Effective Semantic Annotation by Image-to-Concept Distribution Model

Ja-Hwung Su, Chien-Li Chou, Ching-Yung Lin, and Vincent S. Tseng, *Member, IEEE*

***Abstract*—Image annotation based on visual features has been a difficult problem due to the diverse associations that exist between visual features and human concepts. In this paper, we propose a novel approach called Annotation by Image-to-Concept Distribution Model (AICDM) for image annotation by discovering the associations between visual features and human concepts from image-to-concept distribution. Through the proposed image-to-concept distribution model, visual features and concepts can be bridged to achieve high-quality image annotation. In this paper, we propose to use "visual features", "models", and "visual genes" which represent analogous functions to the biological chromosome, DNA, and gene. Based on the proposed models using entropy, tf-idf, rules, and SVM, the goal of high-quality image annotation can be achieved effectively. Our empirical evaluation results reveal that the AICDM method can effectively alleviate the problem of visual-to-concept diversity and achieve better annotation results than many existing state-of-the-art approaches in terms of precision and recall.**

***Index Terms*—Entropy, image annotation, image-to-concept distribution, tf-idf.**

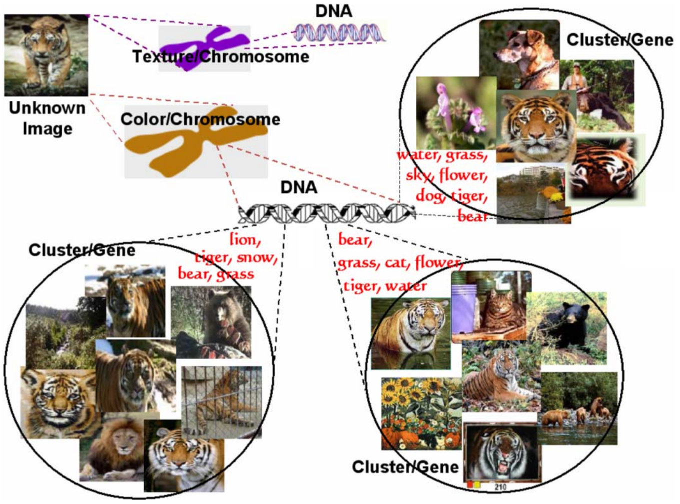
DNA
Texture/Chromosome
Unknown Image
Color/Chromosome
DNA
Cluster/Gene
water, grass, sky, flower, dog, tiger, bear
Cluster/Gene
lion, tiger, snow, bear, grass
bear, grass, cat, flower, tiger, water
Cluster/Gene

Fig. 1. Basic idea of the proposed AICDM.

I. Introduction

ADVANCED digital capturing technologies have led to the explosive growth of image data. To retrieve the desired images from a huge amount of image data, textual query is handier to represent her/his interest than providing visually similar images for query. Most existing successful textual-based image retrieval relies heavily on the related image caption terms, e.g., file-names, categories, annotated keywords, and other manual descriptions. To caption the images effectively, in the last decade, extensive image understanding techniques have been developed to explore semantic concept of images. But, due to the significant diversity of a large amount of image data in daily life, effective image annotation is still a very challenging and open problem. Diverse visual feature versus concept associations indicate that the same visual feature is frequently shared by a set of concepts. The challenge is that the related terms are so diverse that the annotator could not annotate the unknown image accurately. In existing annotation work, this problem, namely the diverse visual-to-concept associations, occurs so frequently that many annotation results are not satisfactory for human users.

To address this problem, in this paper, we propose novel visual-to-concept distribution models that integrate the methods of entropy, $tf - idf$ and association rules to enhance the annotation quality. In molecular biology, genes locating on different chromosomes have similar functions due to the high similarity of their DNA sequences. This is useful for predicting the specific function for a gene. Based on this notion, the purpose of this paper is to annotate the image by discovering the representative and discriminative visual features alike the genes hidden in the images. Fig. 1 is an example for the basic idea of our proposed AICDM. In Fig. 1, we consider each image has a specific number of features to be extracted—similar to chromosome. On each visual feature, a set of models can be applied to divide image collections into several different cluster sets. A specific model of a visual chromosome/feature can be considered as a DNA. For each DNA, a "visual gene" is the visual pattern of a cluster, which includes a set of visually-correlated images of a model on a visual feature and a set of caption terms associated with them. Similar to the biological gene, each "visual gene" carries the information, i.e., the caption terms that have been learned from training corpus. Our intent behind this idea is to identify the conceptual distinctness of each gene. According to the discriminative genes, the image-to-concept distribution model is constructed. For an unknown test image, we shall then

Manuscript received November 03, 2010; revised February 28, 2011; accepted March 08, 2011. Date of publication March 17, 2011; date of current version May 18, 2011. This work was supported by National Science Council, Taiwan, R.O.C. under grants NSC99-2631-H-006-002 and NSC99-2218-E-006-001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhu Liu.

J.-H. Su and V. S. Tseng are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan (e-mail: bb0820@ms22.hinet.net; tsengsm@mail.ncku.edu.tw).

C.-L. Chou is with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan (e-mail: fallwinds@gmail.com).

C.-Y. Lin is with IBM T. J. Watson Research Center, Hawthorne, NY 10532 USA (e-mail: chingyung@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2129502

classify them to find out the likelihood that they are composed of a certain gene. Then, the concept terms of genes are associated to this unknown images based on these genes.

In this paper, we propose four types of models, which can be classified into three categories: 1) from viewpoint of individual images, we adopt tf (term frequency) and entropy to weight the concept terms and genes, respectively; 2) from viewpoint of image sets, we adopt association-rule confidence and idf (inverse document frequency) to weight the concept terms and genes; and 3) by integrating tf, idf, entropy and association rules, we apply late fusion using support vector machines (SVM) [2] to achieve high-quality image annotation. The empirical evaluations on several image sets reveal that our proposed Annotation by Image-to-Concept Distribution Model (AICDM) is very promising on semantic annotation by measuring precision and recall of the annotation accuracy, comparing to several existing algorithms. The remaining of this paper is organized as follows. Several prior works are reviewed in Section II. In Section III, we present the proposed method in detail. The related experimental evaluations are described in Section IV. Finally, the conclusion and future work are stated in Section V.

II. Related Work

In general, image annotation work can be categorized into several types.

Classification-Based Annotation: The first type is the classification-based annotation. In the past, some studies treated annotation as classification using multiple classifiers. Yang *et al.* [24] proposed a region-based annotation method by using SVM. This study presented an extended SVM namely asymmetrical SVM to infer the caption terms of images. Nasierding *et al.* [9] adopted multi-classifier to achieve image annotation by integrating clustering and classification methods. Similarly, Wu *et al.* [18] optimized the bag-of-words to preserve semantic of images. In addition, Bayesian classifier was built to annotate images by integrating regional and global features [10]. Fan *et al.* [5] proposed a structured max-margin learning algorithm to conduct effective inter-related classifiers to support image annotation.

Probabilistic-Based Annotation: The second type is the probabilistic-based annotation. Probabilistic models are constructed by estimating the correlations between images and concepts. Li *et al.* [7] computed the relational probabilities between images and concepts by multi-statistical models, e.g., 2-D Hidden Markov Model, Gaussian, and Gamma distributions. Lavrenko *et al.* [6] calculated the related probabilities between segmentations and concepts by Gaussian Mixture Function. Pan *et al.* [11] developed Mixed Media Graph (MMG) model to annotate the image by Cross-modal Correlation Discovery (CCD) algorithm to calculate the affinities of caption terms and regions. Tang *et al.* [14] proposed the multi-graph-based label propagation approach that integrates multiple instance learning and single instance learning to tag the unknown image.

Retrieval-Based Annotation: The third type is the retrieval-based annotation. The basic notion behind retrieval-based annotation is that semantic-relevant images are composed of similar visual features. Wang *et al.* [22] proposed the AnnoSearch system to bridge the semantic gap by Search Result Clustering (SRC) [25]. Wang *et al.* [23] annotated an image by both of visual and textual search. By using social images with tags and user-generated content, Wu *et al.* [17] presented a retrieval-based method to tag images effectively.

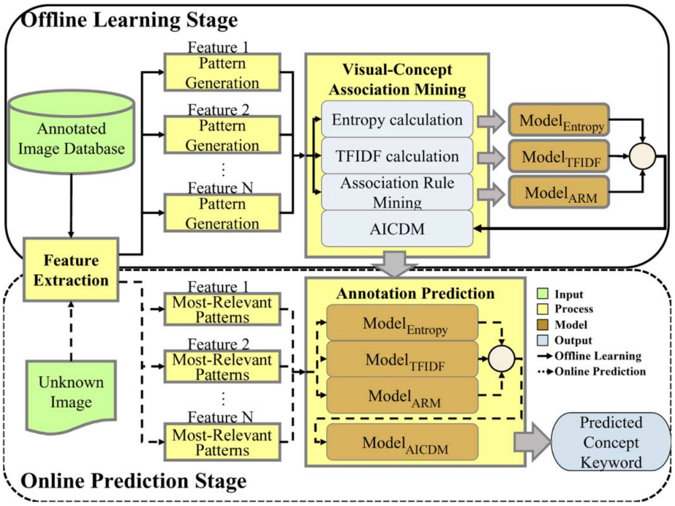

Fig. 2. Framework of the proposed AICDM.

In addition to the above three types of annotation methods that are mainly based on the content modeling, the other type is to use more textual information to enhance the annotation quality [15]. Wong *et al.* [20] made use of the additional metadata, such as aperture, exposure time, subject distance, focal length, and fire activation, to tag the images. Tseng *et al.* [16] integrated decision tree and MMG based on textual and visual information to annotate the web images. Wu *et al.* [19] proposed Flickr distance to achieve effective image annotation. In fact, the Flickr distance work and our proposed method have different advantages. The proposed method in this paper is an extended work of [12]. For the Flickr distance work, it is effective on resolving the problem of concept appearance variation by using spatial information [21]. For our proposed method, by discovering the representative and discriminative patterns, it is effective to alleviate the annotation problem that a feature may occur frequently in many concepts.

III. Proposed Approach

A. Overview of the Proposed Image Annotator

The so-called diverse visual-to-concept associations indicate that similar visual features may frequently occur in different concepts. From another point of view, it says that a semantic concept contains different visual features. In real applications, image annotators encounter difficulties in these diverse associations between visual features and human concepts. To address this problem, in this paper, we propose novel visual-to-concept distribution models that integrate the methods of entropy, $tf-idf$ and association rules to achieve high annotation quality. As shown in Fig. 2, the whole procedure can be decomposed into two stages, namely *offline learning* and *online prediction*.

1) Offline Learning Stage: Overall this stage contains three main sub-procedures, called *feature extraction*, *pattern generation*, and *model construction*.

- **Feature Extraction**: In this paper, six visual features are extracted from the images, including Scalable Color Descriptor, Color Layout Descriptor, Homogeneous Texture

Descriptor, Edge Histogram Descriptor, Grid Color Moment, and Gabor Wavelet Moment, whose dimensionalities are 256, 12, 62, 80, 225, and 72, respectively.

- **Pattern Generation**: After feature extraction, the annotated images are grouped into a set of visual clusters feature by feature. A cluster can be regarded as a representative and discriminative gene hidden in the training images. In other words, images can be described by six visual features.
- **Model Construction**: From the generated patterns, term frequency and inverse document frequency $(tf - idf)$ and cluster entropy are calculated to construct $\mathrm{Model}_{\mathrm{tf-idf}}$ and $\mathrm{Model}_{\mathrm{entropy}}$, respectively. Also, association rules are mined to generate $\mathrm{Model}_{\mathrm{ARM}}$ (Association Rule Mining). Finally, three individual models are integrated into a fusion model, $\mathrm{Model}_{\mathrm{AICDM}}$, by SVM [2].

2) Online Prediction Stage: In this stage, the major aim is to identify the concepts of an unknown image using the proposed models. First, for an unknown image, the most-relevant clusters/patterns are determined feature by feature. Through the most-relevant patterns, potential caption terms can be predicted by the modeled relations between visual features and semantic concepts.

B. Offline Learning

1) Pattern Generation: Before constructing the proposed models, the annotated images are grouped by calculating visual distances. In this work, images are clustered by the well-known k-means algorithm. Thereupon we can obtain a set of clusters, also called patterns or genes in this paper, for each visual feature. A cluster contains a set of images and an image is annotated by a set of keywords. Let us take Fig. 3 as an example. Assume that the images are grouped into five clusters $\{C_1, C_2, C_3, C_4, C_5\}$ by Scalable Color Descriptor, and each image is projected as a set of keywords. In Fig. 3, a box stands for a set of concept terms related to an image. An issue of concern in this work is the quality of clustering since it actually makes a significant impact on the quality of online prediction. To make the clustering quality robust, we perform the validation methods proposed by [13]. After clustering, for each feature, the images are grouped into a set of clusters. Each cluster is viewed as a pattern/gene. For example, the related pattern set for scalable color descriptors is $\{C_1, C_2, \ldots, C_5\}$.

2) Basic Idea: From the generated clusters, we can observe that images in a cluster are very similar on the visual features but containing a number of different concepts. This is a big problem called diverse visual-to-concept associations to confound current annotators. From the bioinformatics point of view, two images may be similar if they share similar visual patterns that are considered as visual genes in this work. Unfortunately, a visual gene perhaps contains lots of concepts. It relates to three important issues: 1) How important is a caption term in a gene, 2) How important is a gene among all visual genes, and 3) How associative is a term-gene pair. To answer these questions, we propose a novel solution that integrates caption term frequency, gene entropy, and association visual-to-concept rule to achieve the high retrieval quality of image annotation.

3) Construction of $Model_{entropy}$: As elaborated above, our intention is to identify the importance of a caption term and a pattern by calculating caption term frequency and pattern entropy. The major idea is that, the higher the frequency of a caption term, the more representative it is. In contrast, if a large number of caption terms occur in a cluster, the related entropy would be too high to disambiguate visual concepts. That is, entropy can be viewed as a local weight for a gene. According to this notion, two related measures [12], called Term Frequency (tf) and Entropy, are defined as follows.

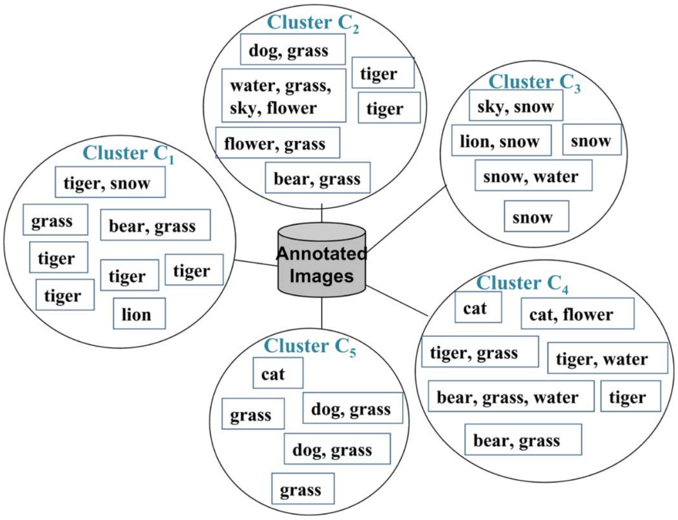

Fig. 3. Example of visual clusters containing the related concepts.

Definition 1: Consider a training data set IDB $= \{img_1, img_2, \ldots, img_k\}$ is divided into t clusters, $\{C_1, C_2, \ldots, C_t\}$, and there are y unique caption terms $\{cp_1, cp_2, \ldots, cp_y\}$. Assume that a cluster contains several images and each image is assigned several caption terms. Hence a cluster can be viewed as a collection of caption terms, $C_j = \cup cp_i$. The entropy of the j^{th} cluster/gene can be defined as

$$Entropy^j = \sum_{cp_i \in C_j} \left[\frac{tf^j_{cp_i}}{\sum_{cp_v \in C_j} tf^j_{cp_v}} \log \left(\frac{\sum_{cp_v \in C_j} tf^j_{cp_v}}{tf^j_{cp_i}} \right) \right] \tag{1}$$

where $tf^j_{cp_i}$ stands for the frequency of caption term cp_i in the jth cluster. For example, based on Fig. 3, the frequency set of {grass, dog, cat} for C_5 is {4, 2, 1} and $Entropy^5$ is $4/7^* \log(7/4) + 2/7^* \log(7/2) + 1/7^* \log(7/1) = 0.415$. Thus, the entropy set is $\{Entropy^1, Entropy^2, Entropy^3, Entropy^4, Entropy^5\} = \{0.59, 0.778, 0.466, 0.755, 0.415\}$.

4) Construction of $Model_{tf-idf}$: Similar to the above model, tf is also generated first. Another way to determine the discrimination of a gene is inverse document frequency, namely idf. In this paper, the idf of the jth cluster/gene can be defined as the following [12].

Definition 2: Following the above definitions, the "inverse document frequency (idf)" for the jth cluster/pattern is

$$idf^j = \log \left(\frac{y}{\sum_{1 \le g \le y} df^j_{cp_g}} \right) \tag{2}$$

TABLE I
DEFINITIONS OF VISUAL PATTERN SETS

Feature	Pattern Set
Scalable Color Descriptor	KC={C_1, C_2, ..., C_{kc}}
Color Layout Descriptor	KL={L_1, L_2, ..., L_{kl}}
Homogeneous Texture Descriptor	KH={H_1, H_2, ..., H_{kh}}
Edge Histogram Descriptor	KE={E_1, E_2, ..., E_{ke}}
Grid Color Moment	KG={G_1, G_2, ..., G_{kg}}
Gabor Wavelet Moment	KW={W_1, W_2, ..., W_{kw}}

where

$$df^j_{cp_g} = \begin{cases} 1, & if\ C_j\ contains\ caption\ cp_g \\ 0, & otherwise. \end{cases}$$

In this model, if the pattern/gene contains most of unique caption terms, it would be a general pattern/gene. Therefore, its discrimination with respect to *idf* is low. Let us take an example based on Fig. 3. The set of unique caption terms in this example is {tiger, grass, bear, lion, snow, sky, flower, water, cat, dog}. For cluster C_1, it contains the caption term set {tiger, grass, bear, lion, snow}. The *idf* of C_1 is $\log(10/5) = 0.301$. For cluster C_2, the *idf* of C_2 is $\log(10/7) = 0.155$. In this case, C_1 is more discriminative than C_2. Overall *idf* can be viewed as the global weight for a gene. Thus, the final *idf* set for $\{C_1, C_2, C_3, C_4, C_5\}$ is {0.301, 0.155, 0.398, 0.222, 0.523}.

5) Construction of $Model_{ARM}$: In summary, the above two models are constructed feature by feature. That is, regarding Table I, six entropy models and six $tf - idf$ models are generated. In contrast to the above models, the main concern of $\mathrm{Model}_{\mathrm{ARM}}$ is to discover the associations between visual features and concept keywords by considering all features simultaneously. Before mining the associations, it is necessary to define the items. To fit association mining, a pattern or a concept keyword is regarded as an item and an image perhaps contains a set of keywords (caption terms). Furthermore, in this model, a transaction divided can be defined as follows.

Definition 3: Based on the definitions in Definition 1 and Table I, the ith transaction T_i for the kth image img_k is

$$T_i = \langle \{C_{t_1}, L_{t_2}, H_{t_3}, E_{t_4}, G_{t_5}, W_{t_6}\}, \{cp_g\}\rangle \tag{3}$$

where

$$C_{t_1} \in KC,\ \ L_{t_2} \in KL,\ \ H_{t_3} \in KH,\ \ E_{t_4} \in KE,\ \ G_{t_5} \in KG,\ \ W_{t_6} \in KW$$

and cp_g is one of the concept keywords related to img_k.

For example, assume image_k contains two keywords {tiger, grass}. Thus, the referred transactions are $\langle\{C_2, L_1, H_4, E_5, G_1, W_2\}, \{\mathrm{tiger}\}\rangle$ and $\langle\{C_2, L_1, H_4, E_5, G_1, W_2\}, \{\mathrm{tiger}\}\rangle$ where the set of patterns for img_k is $\langle\{C_2, L_1, H_4, E_5, G_1, W_2\}, \{\mathrm{grass}\}\rangle$. Finally, the annotated image database can be transformed into a transaction database. After database transformation, discovering the association rules from transaction data is our next intention in this work. From the transactions, a rule and related confidence can be defined in the following.

Definition 4: Consider that there are n rules in the rule set $\{\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_\mathrm{n}\}$ mined from the transaction database. Thus, a rule in the rule set can be defined as

$$R_i : Feature\ Patterns \rightarrow Concept\ Keyword. \tag{4}$$

The confidence of rule $\mathrm{R_i}$ can be defined as

$$Confidence(R_i) = \frac{Sup(Feature\ Patterns \cup Concept\ Keyword)}{Sup(FeaturePatterns)} \tag{5}$$

where the $Sup(\mathrm{itemset})$ is the normalized frequency of the itemset in transaction database. For example, the rule $\mathrm{R}_1\ :\ \{C_2, E_5, G_1, W_2\} \rightarrow \{\mathrm{tiger}\}$ indicates that the image whose features can be assigned to the pattern set $\{C_2, E_5, G_1, W_2\}$ always contains a concept keyword, "tiger". The confidence of rule R can be calculated by (5):

$$Confidence(R_1) = \frac{Sup\left(\{C_2, E_5, G_1, W_2, \mathrm{tiger}\}\right)}{Sup\left(\{C_2, E_5, G_1, W_2\}\right)}$$

where $Sup(\{C_2, E_5, G_1, W_2, \mathrm{tiger}\})$ indicates the count of itemset $\{C_2, E_5, G_1, W_2, \mathrm{tiger}\}$ and $Sup(\{C_2, E_5, G_1, W_2\})$ indicates the count of itemset $\{C_2, E_5, G_1, W_2\}$.

In addition to the confidence value of a rule, another considerable factor τ_{R_i} named *pattern-concept count* is how many concept keywords are implied by the same Feature Patterns set. The basic idea is that, if a feature itemset is shared with lots of concept keywords, the discrimination of the rule is relatively low. From another viewpoint, if lots of rules contain the same feature itemset, the related weights are low. Thus, for each rule, we count the number of implied keywords. For example, suppose that three rules

$$\begin{aligned} &\mathrm{R}_1 : \{C_2, E_5, G_1, W_2\} \rightarrow \{\mathrm{tiger}\}, \\ &\mathrm{R}_2 : \{C_2, E_5, G_1, W_2\} \rightarrow \{\mathrm{grass}\}\ \mathrm{and} \\ &\mathrm{R}_3 : \{C_1, E_3, G_5, W_1\} \rightarrow \{\mathrm{tiger}\} \end{aligned}$$

are mined from the transaction database. The set $\{C_2, E_5, G_1, W_2\}$ implies two keywords, "tiger" and "grass". Thus, the pattern-concept counts τ_{R_1} and τ_{R_2} are both 2. Comparatively, the set $\{C_1, E_3, G_5, W_1\}$ implies only one keyword, "tiger". Therefore, the related pattern-concept count τ_{R_3} is 1. From discrimination point of view, R_3 is better than R_1 and R_2. At last, the rules, confidences, and pattern-concept counts are all stored into rule database.

C. Online Prediction for Annotation

The prediction procedure starts when an unknown image QI is submitted to this system. First, for each feature, the most-relevant clusters are determined by calculating visual similarities/distances. Assume that the most-relevant cluster set CS to QI is determined by visual distance calculations. The visual similarity (visual distance) between the unknown image and the jth cluster is defined as dis^j in this paper. Actually, the most-relevant clusters can be regarded as a kind of genes for the unknown image. Once the genes of the unknown image are determined, three prediction models are triggered to predict the

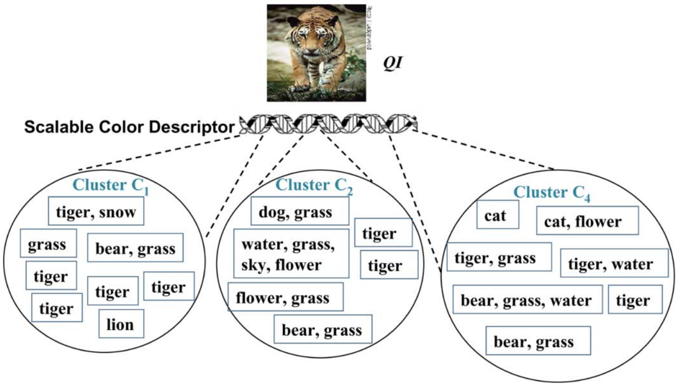

Fig. 4. Example of three relevant clusters to the unknown image.

potential caption terms. In our current system, the number of the most-relevant clusters is determined by experimental evaluations based on training set.

1) Prediction by $Model_{entropy}$: In this model, our intent is to weight caption terms by using gene entropy and caption term frequency. The major notion is that, if the caption term occurs frequently in a gene with low entropy, its referred degree would be high. Otherwise, its degree would be low. Finally, the caption terms are ranked by the related degrees. In this paper, the entropy-based degree [12] is defined as

$$EDegree_{cp_x} = \sum_{C_j \in CS} \left[\left(\frac{tf^j_{cp_x}}{\sum_{cp_v \in C_j} tf^j_{cp_v}} \right) \times \frac{1}{Entropy^j} \times \left(\frac{\left(\sum_{1 \leq k \leq |IDB|} dis^k \right)}{dis^j} \right) \right] \quad (6)$$

where $1 \leqq x \leqq y$. Afterwards the EDegree is normalized. The normalization is defined as

$$NormEDegree_{cp_x} = \frac{EDegree_{cp_x}}{\sum_{i=1}^{y} EDegree_{cp_i}}. \quad (7)$$

The EDegrees referred to the selected six features are aggregated and normalized as the final degrees for a caption term. Let us take an example based on Figs. 3 and 4. Assume that the most-relevant cluster/gene set for an unknown image QI is $\{C_1, C_2, C_4\}$. Accordingly, these three most-relevant clusters contain 21 images and ten unique caption terms. If the referred distance set for $\{dis^1, dis^2, dis^4\}$ is {150, 500, 250} and the sum of distances for $\{dis^1, dis^2, dis^3, dis^4, dis^5\}$ is 2700, the normalized distance set is {18, 5.4, 10.8}. Therefore, the entropy set of $\{C_1, C_2, C_4\}$ is {0.59, 0.778, 0.755}. Thus, the EDegree for {tiger} is ((5/10)*(1/0.59)*(18)) + ((2/12)*(1/0.778)*(5.4))+((3/13)*(1/0.755)*(10.8)) = 19.714. Finally, the entropy-based degree set for caption term set {tiger, grass, bear, lion, snow, sky, flower, water, cat, dog} is {19.714, 11.717, 5.831, 3.051, 3.051, 0.578, 2.257, 2.78, 2.202, 0.578}. In this example, the correct caption term set {tiger, grass} regarding Fig. 1 is successfully inferred from top 2 results.

2) Prediction by $Model_{tf-idf}$: To weight caption terms by considering the global weight, we adopt idf to reveal the degrees of the caption terms in the most-relevant genes. The major notion behind this model is that, if the caption term occurs frequently in a gene with high idf, its related degree would be high. Otherwise, its degree would be low. At last, the caption terms are ranked by the related degrees. We define the idf-based degree [12] as

$$FDegree_{cp_x} = \sum_{C_j \in CS} \left[\left(\frac{tf^j_{cp_x}}{|C_j|} \right) \times idf^j \times \left(\frac{\left(\sum_{1 \leq k \leq |IDB|} dis^k \right)}{dis^j} \right) \right] \quad (8)$$

where $1 \leqq x \leqq y$. Afterwards the EDegree is normalized. The normalization is defined as

$$NormFDegree_{cp_x} = \frac{FDegree_{cp_x}}{\sum_{i=1}^{y} FDegree_{cp_i}}. \quad (9)$$

At last, the FDegrees referred to the selected six features are aggregated and normalized as the final degrees for a caption term. For example, based on above examples, for caption term "tiger", the occurring cluster set is $\{C_1, C_2, C_4\}$ and the related FDegree is (5/5*0.301*2700/150) + (2/7*0.155*2700/500)+(3/6*0.222*2700/250) = 6.854. The final FDegree set for {tiger, grass, bear, lion, snow, sky, flower, water, cat, dog} is {6.854, 3.839, 1.998, 1.08, 1.08, 0.119, 0.637, 0.918, 0.799, 0.119}. Therefore, the correct caption term set {tiger, grass} is successfully derived by this model.

3) Prediction by Fusion $Model_{ARM}$: In addition to the above degrees, the confidences of rules are adopted to reveal the degrees of the concept keywords in the most-relevant pattern. The major notion behind this prediction is that, if a concept keyword occurs in lots of rules, it is a general caption term in the global feature space. As a result, its related degree is high. In this prediction, the six patterns for an unknown image QI are first determined for six selected features, respectively. Then the matched association rules for QI are found. Based on Definition 4, the matched rule set RQI can be defined as $RQI = \cup\{R_i^{QI}\}$, where R_i^{QI} denotes the matched rule for QI. Then the length of rule R_i^{QI}, defined as $len(R_i^{QI})$, is $|FeaturePatterns|$. For example, the $len(R_i^{QI})$ of rule $\{C_2, E_5, G_1, W_2\} \rightarrow \{\text{tiger}\}$ is 4 because there are four items in the left-hand side of the rule. In this work, we have to find the maximum matching rules. That is, if $len(R_i^{QI})$ is maximum among all matching rules, the rule R_i^{QI} is added into the longest rule set $LRS(RQI)$. Moreover, the related sub-rule sets, which are the combinations of a feature pattern and a caption term, are chosen. For example, there is a maximum matching rule R_1^{QI}: $\{C_2, E_5, G_1, W_2\} \rightarrow \{\text{tiger}\}$, and the related sub-rule set $sub(R_1^{QI})$ is $\{\{C_2\} \rightarrow \{\text{tiger}\}, \{E_5\} \rightarrow \{\text{tiger}\}, \{G_1\} \rightarrow \{\text{tiger}\}, \{W_2\} \rightarrow \{\text{tiger}\}\}$. After determining the matching rules and the related sub-rules, the degrees of concept keywords are calculated by the pattern-concept counts of these rules. Finally, the concept keywords are ranked

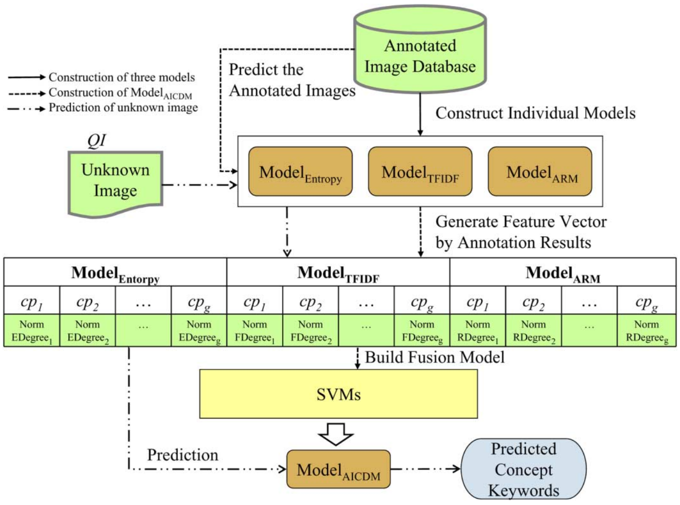
Construction of three models
Construction of Model$_{\text{AICDM}}$
Prediction of unknown image
Annotated Image Database
Predict the Annotated Images
Construct Individual Models
QI
Unknown Image
Model$_{\text{Entropy}}$
Model$_{\text{TFIDF}}$
Model$_{\text{ARM}}$
Generate Feature Vector by Annotation Results
Model$_{\text{Entorpy}}$
Model$_{\text{TFIDF}}$
Model$_{\text{ARM}}$
cp_1 cp_2 … cp_g
Norm EDegree$_1$ Norm EDegree$_2$ … Norm EDegree$_g$
Norm FDegree$_1$ Norm FDegree$_2$ … Norm FDegree$_g$
Norm RDegree$_1$ Norm RDegree$_2$ … Norm RDegree$_g$
Build Fusion Model
SVMs
Prediction
Model$_{\text{AICDM}}$
Predicted Concept Keywords

Fig. 5. Procedure of fusing $\text{Model}_{\text{entropy}}$, $\text{Model}_{\text{tf}-\text{idf}}$, and $\text{Model}_{\text{ARM}}$.

by the related degrees. The rule-based degree can be defined as (10) at the bottom of the page, where $Lhs(R_i^{QI})$ is the ancestor of R_i^{QI} and $Rhs(R_i^{QI})$ is the descendant of R_i^{QI}, and $Sup(Lhs(R_i^{QI}) \cup Rhs(R_i^{QI}))$ is the support count of itemset "$Lhs(R_i^{QI}) \cup Rhs(R_i^{QI})$". Then, the RDegrees are normalized further. The normalization is defined as

$$NormRDegree_{cp_x} = \frac{RDegree_{cp_x}}{\sum\limits_{i=1}^{y} RDegree_{cp_i}}. \tag{11}$$

For example, there are three longest matching rules:

$$R_1^{QI} : \{C_2, E_5, G_1, D_2\} \rightarrow \{\text{tiger}\},$$
$$R_2^{QI} : \{C_2, E_5, G_1, D_2\} \rightarrow \{\text{grass}\} \text{ and}$$
$$R_3^{QI} : \{C_2, L_3, H_5, D_2\} \rightarrow \{\text{tiger}\}.$$

The related confidence set and the support set are {0.7, 0.35, 0.25} and {10, 20, 8}, respectively. The confidence sets of $sub(R_1^{QI})$, $sub(R_2^{QI})$, and $sub(R_3^{QI})$ are {0.2, 0.4, 0.5, 0.3}, {0.3, 0.2, 0.15, 0.15}, and {0.15, 0.2, 0.1, 0.15}. According to the descriptions mentioned in above sections, the pattern-concept counts, τ_{R_1}, τ_{R_2}, and τ_{R_3}, can be calculated as 2, 2, and 1, respectively. Then, we calculate the RDegree for each concept keyword. For concept keyword "tiger", the related RDegree is $(1/2)*10*(0.7 + 0.2 + 0.4 + 0.5 + 0.3) + (1/1)*8*(0.25 + 0.15 + 0.2 + 0.1 + 0.15) = 17.3$. For concept keyword "grass", the related RDegree is $(1/2)*20*(0.35 + 0.3 + 0.2 + 0.15 + 0.15) = 11.5$.

4) Prediction by Fusion Model $Model_{AICDM}$: To achieve better annotation quality, we approximate a near-optimal fusion model. Fig. 5 reveals the procedure of constructing $\text{Model}_{\text{AICDM}}$. The annotated images are first used as the learning set to construct $\text{Model}_{\text{entropy}}$, $\text{Model}_{\text{tf}-\text{idf}}$, and $\text{Model}_{\text{ARM}}$. Meanwhile, the annotation results of the annotated images are generated by the three above models, respectively. Eventually, the derived annotation results and related concept degrees are used as feature vectors to build the fusion model, $\text{Model}_{\text{AICDM}}$, by utilizing SVM [2]. For each concept keyword, we build a SVM, with respect to radial basis function (RBF) kernel function, to perform the binary classification. The number of dimensions for each model is the number of keyword categories, and total number of dimensions for SVM is triple the number of keyword categories. The whole procedure shown in Fig. 5 starts with an unknown image QI submitted to our proposed annotator. The related EDegree, FDegree, and Rdegree for each concept keyword are derived by the individual prediction models first. Then, the EDegrees, FDegrees, and RDegrees regarded as the feature vectors of the unknown image QI are sent to the SVMs in $\text{Model}_{\text{AICDM}}$. Thereupon the classification confidence of each concept keyword is derived. At last, the concept keywords are ranked by the related classification confidences.

IV. Empirical Evaluations

A. Experimental Data and Parameter Settings

To make the experiments complete, the experimental data came from the collections of WebImage, PascalVOC07 (Pascal Visual Object Classes Challenge 2007) [3], and ESP [1]. For WebImage, the experimental data is a collection of ten categories gathered from Google, including Bear, Cat, Dog, Lion, Tiger, Flower, Grass, Sky, Snow, and Water. Each category contains 100 unique web images occurring in 100 different web pages. On average, an image contains 1.574 caption terms in this dataset. We select 50% of experimental data as the training set and the others are adopted to serve the testing experiments. For PascalVOC07, it contains 9963 images. We adopt 5011 images as the training set and 4952 images are adopted as the testing set. There are 20 unique concepts in this dataset and an image, on average, contains 1.71 caption terms. For ESP, the set we obtained contains 67 769 images. However, we removed the images with infrequent annotations and then split the set into a training set and a testing set according to [8]. Finally, there are 269 concepts left in this set. The training set contains 18 689 images and the testing set contains 2081 images. Overall there are 269 unique caption terms and the average of caption terms for an image is 4.7. To investigate the effectiveness of our proposed models, three measures, namely precision, recall, and F_1-measure, are used in the experiments. Note that the definitions of precision and recall [16] here are different from that in PascalVOC07 [3]. In this work, the number of the clusters is approximated for each

$$RDegree_{cp_x} = \sum_{R_i^{QI} \in LRS(RQI)} \left[\frac{1}{\tau_{R_i}} \times Sup\left(Lhs\left(R_i^{QI}\right) \cup Rhs\left(R_i^{QI}\right)\right) \times \sum_{R_j \in \left(sub\left(R_i^{QI}\right) \cup R_i^{QI}\right)} Confidence(R_j) \right] \tag{10}$$

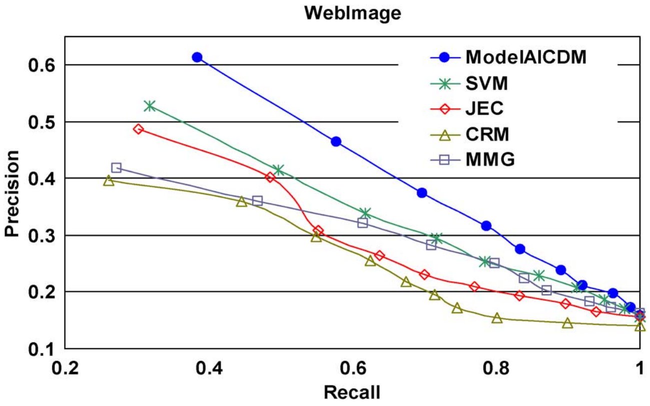

Fig. 6. Precision-recall curves of the proposed and other annotators on WebImage.

dataset and each model by golden search algorithm. Additionally, in the experiments, the numbers of the most-relevant clusters of $\mathrm{Model}_{\mathrm{entropy}}$, $\mathrm{Model}_{\mathrm{tf-idf}}$, and $\mathrm{Model}_{\mathrm{ARM}}$ are 2, 2, and 1, respectively, for all datasets. The experiments were carried out under hardware environment of Intel(R) Xeon(R) E3113 CPU 3.00 GHz, 4 GB memory with Windows Server 2003 R2 SP2 operating system.

B. Experimental Results

The main experiments we explore are the comparisons between our proposed AICDM and other well-known annotators, including CRM [6], SVM [4], MMG [11], and JEC [8], in terms of precision, recall, and execution time. We made our best effort to implement those algorithms based on their papers and got similar performance if their datasets are available. Basically, CRM and MMG are probabilistic-based approaches using image segmentation. Without image segmentation, SVM is classification-based approach and JEC is KNN-based approach. In this experiment, the area under curve (AUC) is the additional measure.

Fig. 6 reveals some interesting results to discuss in detail. First, SVM performs better than CRM, MMG, and JEC for WebImage dataset, and the related AUCs are 0.3934, 0.2925, 0.3321, and 0.3498, respectively. Second, JEC is better than MMG, and MMG is better than CRM. It says that the segmentation-based annotation models are not really better than the models without segmentation. Third, our proposed $\mathrm{Model}_{\mathrm{AICDM}}$ is the best one, and the related AUC is 0.4706. It tells us the truth that the special genes in images can be identified effectively to imply the visual-concept associations. Fig. 7 reveals that the precision-recall curves on PascalVOC07 dataset. In this dataset, CRM and MMG fail to execute because the required memory size is out of the resource. From the remaining three approaches, we can observe that SVM $(\mathrm{AUC} = 0.1912)$ does not work well in this dataset due to the higher diversities of images and concepts. In contrast, JEC $(\mathrm{AUC} = 0.2571)$ can still keep the good performance through KNN strategy. Compared with above methods, our proposed $\mathrm{Model}_{\mathrm{AICDM}}$ $(\mathrm{AUC} = 0.2892)$ can achieve the highest effectiveness for this dataset.

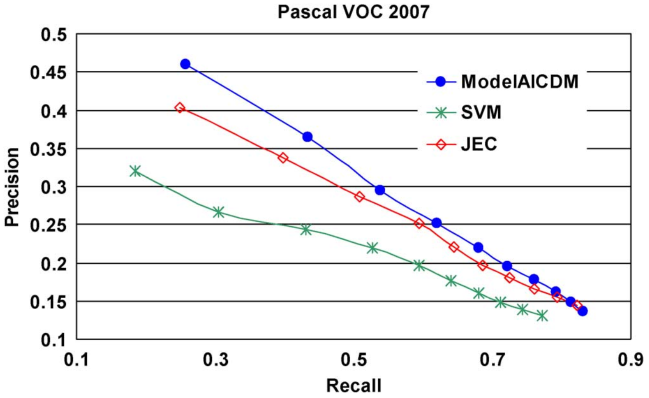

Fig. 7. Precision-recall curves of the proposed and other annotators on PascalVOC07.

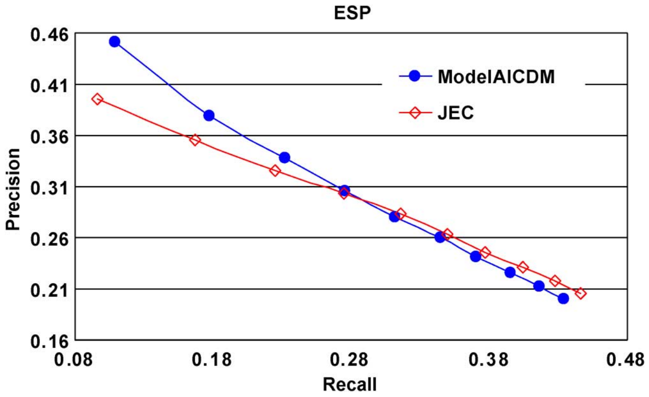

Fig. 8. Precision-recall curves of the proposed and other annotators on ESP.

TABLE II
PERFORMANCE AMONG COMPARED METHODS

Dataset Method	WebImage	PascalVOC07	ESP
AICDM	0.059 sec.	0.421 sec.	0.632 sec.
CRM	0.112 sec.	*out of memory*	*out of memory*
MMG	0.153 sec.	*out of memory*	*out of memory*
SVM	0.031 sec.	0.109 sec.	*out of training time*
JEC	0.076 sec.	0.477 sec.	4.8495 sec.

Fig. 8 reveals the comparisons among different approaches on ESP dataset. In this dataset, CRM and MMG also fail to execute because the required memory is out of the resource. In addition, the training cost of SVM is too large, exceeding one month, probably caused by a large number of outliner image features and keywords. Therefore, we only compare $\mathrm{Model}_{\mathrm{AICDM}}$ with JEC. In this experiment, the AUCs of $\mathrm{Model}_{\mathrm{AICDM}}$ and JEC are 0.1512 and 0.1440, respectively. In detail, JEC performs slightly better as the recall is larger than 0.28. However, on average, our proposed $\mathrm{Model}_{\mathrm{AICDM}}$ is much better than JEC in terms of AUC. For each dataset, $\mathrm{Model}_{\mathrm{AICDM}}$ outperforms other well-known annotation approaches in terms of precision, recall, and AUC. That is, from the viewpoint of dataset sensitivity, SVM is highly sensitive to the dataset distribution. In contrast, JEC is more stable than SVM. From all experimental results, we can observe that our proposed $\mathrm{Model}_{\mathrm{AICDM}}$ is insensitive for different datasets.

In addition to the effectiveness, another issue is how efficient the proposed model is by comparing with other annotators. Table II depicts the execution time of each annotator for predicting an image, and there are some observations to discuss.

First, it shows that our approach, AICDM, is very efficient for generating real-time annotation results. Second, JEC is efficient for small dataset. However, the execution time increases explosively as the training data size increases. For ESP dataset, JEC needs about 4.85 s such that it is not suitable for real applications. Third, SVM is the most efficient, but it does not provide the adequate annotation accuracy.

V. Conclusion

Indeed, an optimal solution to achieve high accuracy annotator is very difficult. This paper constitutes a novel approach to discover the visual-to-concept associations from the image-to-concept distribution. The experimental results show that our proposed annotation approach is effective and efficient in facing data consisting of the diverse relations between visual features and human concepts. On one hand, entropy and tf reflect the local weights of patterns. On the other hand, idf and association rules reflect the global weights of patterns. By making use of both the global and local weights, the fusion model can successfully achieve high annotation quality. In the future, there remain some issues for further investigation. First, we shall explore more visual features to enhance the annotation quality. Second, the spatial information will be a further consideration to enhance our proposed method. Third, we shall further investigate the better fusion methods to reach higher annotation quality. Furthermore, in the future, we shall also explore the proposed algorithms to domains other than multimedia.

References

[1] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Human Factors on Computer Systems*, 2004, pp. 319–326.

[2] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Classes Challenge 2007 (VOC2007) Results, 2009.

[4] H. Feng, R. Shi, and T. S. Chua, "A bootstrapping framework for annotating and retrieving WWW images," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 960–967.

[5] J. Fan, Y. Shen, C. Yang, and N. Zhou, "Structured max-margin learning for inter-related classifier training and multi-label image annotation," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 837–854, Mar. 2011.

[6] V. Lavrenko, S. L. Feng, and R. Manmatha, "Statistical models for automatic video annotation and retrieval," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2004, pp. 17–21.

[7] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Sep. 2003.

[8] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for image annotation," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 88–105, 2010.

[9] G. Nasierding, G. Tsoumakas, and A. Z. Kouzani, "Clustering based multi-label classification for image annotation and retrieval," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 2009.

[10] L.-D. Nguyen, G.-E. Yap, Y. L., A.-H. Tan, L.-T. Chia, and J.-H. Lim, "A Bayesian approach integrating regional and global features for image semantic learning," in *Proc. 2009 IEEE Int. Conf. Multimedia and Expo*, 2009, pp. 546–549.

[11] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," in *Proc. 10th Int. Conf. Knowledge Discovery and Data Mining*, 2004, pp. 653–658.

[12] J.-H. Su, C.-L. Chou, C.-Y. Lin, and V. S. Tseng, "Effective image semantic annotation by discovering visual-concept associations from image-concept distribution model," in *Proc. 2010 IEEE Int. Conf. Multimedia and Expo*, 2010, pp. 42–47.

[13] J.-H. Su, Y.-T. Huang, H.-H. Yeh, and V. S. Tseng, "Effective content-based video retrieval using pattern indexing and matching techniques," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5068–5085, 2010.

[14] J. Tang, H. Li, G.-J. Qi, and T.-S. Chu, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[15] V. S. Tseng, J.-H. Su, J.-H. Huang, and C.-J. Chen, "Integrated mining of visual features, speech features and frequent patterns for semantic video annotation," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 260–267, Jan. 2008.

[16] V. S. Tseng, J.-H. Su, B.-W. Wang, and Y.-M. Lin, "Web image annotation by fusing visual features and textual information," in *Proc. 22nd Annu. ACM Symp. Applied Computing*, 2007, pp. 1056–1060.

[17] L. Wu, S. C. H. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 135–144.

[18] L. Wu, S. C. H. Hoi, and N. Yu, "Semantics-preserving bag-of-words models for efficient image annotation," in *Proc. 1st ACM Workshop Large-Scale Multimedia Retrieval and Mining in Conjunction With ACM Multimedia*, 2009, pp. 19–26.

[19] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 31–40.

[20] R. C. F. Wong and C. H. C. Leung, "Automatic semantic annotation of real-world web images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1933–1944, Nov. 2008.

[21] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu, "Visual language modeling for image classification," in *Proc. ACM SIGMM Int. Workshop Multimedia Information Retrieval in Conjunction With ACM Multimedia*, 2007, pp. 115–124.

[22] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma, "Annosearch: Image auto-annotation by search," in *Proc. 2006 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1483–1490.

[23] X. J. Wang, L. Zhang, X. Li, and W. Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[24] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple instance learning," in *Proc. 2006 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2057–2063.

[25] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma, "Learning to cluster web search results," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2004, pp. 210–217.

Ja-Hwung Su received the Ph.D. degree from the Department of Computer Science and Information Engineering at National Cheng Kung University, Tainan, Taiwan, in 2010.

He is a postdoctoral fellow in the Department of Computer Science and Information Engineering in National Cheng Kung University. His research interests include data mining, multimedia mining, web mining, and data warehousing.

Chien-Li Chou received the M.S. degree in the Department of Computer Science and Information Engineering at National Cheng Kung University, Tainan, Taiwan. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Information Engineering at National Chiao Tung University, Hsinchu, Taiwan.

His research interests include data mining and multimedia mining.

Ching-Yung Lin received the B.S. and M.S. degrees from National Taiwan University, Taipei, in 1991 and 1993, respectively, and the Ph.D. from Columbia University, New York, in 2000, all in electrical engineering.

He is a Research Staff Member at IBM T. J. Watson Research Center and the IBM Lead of Social and Cognitive Network Science Academic Research Center, Hawthorne, NY. He was an Affiliate Assistant and Associate Professor in the University of Washington, Seattle, from 2003 to 2009, and is an Adjunct Associate Professor in 2005–2006 and Adjunct Professor since 2010 in Columbia University. His research interest includes network science, multimedia security, and multimedia retrieval. He has been focusing on innovating new research paradigms of utilizing multimodality analysis on the social science domain.

Vincent S. Tseng (M'11) is a Professor in the Department of Computer Science and Information Engineering at National Cheng Kung University (NCKU), Tainan, Taiwan. Before this, he was a postdoctoral research fellow in thw University of California at Berkeley during January 1998 and July 1999. He has also acted as the Director for Institute of Medical Informatics of NCKU since August 2008. He has a wide variety of research interests covering data mining, biomedical informatics, multimedia databases, mobile, and Web technologies. He has published more than 200 research papers in referred journals and international conferences, and has held/filed more than 15 patents in the USA and Taiwan.