vision perception into consideration and define an extended tag set to enrich the initialization.
3) By considering multi-auxiliary information including multi-visual content, tag co-occurrence and user interest similarity, we propose the Multi-Information All-Labels model (MIAL, for short) to refine tags. Experimental results demonstrate that our work has better performance than many other methods.

The rest of this paper is organized as follows. In the part of Related Work, we will introduce related work, including multimodal fusion and tag refinement. In the part of Framework, we will describe our framework and its solution in detail. In the part of Experiment, we will introduce experiments, including experimental settings and results. Finally, we will conclude this paper in the part of Conclusion.

II. RELATED WORK

Visual feature selection and extraction are quite important in automatic image annotation. Unlike humans, computers rely on visual features to assign tags to images. There exist quantities of visual features and the most effective feature may vary for different images. For example, for images of sunset and sunrise, color features may perform well. However, for some images of buildings and streets, edge and texture features may be more effective. Therefore, as for image annotation, a simple feature will not meet our needs; instead, feature fusion may be a good solution to this problem. Early and late fusion are the most popular approaches for using various features [1]. Early fusion means extracting several features and concatenating them into a long feature vector. Iyengar *et al.* [8] and Snoek *et al.* [1] accomplished fusion with Support Vector Machine (SVM). Late fusion means integrating the results obtained by different features. A natural method [9] is to replace the high-dimensional learning task by multiple low-dimensional learning tasks, separately applying different features to learning algorithms and then fusing the results. Xia *et al.* [10] proposed a multi-feature fusion method for automatic image annotation by using weighted histogram integral and closure regions counting. However, the two fusion methods have their own shortcomings. As for early fusion, it usually suffers from the problem of ''curse-of-dimensionality''. For late fusion, it may not perform well due to the poor results obtained by each single feature. Moreover, it is quite difficult for us to assign appropriate weights to different features. Fortunately, Wang *et al.* [31] proposed an approach which can automatically assign different features with different weights by evaluating the importance of different features.

Apart from feature fusion, many effective methods were proposed in automatic image annotation. Xu *et al.* [11] proposed an ensemble approach based on Conditional Random Fields. In this method, multiple models are first trained for each tag, then the predictions of these models and the correlations between tags are incorporated into a Conditional Random Field. Deschacht and Moens [12] used salience (the importance of an entity) and visualness (the extent to which an entity can be perceived) to assign image with tags. Shivdikar *et al.* [13] proposed a hybrid engine that uses a combination feature detection algorithms coupled with context free grammar to describe an image in its entirety. Murthy *et al.* [14] made use of Convolutional Neural Network features and word embedding vectors to represent their associated tags.

Image tag refinement is an important step in enhancing the results of initial labeling in automatic image annotation. Many efforts have been devoted to the problem of image tag refinement. Solutions may be classified into two main categories: statistical modeling techniques and data-driven approaches [15]. Wang *et al.* [16] used the algorithm named random walk with restart (RWR) to refine the original annotations of images; the algorithm leverages both corpus information and original confidence score of each candidate annotation. Jia *et al.* [17] proposed a multi-graph similarity reinforcement method; image visual contents were used to explore better correlations. An image retagging framework was proposed by Liu *et al.* [18]; this framework consists of filtering, refinement and enrichment and it shows good results on images from Flickr. A Bayesian network structure was proposed by Zhang and Zhang [19] to efficiently encode the conditional dependencies of the labels as well as the feature sets. Instead of mining tag correlation from co-occurrence or WordNet [20], Xu *et al.* extracted it from a graphical model-rLDA (regularized Latent Dirichlet Allocation). It facilitated topic modeling by exploiting both the statistics of tags and visual features of images [21]. An evaluation of

nearest-neighbor methods for tag refinement was performed by Tiberio Uricchio *et al.* [22]. In [26], the tag refinement problem was formulated as a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix and finally constituted a constrained yet convex optimization problem. The results showed that methods based on the nearest-neighbor algorithm can give comparable results to those of more complex and advanced methods, while being more flexible and less consuming.

Most popular papers focus on one or two correlations, such as image-tag or image-image correlation while some other important correlations are often ignored. In our work, we make full use of multi-auxiliary information in initial labeling and tag refinement. In our previous work [7], we use three kinds of consistencies and various features in tag refinement, but we don't pay much attention to initial labeling before tag refinement. In this paper, we add visibility of words and extended tag set into consideration in initial labeling and propose the TIVETS model. Second, in tag refinement, we propose the MIAL model, the input is the result of TIVETS rather than original image matrix. In this way, the result can be largely improved. Additionally, we show more results and experiments. Experimental results demonstrate that, compared with the state-of-the-art methods, our method achieves the best performance on MIR-Flickr datasets, outperforming the second best by 2%.

III. FRAMEWORK

In this section, we describe our framework. We first give a brief description of the problem, including some notations and definitions we will use later. Then we introduce the extraction of multi-auxiliary information, including three kinds of correlation consistency. In initial labeling, we describe the TF-IDF, Visibility and Extended Tag Set model (TIVETS). Finally, we introduce the Multi-Information All Labels model (MIAL).

A. PROBLEM DESCRIPTION

We define image annotation problem as follows: Given a training image set $I_{train} = \{I_1, I_2, \ldots I_{N\_train}$, a test set $I_{\text{test}} = \{I_1, I_2, \ldots, I_{N\_test}\}$, a tag set $T = \{T_1, T_2, \ldots, T_M\}$, the training image-tag matrix $Y_{train} \in \{0, 1\}^{N_{train} \times M}$ (where $Y_{i,j} = 1$ means tag j is assigned to image I_i) and a list of auxiliary information (such as tag-tag correlation). Then we can get the refined image-tag matrix by refinement function ϕ:

$$Y_{test} = \phi(I_{train}, I_{test}, T, Info, Y_{train}, \varphi) \tag{1}$$

Where φ is set of parameters.

In this paper, the approach we propose can be divided into three parts: multi-auxiliary information extraction, initial labeling and tag refinement. For clarity, we illustrate important notations and definitions used throughout this paper in Table 1.

TABLE 1. Notations and definitions.

Notation	Definition
N, M, K	Number of images, tags and visual features
S, W	The similarity matrix and the corresponding Laplace matrix
X_i^k	The k-th visual feature of image i
$\overline{Y}_{i,j}$	The initial score on tag j for image i
$Y_{i,j}$	Score on tag j for image i after refinement
$\hat{Y}$	Image-tag matrix after refinement, $\hat{Y} = \{0,1\}^{N \times M}$
Y_i	A vector, scores of image i on all the tags after refinement
$(Y^T)_i$	A vector, scores of tag i on all the images after refinement
U_i	A vector, interest scores of user i on all the tags after refinement
$S_{i,j}^k$	The similarity between image i and j on the k-th feature
$S_{i,j}^t$	The similarity between image i and j on the k-th feature
$S_{i,j}^u$	The similarity of user i and j
D^k, D^t, D^u	For regularization, $D_{i,j}^k = \sum_i S_{i,j}^k$, the same with D^t, D^k
L_k^m, L^t, L^u	The Laplace matrix, $L_k^m = D^k - S^k$, the same with L^t, L^u
λ_k	Weight of the k-th visual feature
θ, ξ, α	Parameters

B. MULTI-AUXILIARY INFORMATION EXTRACTION

As we mentioned above, most authors obtain information from only one or two correlations, ignoring some other important information such as tag-tag correlation and user-user correlation. In our framework, three elements (image, tag and user) and three kinds of correlation consistency are considered. Below are the definitions.

1) IMAGE-IMAGE CORRELATION CONSISTENCY

Visually similar images should have similar confidence score on the same label. In our framework, this consistency is defined by calculating similarities between images with different visual features.

We extract K kinds of low level visual features (K modalities) for each image, so there are K visual similarity matrices. The similarity of the i-th and the j-th image on the k-th modality (feature) is defined as follow:

$$S_{i,j}^k = e^{(-||X_i^k - X_j^k||^2/\sigma_k^2)} \tag{2}$$

where σ_k is the median of the Euclidean distance matrix of samples on the k-th modality, X_i^k represents the k-th visual feature of image i.

2) TAG-TAG CORRELATION CONSISTENCY

actually relationships between tags are usually complicated, therefore, the assumption that tags are independent does not hold in traditional annotation methods. Co-occurrence is a kind of relation between tags that means two or more tags often appear in the same image. For example, ''sky'' and ''cloud'', ''car'' and ''road''. Like [16], we define S_{t_i,t_j}^t as the

similarity of tag i and tag j:

$$S^t_{i,j} = \frac{num(t_i, t_j)}{\min(num(t_i), num(t_j))} \tag{3}$$

where $num(t_i, t_j)$ represents the number of results when we search for tag i and j simultaneously; $num(t_i)$ means the number of results when we only search for tag i.

3) USER-USER CORRELATION CONSISTENCY

different users focus on different aspects and have different interests. So the similarity and interests of different users reflected by image-tag matrix after refinement should be consistent with real life. In order to represent users' preference and interests of different images, we define a user interest vector whose dimensionality is the number of tags. in matrix means the total number of images user i assign to tag j over the entire data set. So the user-user similarity can be defined as below:

$$S^u_{i,j} = \exp(-||R \times U_{T_i} - R \times U_{T_j}||^2/\sigma_u^2) \tag{4}$$

where U_T_i is the vector for user i, R is a diagonal matrix and $R_{i,i} = \frac{1}{\sum_{j=1}^{m} U_T(i,j)}$, m represents the dimensionality of U_T matrix, σ_u is the median of the Euclidean distance matrix of different users.

C. INITIAL LABELING

Before introducing the proposed model TIVETS (TF-IDF, Visibility and Extended Tag Set), we will give a brief introduction to TF-IDF.

1) TF-IDF

TF-IDF is a traditional method of initial labeling, whose main idea is that a term is more important and can be the tag of the article if it appears more frequently than other terms.

TF (Term Frequency) describes the frequency of a term appearing in an article. It can be defined as follows.

$$TF_{i,j} = \frac{N_{i,j}}{\sum_k N_{k,j}} \tag{5}$$

where $N_{i,j}$ means the frequency of word W_i appearing in document D_j, $\sum_k N_{k,j}$ means the total appearance of all the words in document D_j.

When the frequencies of two terms in an article are equal, we use a new term, IDF (Inverse Document Frequency), to evaluate the importance of the two terms for the article. It can be defined as follows.

$$IDF_i = \log \frac{|D|}{|\{j : W_i \in D_j\}| + 1}, \tag{6}$$

Where |D| represents the total number of articles, $|\{j : W_i \in D_j\}|$ represents the total number of articles that contains term W_i, we add 1 to $|\{j : W_i \in D_j\}|$ to prevent the denominator from being 0. After calculating TF and IDF, we can get TF-IDF:

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i. \tag{7}$$

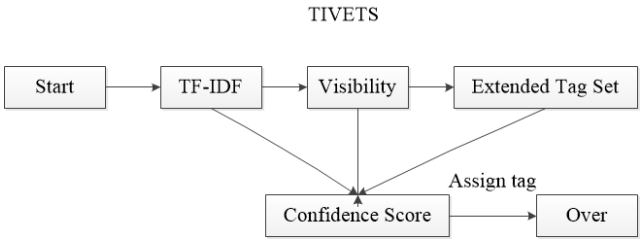

FIGURE 2. Framework of TIVETS.

2) TF-IDF, VISIBILITY AND EXTENDED TAG SET (TIVETS)

In this paper, we propose TIVETS to enhance TF-IDF: we take human vision perception into consideration and define an extended tag set to enrich the initialization. TIVETS can make the result of initial labeling more consistent with human vision perception. Meanwhile, it not only breaks the assumption that tags are independent, but also defines the concept of Extended Tag Set which balances the preference of image provider. Therefore, the coverage and accuracy of initial labeling are greatly improved. The flowchart of TIVETS is described in Fig. 2.

Visibility represents the probability of the tags that can be perceived by human vision in the images. It evaluates the description ability of a word. Visibility has various definitions [12], [23], [24], one [24] of which is the following.

$$vis(w) = \left(\frac{C_1 + 10^{-9}}{C_2 + 10^{-9}}\right)^{-IDF_{Google}(w)} \tag{8}$$

$$IDF_{Google}(w) = \log \frac{|D|}{C_2} \tag{9}$$

where C_1 represents the appearing times of word in Google Image, C_2 represents the appearing times of word in Google Web, |D| represents the total number of websites in Google Web. $\frac{C_1+10^{-9}}{C_2+10^{-9}}$ indicates the probability of the word that appears both in the image and in the website. $IDF_{Google}(w)$ evaluates the importance of the word in Google Web. The higher the visibility of a word, the better its description ability is.

The Extended Tag Set is established by the theory that the assumption of independence between tags does not hold and the relationship between tags is complicated. People will often focus on specific objects but ignore some other abstract concepts, so the user-provided tags are biased. For example, as for an image that contains a cat, people may give tags such as "cat" or the name of it, "Tom", but they will often forget the hypernym, "animal" and other synonyms. However, the integrity of a tag set for an image makes a big difference in text-based image retrieval system and image annotation. Enrichment is a process which adds hypernyms and synonyms to a tag i for an image which enhance the coverage and accuracy of initial labeling. The Extended Tag Set of tag_i consists of tag_i itself, its hypernym set $\Psi(tag_i)$ and its synonym set $\Upsilon(tag_i)$.

$$\Omega(tag_i) = \{tag_i, \Psi(tag_i), \Upsilon(tag_i)\}$$

TIVETS combines TF-IDF with visibility and Extended Tag Set and assign initial tags to each image by calculating the confidence score of all the tags on each image. The confidence score is defined as

$$\overline{Y}_{i,j} = ((TF_{i,j} + \xi) \times IDF_j) \times V_j \times \left(1 - e^{-|\Omega(tag_i)|}\right) \quad (10)$$

where, $\overline{Y}_{i,j}(\overline{Y}_{i,j} \in (0, 1))$ represents the confidence score of tag_j on image i, $(TF_{i,j}+\xi)\times IDF_j$ represents the TF-IDF value of tag_j, V_j represents the visibility of tag_j, and $1 - e^{-|\Omega(tag_i)|}$ means the enriching process of extended tag set. The larger the extended tag set is, the less $e^{-|\Omega\ (tag_i)|}$ will be, therefore, $1 - e^{-|\Omega\ (tag_i)|}$ will be larger. The values of the three terms in the confidence score are between 0 and 1. We multiply them and get the confidence score. The higher the confidence score, the more the tag matches the image. Parameter ξ is a small positive number in case that tag_j does not appear but its hypernym or synonym appears, that is to say, we guarantee that $TF_{i,j}$ will not be 0.

D. TAG REFINEMENT

In order to promote the result of initial labeling, we propose a tag refinement model. We introduce a method named Multi-Information All Labels which integrates multi-visual content, tag co-occurrence and user interest similarity together to refine tags. The details of the model is almost the same as our previous work [7], the main difference is that the input of MIAL algorithm is the result of TIVETS rather than original image matrix.

Many popular methods integrate various visual features to achieve the goal of tag refinement. But they simply use the visual features and ignore the relationship between tags and the important role of user in tag annotation. Therefore, we put forward the model of Multi-Information All Labels. It takes various visual features as well as tag co-occurrence and user interest into consideration and refine the results provided by TIVETS. In this model, the three kinds of consistency should be ensured: Image-image, tag-tag and user-user correlation consistency. The specific detail about these kinds of consistency can be found in Part B of this section, multi-auxiliary information extraction. The framework of MIAL is shown in Fig. 3.

With the constraints above, we can get an optimization function:

$$\begin{aligned} \min J(Y, \lambda; \theta, \xi) = & \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j}\lambda_k S_{i,j}^k ||Y_i - Y_j||^2 \\ & + \frac{1}{2}\sum_{i,j} S_{i,j}^t ||(Y^T)_i - (Y^T)_j||^2 + \frac{1}{2}\sum_{i,j} S_{i,j}^u ||U_i - U_j||^2 \\ & + \theta \sum_i ||Y_i - \overline{Y}_i||^2 + \xi||\lambda||^2 \\ s.t. & \sum_{k=1}^{K}\lambda_k = 1, \quad 0 \le \lambda_k \le 1,\ k = 1, 2, ...K. \end{aligned} \quad (11)$$

The function consists of three parts. The first part contains three terms: the first term means that visually similar images should be annotated with the same tag. The second term means the relation between similar tags should be guaranteed. The third term ensures that the similarity and interests of different users reflected by image-tag matrix after refinement should be consistent with real life. The second part means the results obtained by initial labeling are reliable to some extent, the results after refinement should not change too much. The third part is used to prevent over-fitting.

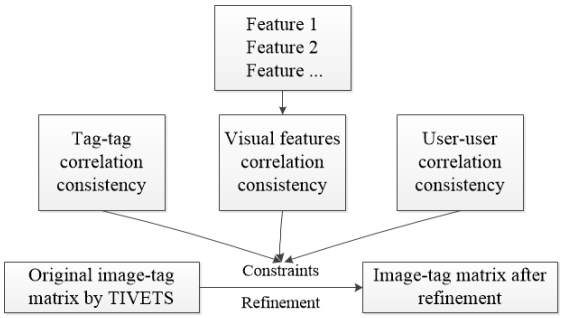

FIGURE 3. Framework of MIAL.

In (11), U_i is the score vector of user i, and $U = R \times G \times Y$, where R is a diagonal matrix, matrix G can be defined as (12), $G(i, j)$ will be 1 on condition that image j belongs to user i.

$$G(i,j) = \begin{cases} 1, & I_j \in \mathrm{u}_i \\ 0, & else \end{cases} \quad (12)$$

As for a symmetric matrix, $\sum_{i,j} H_{i,j}||Y_i - Y_j||^2 = 2Tr(Y^T H_L Y)$, H_L is the corresponding Laplace matrix of matrix H. Therefore, we can simplify J:

$$\begin{aligned} J = & \sum_{k=1}^{K}\lambda_k Tr\left(Y^T L_k^m Y\right) + Tr\left(YL^t Y^T\right) \\ & + Tr\left((RGY)^T L^u (RGY)\right) + \theta Tr\left(\left(Y - \overline{Y}\right)\left(Y - \overline{Y}\right)^T\right) \\ & + \xi||\lambda||^2 \\ s.t.\ & 0 \le \lambda_k \le 1, \quad k = 1, 2, ...K, \sum_{k=1}^{K}\lambda_k = 1 \end{aligned} \quad (13)$$

Here we give an iterative algorithm to solve the optimization problem.

Fix λ_k to Solve Y:

$$\begin{aligned} \frac{\partial J}{\partial Y} = & \sum_{k=1}^{K}\lambda_k\left(Y^T L_k^m + Y^T\left(L_k^m\right)^T\right)^T + \left(YL^t + Y(L^t)^T\right) \\ & + \left(Y^T Q + Y^T Q^T\right)^T + \theta\left(\left(Y - \overline{Y}\right) + \left(Y - \overline{Y}\right)\right) \\ \frac{\partial J}{\partial Y} = 0 \Rightarrow & \left(\sum_{k=1}^{K}(\lambda_k L_k^m) + Q + \theta I\right) Y + YL^t - \theta\overline{Y} = 0 \end{aligned} \quad (14)$$

TABLE 2. Algorithm of MIAL.

Algorithm: Multi-Information All Labels
Input: the training images and corresponding tags.
Step 1: initialization. Extract various visual features.
Step 2: initial labeling, calculate the initial confidence score and similarity matrix, including S^k, S^t, S^u and initial loss J^0.
Step 3: fix λ_k to solve Y. With Y, $\hat{Y}_{i,j} = \begin{cases} 1, if Y_{i,j} > 0 \\ 0, else \end{cases}$
Step 4: fix Y to solve λ_k. Utilize the active-set algorithm to solve the nonlinear quadratic programming.
Step 5: Calculate J^t, if $J^t < J^{t-1}$, go to step 3, otherwise quit iteration.
Output: the image-tag matrix after refinement, $\hat{Y}_{i,j}$.

where $Q = G^T R^T L^u RG$ is a symmetric matrix, and we make equations as

$$\begin{cases} A = \sum_{k=1}^{K} (\lambda_k L_k^m) + Q + \theta I \\ B = L^t \\ C = \theta \overline{Y} \end{cases} \Rightarrow AY + YB = C \qquad (15)$$

This is a Sylvester equation, MATLAB provides a function to solve it.

Fix Y to Solve λ_k: When Y is fixed, the original optimization problem can be transformed into a nonlinear quadratic programming problem:

$$\min Z = \sum_{k=1}^{K} Tr\left(Y^T L_k^m Y\right) \lambda_k + \xi \|\lambda\|^2$$
$$s.t. \sum_{k=1}^{K} \lambda_k = 1, \quad 0 \le \lambda_k \le 1, \ k = 1, 2, ...K. \qquad (16)$$

There are many tools and methods for nonlinear quadratic programming; the active-set algorithm, for instance, can solve this problem. Table 2 shows the algorithm of MIAL.

In our previous work, we took three kinds of consistency into consideration in tag refinement, but we didn't pay much attention to initial labeling before tag refinement. Therefore, the input of the algorithm will be the original matrix. But in this paper, we pay much attention to initial labeling. First and foremost, we utilize TF-IDF model to calculate the possibility that a tag belongs to the image. In order to prevent the value of TF to be 0, we use the deformation formula of TF-IDF and add a small positive number. Moreover, we use the visibility of words to evaluate the description ability of a tag. Words with higher visibility are more likely to form visual images in the human brain than those with low visibility. Thus, visibility can be used to represent a strong or weak association between words and images. Thirdly, different from the previous work, we get the extended tag set by considering the tag itself and its hypernym and synonym. This will be helpful to make the annotation results more precise and abundant. Then, we combine the value of TF-IDF, visibility and extended tag set to calculate the confidence scores of tag on the images. With this initial labeling matrix, we use it as the input matrix of refinement process instead of original matrix in previous work. Last but not least, we do much more comparison experiments to test the efficiency and validity of our algorithm.

IV. EXPERIMENTS

In this section, we first introduce our experimental settings, and then we present the experimental results that validate the effectiveness of our approach. The experimental result contains two parts. In the first part, we compared the results obtained by the proposed TIVETS model with the traditional method, TF-IDF. In the second part, we compared our MIAL model with other popular methods in tag refinement.

A. EXPERIMENTAL SETTINGS

1) DATASET

To empirically evaluate our proposed approach, we conducted experiments on the data sets from the photo-sharing websites Flickr and MIR-Flickr 25K [25], which contains 25000 images and 1386 unique social labels. The data set derives from a photo sharing website, therefore, it contains various kinds of images, an example set is shown in Fig. 4.

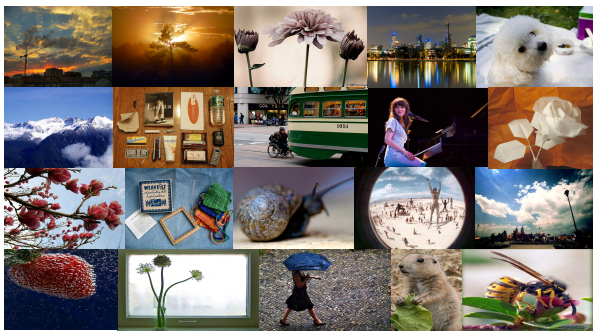

FIGURE 4. Images of MIR-Flickr 25K.

Learning from [26], taking misspellings and meaningless tags into consideration, we deleted infrequent tags (less than 50) and those entries that cannot be matched in Wikipedia. Finally we obtained a tag set with 194 tags.

2) VISUAL FEATURE EXTRACTION

We extracted six kinds of visual features from the images: 1) MPEG-7 edge histogram; 2) MPEG-7 homogeneous texture; 3) 5-by-5 block-wise color moment; 4) HSV color histogram; 5) Wavelet-texture; 6) Edge direction histogram.

3) EVALUATION METRIC

In this paper, we used the F-score macro as our evaluation metric. We first computed the F-score on each of the tags and

then calculated the average score.

$$F = \frac{2PR}{P+R} \tag{17}$$

$$F - macro = \frac{1}{M}\sum_{1}^{M} F_i \tag{18}$$

where P represents the precision rate and R represents the recall rate.

4) EXPERIMENTAL ENVIRONMENT

We conducted our experiments in Windows 10(64bit), with MatlabR2013a, g++ complier.

B. INITIAL LABELING

1) VERIFICATION OF VISIBILITY

With equation (8) and (9), we calculated the visibility of some tags which are commonly used, the result shows as Table 3.

TABLE 3. Visibility of different tags.

Tag	Visibility	Tag	Visibility
fog	0.964	ocean	0.353
clouds	0.838	snow	0.307
grass	0.735	wood	0.227
sunrise	0.698	animal	0.219
mountains	0.687	spring	0.200
landscape	0.496	winter	0.181
sunset	0.475	art	0.053
leaves	0.471	size	0.051
boat	0.429	type	0.041
trees	0.426	love	0.027

Table 3 shows that, for tags ''fog'', ''clouds'' and ''grass'', their visibilities are higher than the visibilities of tags such as ''size'', ''type'' and ''love''. The reason for this result is that these former tags are more specific and they can be perceived more easily by human brains than the latter ones. Therefore, the visibility of a tag can be used as an effective metric in initial labeling.

2) OVERALL PERFORMANCE

We conducted our experiments on MIR-Flickr 25K with TIVETS and TF-IDF. We used average F based on images (mF_I) and average F based on tags (mF_T). We assigned a tag to an image only if the confidence score is higher than the threshold. Here we set $\xi = 0.2$. Fig. 4 shows the overall results of TIVETS and TF-IDF.

Fig. 5 demonstrates that TIVETS performs much better than TF-IDF from both mF_I and mF_T. For either mF_I or mF_T, the overall result of the TIVETS is twice that of TF-IDF. Therefore, the coverage of TIVETS is much better.

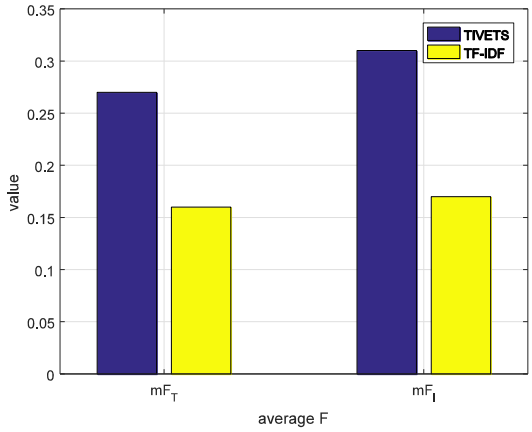

FIGURE 5. Overall results of TIVETS and TF-IDF.

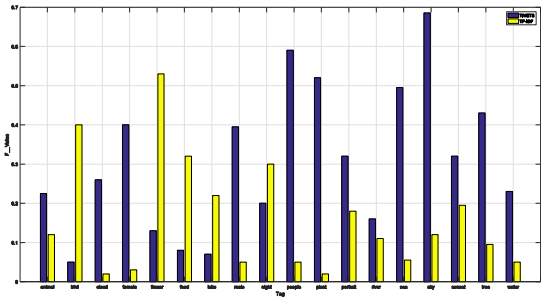

FIGURE 6. Comparison of TIVETS and TF-IDF on different tags.

3) PERFORMANCE ON DIFFERENT TAGS

Then we compared the performance based on each of tags, the results are shown in Fig. 6.

From Fig. 6, we can see that for most tags, the results of TIVETS is much better than TF-IDF. But for some special tags, such as ''flower'', flowers have large quantities of hyponyms, ''Rose'', for instance. However, ''Rose'' can also be a name of a person. In this situation, TIVETS may assign ''Rose'' to tag ''person'' instead of tag ''flower''. The reason that accounts for this situation is that the number of images that belong to different categories is unbalanced. A solution to this problem can be enlarging the dataset. On the whole, the result is still promising because for most situations, TIVETS is more robust than TF-IDF.

4) SENSITIVITY TO PARAMETER

We also tested the sensitivity to parameter ξ, Fig. 7 demonstrates that, with different thresholds, the higher the threshold is, the lower the F-score of TIVETS. But we can see that the F-score of TIVETS is always higher than that of TF-IDF according to both mF_I and mF_T. Fig. 8 is an example of comparison of TF-IDF and TIVETS. We can see that besides accuracy rate, the coverage rate of TIVETS is much better.

C. TAG REFINEMENT

In this part, we firstly verified the validity of multi-auxiliary information and various visual features. Then we compared

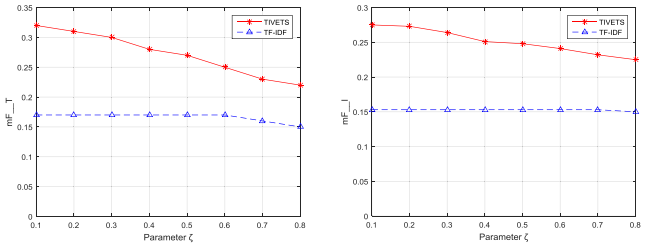

FIGURE 7. Comparison of TIVETS and TF-IDF with different threshold.

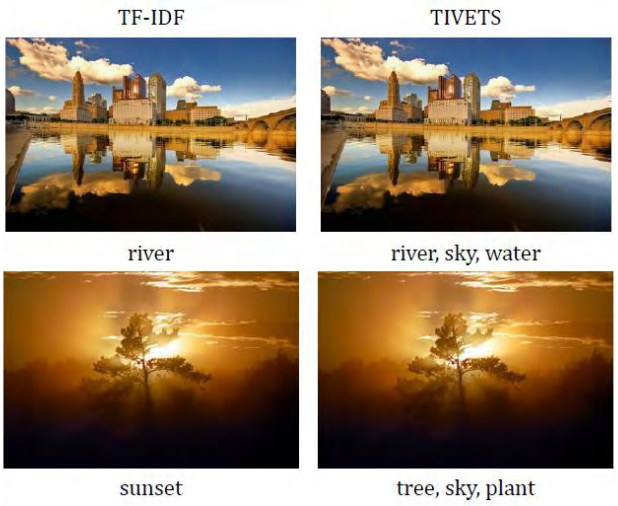

FIGURE 8. Comparison based on actual examples.

TABLE 4. F-macro with different kinds of consistency.

Consistency	TIVETS (baseline)	C	C+T	**C+T+U**
F-macro	0.30	0.39	0.43	**0.44**

the results of our proposed models with those of other popular models from both overall results and each kind of tags.

1) ADVANTAGE OF MULTI-CORRELATION CONSISTENCY

In the proposed MIAL model, three kinds of correlation consistency are taken into consideration in tag refinement. They are consistency between visual content and semantic correlation (C), tag-tag correlation consistency (T) and consistency between user interests (U). In comparison, we took the results of TIVETS as the baseline.

Table 4 shows that multi-auxiliary information in tag refinement can largely enhance the results obtained by TIVETS. Moreover, the more multi-auxiliary information or the more consistency we consider, the better refinement performance we will get.

2) ADVANTAGE OF FEATURE FUSION

For individual feature, we averaged the results of five-by-five Block-Wise Color Moment, Edge Direction Histogram and Wavelet Texture. For connected feature vector, we connected six kinds of visual features mentioned above. From Table 5, it is obvious that multi-feature fusion achieves much better performance than the other two methods.

TABLE 5. Refinement result with different feature fusion methods.

Feature	Individual Feature	Connected Feature	Feature Fusion
F-macro	0.31	0.35	0.44

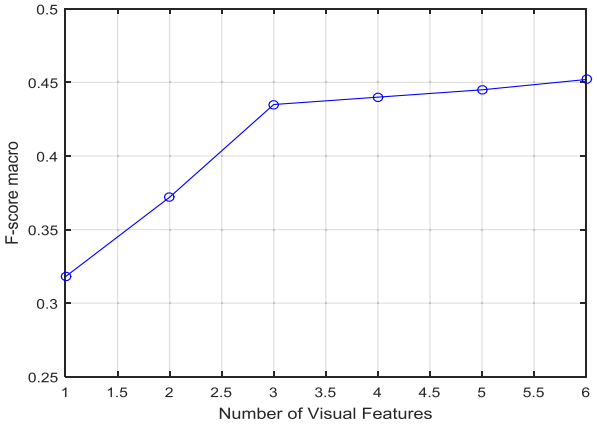

FIGURE 9. Refinement result with different number of features.

TABLE 6. Weights of different features.

Feature	MPEG-7 edge histogram	MPEG-7 homogeneous texture	Wavelet-texture
Weight	0.21	0.12	0
Feature	**HSV color histogram**	**Edge direction histogram**	**Color moment**
Weight	0.15	0.29	0.23

For multi-feature fusion, we also tested whether the number of features will influence the result. The result is shown in Fig. 9.

From Fig. 9, we can see that the performance gets better when increasing the number of visual features. Therefore, multi-feature fusion is efficient for tag refinement.

3) COMPARISON OF DIFFERENT REFINEMENT METHODS

In this part, we conducted our experiments on dataset MIR-Flickr 25K and we compared our proposed approach with several popular methods of tag refinement. On the parameter of MIAL, we set $\theta = 0.8$, $\xi = 250$, and when algorithm terminates, the weights of visual features shows as Table 6.

In Table 7, we compared the result of our models with those of RWR, TRVSC and LR. The results of initial labeling, TIVETS, can be regarded as the baseline.

We can see that the proposed MIAL method can enhance the result of TIVETS and outperforms the best state-of-the-art

TABLE 7. Comparison of different methods.

Method	TIVETS	RWR[15]	TRVSC[17]	LR[26]	**MIAL**
F-macro	0.30	0.34	0.41	0.42	**0.44**

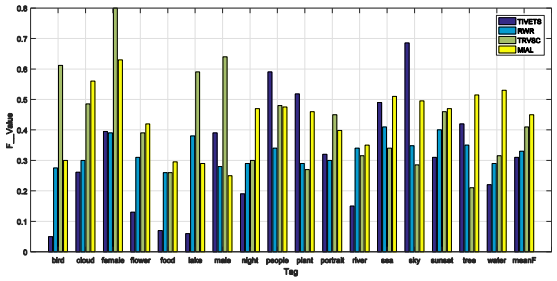

FIGURE 10. Comparison of refinement result on different tags.

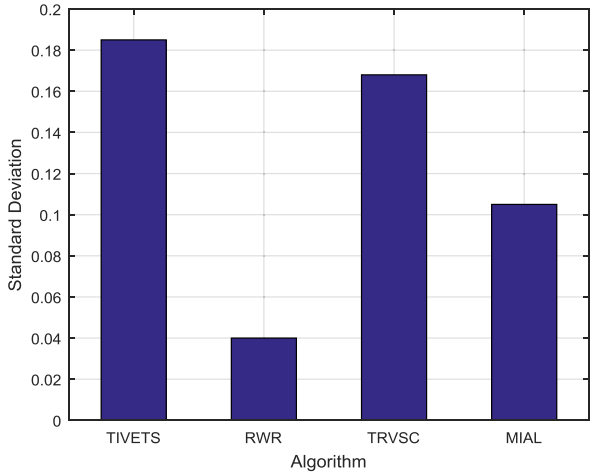

FIGURE 11. Standard deviation.

by 2%. Compared with our previous work [7], with TIVETS, the MIAL model outperforms the previous algorithm by 3%. Fig. 10 shows the F-macro of different methods on different tags. We can see that, in contrast with initial labeling, other refinement methods improve the result to different extent. The proposed model MIAL achieves excellent results almost on all tags.

4) STABILITY OF DIFFERENT METHODS

We calculated the standard deviation of different algorithms on various tags in order to test the stability of different algorithms.

Fig. 11 shows that RWR is the most stable algorithm compared to TRVSC, MFLW and MIAL. In all the tag refinement methods, TRVSC has the largest standard deviation and performs well in tag ''bird'', ''female'', ''lake'' and ''male''. However, when it comes to tag ''animal'' and ''tree'', the F-macro is pretty low, even lower than that of TIVETS. The standard deviation of our proposed MIAL models is between them.

Meanwhile, as for initial labeling, TIVETS, we can see that the standard deviation is quite high. The reason for this phenomenon may be the extended tag set. For those tags, with large extended tags set, the result of initial labeling may be good. Similarly, for those which have a small extended tag set, the result may be much poorer than those of excellent ones. Therefore, the standard of TIVETS is large and it is not stable.

V. CONCLUSION

We introduced an automatic image annotation framework consisting of multiple auxiliary information, initial labeling and tag refinement. We took multiple information and several kinds of correlation consistency into consideration. Moreover, in initial labeling, we proposed TIVETS, which enhanced TF-IDF model by considering the visibility of words and extended tag set. For tag refinement, by considering multiple auxiliary information including multi-visual content, tag co-occurrence and user interest similarity, we proposed the Multi-Information All-Label (MIAL) model. To test the performance of the proposed approach, we conducted experiments on dataset MIR-Flickr 25K. The effectiveness of TIVETS and MIAL was demonstrated. TIVETS has higher accuracy and coverage than traditional methods. MIAL takes multi-feature fusion, tag-tag correlation consistency and user interests into consideration, which shows more robustness than many other popular refinement methods. Through comparing with several existing refinement methods, the results demonstrate the superiority of our approach and outperforms the second best by 2%.

Future work will focus on the feature-tag correlation consistency and feature extraction. In our paper, we considered image-image, tag-tag and user-user correlation consistency. Actually, the relation between tags and features is also complicated. Some features may achieve a better result for specific tags. For example, the color feature may perform better than some other features for tag ''cloud'', while some texture features may obtain better results for tag ''street''. Second, the features we use now are artificial ones, and we always select features according to our preference which may influence the tag annotation. Feature extraction using deep learning [27]–[30] may be more suitable for image annotation.

REFERENCES

[1] C. G. Snoek, M. Worring, and A. W. Smeulders, ''Early versus late fusion in semantic video analysis,'' in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.
[2] J. Z. Wang, J. Li, and G. Wiederhold, ''SIMPLIcity: Semantics-sensitive integrated matching for picture libraries,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947–963, Sep. 2001.
[3] A. R. Smith, ''Color gamut transform pairs,'' in *Proc. 5th Annu. Conf. Comput. Graph. Interact. Techn.*, vol. 12. 1978, pp. 12–19.
[4] X. Gao, B. Xiao, D. Tao, and X. Li, ''Image categorization: Graph edit distance+edge direction histogram,'' *Pattern Recognit.*, vol. 41, no. 10, pp. 3179–3191, 2008.
[5] D. Zhang, M. M. Islam, and G. Lu, ''A review on automatic image annotation techniques,'' *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, 2012.

[6] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. Hoboken, NJ, USA: Wiley, 2002.
[7] Y. Li, D. Miao, and Z. Wei, "Social tagrefinement model based on feature fusion and multi-correlation consistency," *J. Nanjing Univ.*, vol. 52, no. 2, pp. 244–252, 2016.
[8] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 255–258.
[9] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. ACM Int. Conf. Multimedia*, 2004, pp. 572–579.
[10] S. Xia, P. Chen, J. Zhang, X. P. Li, and B. Wang, "A multi-feature fusion method for automatic multi-label image annotation with weighted histogram integral and closure regions counting," in *Proc. Int. Conf. Intell. Comput.*, 2015, pp. 323–330.
[11] X. S. Xu, Y. Jiang, L. Peng, X. Xue, and Z.-H. Zhou, "Ensemble approach based on conditional random field for multi-label image and video annotation," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1377–1380.
[12] K. Deschacht and M. F. Moens, "Text analysis for automatic image annotation," in *Proc. 45th Annu. Meet. Assoc. Comput. Linguistics*, 2007, pp. 1000–1007.
[13] K. Shivdikar, A. Kak, and K. Marwah, "Automatic image annotation using a hybrid engine," in *Proc. IEEE INDICON*, Dec. 2015, pp. 1–6.
[14] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 603–606.
[15] D. Liu, X. S. Hua, and H. J. Zhang, "Content-based tag processing for Internet social images," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 723–738, 2011.
[16] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 647–650.
[17] J. Jia, N. Yu, X. Rui, and M. Li, "Multi-graph similarity reinforcement for image annotation refinement," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 993–996.
[18] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proc. Int. Conf. Multimedia*, 2010, pp. 491–500.
[19] M. L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, ACM*, 2010, pp. 999–1008.
[20] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[21] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Tag refinement by regularized LDA," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 573–576.
[22] T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo, "An evaluation of nearest-neighbor methods for tag refinement," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.
[23] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, "A text-to-picture synthesis system for augmenting communication," in *Proc. AAAI*, vol. 7. 2007, pp. 1590–1595.
[24] F. Wu, Y.-H. Han, Y.-T. Zhuang, and J. Shao, "Clustering Web images by correlation mining of imagetext," *J. Softw.*, vol. 21, no. 7, pp. 1561–1575, 2010.
[25] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[26] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. Int. Conf. Multimedia*, 2010, pp. 461–470.
[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[29] D. Ciresan, U. Meier, and J. Schmidhuber. (Feb. 2012). "Multi-column deep neural networks for image classification." [Online]. Available: https://arxiv.org/abs/1202.2745
[30] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. (Feb. 2011). "High-performance neural networks for visual object classification." [Online]. Available: https://arxiv.org/abs/1102.0183
[31] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "multimodal graph-based reranking for Web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2013.

PENGYU ZHANG received the bachelor's degree from the Department of Computer Science and Technology, Tongji University, China, in 2015, where he is currently pursuing the master's degree. His research interests include machine learning and image processing.

ZHIHUA WEI received the Ph.D. degree from the Department of Computer Science and Technology, Tongji University, China, in 2010. She is currently an Associate Professor with Tongji University. Her research interests include machine learning, image processing, and natural language processing.

YUNYI LI received the master's degree from the Department of Computer Science and Technology, Tongji University, China, in 2015. Her research interests include image processing and natural language processing.

CAIRONG ZHAO received the B.S. degree from Jilin University, the M.S. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, and the Ph.D. degree from the Nanjing University of Science and Technology in 2011, 2006, and 2003, respectively. He is currently an Associate Professor with Tongji University. He has authored over 30 scientific papers in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.