Neurocomputing 95 (2012) 117–125

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning by expansion: Exploiting social media for image classification with few training examples

Sheng-Yuan Wang, Wei-Shing Liao, Liang-Chi Hsieh, Yan-Ying Chen, Winston H. Hsu*

National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Available online 26 March 2012

Keywords:
Object recognition
Image classification
Web image search
Crowdsourcing
Semantic query expansion

ABSTRACT

Witnessing the sheer amount of user-contributed photos and videos, we argue to leverage such freely available image collections as the training images for image classification. We propose an image expansion framework to mine more semantically related training images from the auxiliary image collection provided with very few training examples. The expansion is based on a semantic graph considering both visual and (noisy) textual similarities in the auxiliary image collections, where we also consider scalability issues (e.g., MapReduce) as constructing the graph. We found the expanded images not only reduce the time-consuming (manual) annotation efforts but also further improve the classification accuracy since more visually diverse training images are included. Experimenting in certain benchmarks, we show that the expanded training images improve image classification significantly. Furthermore, we achieve more than 27% relative improvement in accuracy compared to the state-of-the-art training image crowdsourcing approaches by exploiting media sharing services (such as Flickr) for additional training images.

1. Introduction

Image classification is challenging and one of the enabling techniques for semantic understanding and effective manipulation over large-scale images and videos [1]. The intent of the classification process is to decide whether an image belongs to a certain category or object. The problem has been an active research subject for recent works (e.g., [2–7], etc.). Witnessing the challenging problems in large-scale image and video datasets, where the subject of interest might suffer from noises such as changes in viewpoints, lighting conditions, occlusions, etc., the prior works focus more on various feature representations and sophisticated classification algorithms. However, a promising and orthogonal avenue for the problem is to leverage the web scale images for increasing the training data for image classification. It has been hypothesized that increasing the quantity and diversity of hand-labeled images improves classification accuracy [8] even with preliminary features and learning methods. Although many supervised learning methods prevail in the literatures, manually collecting the required training images/videos is still very painful and time-consuming.

With the prevalence of capture devices and the ease of social media sharing services such as Flickr and YouTube, the amount of freely available image collections on the web is ever increasing. Aside from the visual signals in social media, there are also rich textual and visual cues, device metadata, and user interactions for context-aware social and organizing purposes. The textual cues come from user-provided tags, descriptions for the media, and so on. For the social media, the viewers may leave comments or ratings, bookmark as favorites, or even mark “notes” (visual annotation) surrounding certain regions in the images or videos. The capture devices can also provide geo-location, time, camera settings (e.g., shutter speed, focal length, flash, etc.), which reflect the capturing environment for the scene. These associated information are accumulation of human interactions and (noisy) knowledge. They are promising for solving practical problems (e.g., image classification, question and answering) in the manner of *crowdsourcing*—exploiting the mass interactions on the web. More promising applications are illustrated in [9].

Along with the early attempts (see the reviews in Section 2), in this work, we aim to leverage the rich social media for improving image classification by automatically providing supplemental and diverse training images and auxiliary semantic features. However, we observe the following issues as leveraging the web resources for image classification.

- *Noises*: The crowdsourcing nature of social annotations brings the challenge when learning media semantics, i.e., the inaccuracies and incompleteness of the annotations. It stems from several factors including user subjectivity, video- and album-level annotation, and lack of control on annotation quality. This setback can greatly affect the learning performance if “noisy” tags are directly applied.

[*] Corresponding author.
E-mail addresses: libby@cmlab.csie.ntu.edu.tw (S.-Y. Wang), peegoo77@gmail.com (W.-S. Liao), viirya@gmail.com (L.-C. Hsieh), yanying@cmlab.csie.ntu.edu.tw (Y.-Y. Chen), winston@csie.ntu.edu.tw (W.H. Hsu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2011.05.043

- *Diversities*: It is believed that training data with high diversity would benefit classification performance (e.g., a landmark might have certain variant appearances, cf. Fig. 1(c)). Though having enormous images on the web, the automatic mining methods for acquiring the diverse training images pose another great challenge.

To leverage user-contributed images for reducing human labors in acquiring training data and further improving diversity (completeness) in training image pools, we propose *learning by semantic expansion* to augment training data with supplemental diverse images and additional semantic features. Lacking textual information in the training images, we correlate them to visually similar images from the social media for further graph-based expansion. The key is that given the *auxiliary image collection* from social media (i.e., Flickr), we can find very similar images to the training image even when matching with simple image features as demonstrated in [8]. Given an original training example (Fig. 1(a)), Fig. 1(b) demonstrated the expanded results from the auxiliary image collection using visual features only. However, they bear strong visual similarity with each other (i.e., low diversity) by content-based image retrieval (CBIR) methods. By incorporating rich (but noisy) textual information from the auxiliary image collection, we can further improve the diversity of the expanded images. Fig. 1(c) shows the top 15 images expanded by semantic expansion from the initial CBIR results.

The proposed expansion method, provided the few training images, aims to mine semantically related (probably visually diverse) images by expansion over effective image graphs (textual and visual). Note that we do not need to provide a list of category names of interest required in previous works (e.g., [10–12]) but only few training images. Evaluating in common benchmarks, our proposed method shows significant classification gains by expanding training images to bring more diverse external images into the training process (as shown in Fig. 1(c)), and also outperforms the prior state-of-the-art in crowdsourcing for image classification. Meanwhile, we also consider scalability issues as constructing the image graphs. We will show the potential reduction for human annotations for supervised learning and further investigate parameterizing factors for the proposed framework (cf. Fig. 2).

The key contributions of this paper include:

- We propose the novel semantic expansion method which exploits social media for training example expansion. In comparison to other state-of-the-art methods, our method improves classification accuracy by more than 27%.
- We incorporate efficient image graph construction and such graphs have been shown effective for image expansion by

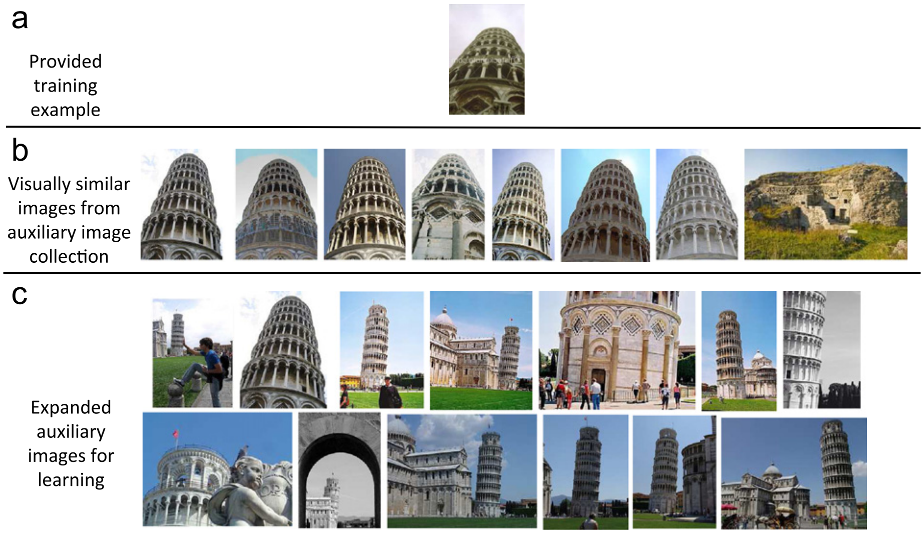
a
Provided training example
b
Visually similar images from auxiliary image collection
c
Expanded auxiliary images for learning

Fig. 1. The need for training image expansion. The top ranked images retrieved from the auxiliary image collection with the training image (a) as the query, by fusing multiple visual features (b) and by the proposed (automatic) semantic expansion framework (c). The latter yields more visually diverse but semantically related images, shown very effective for augmenting image classification.

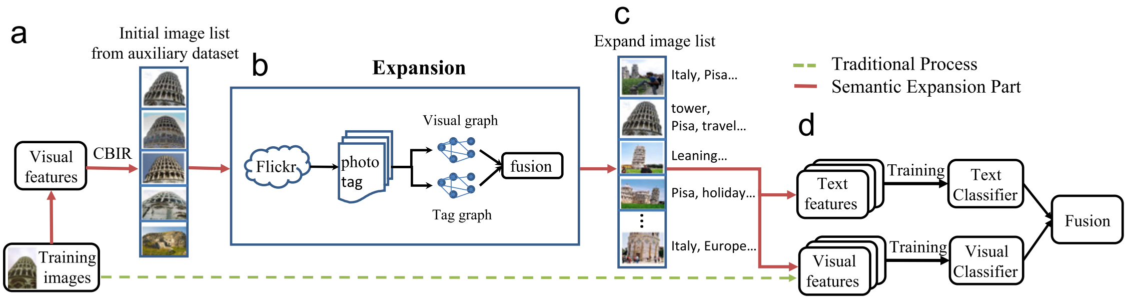

Fig. 2. We leverage the social media as the *auxiliary image collection* and build a combinational image graph from them in the offline stage. The few training images first derive an initial image list (a) by the CBIR method the auxiliary image collection and further refined by our expansion scheme (b). The expanded images (c) are used as extra training examples for classification in (d). Our expansion method additionally derives semantic text features as byproducts and further facilitates multimodal classification though provided visual training images only. Comparing with the conventional learning model, we can expand more visually diverse (but semantically related) images for robust learning.

efficient random-walk-like methods. The combinational (multimodal) image graph provides a mean to efficiently merge web-based knowledge into the semantic expansion scheme.

- The expansion method promisingly improves the diversity issue for image classification. We also investigate impacts of diversity in both image retrieval and classification.

The remainder of this paper is organized as follows. We will discuss related approaches in crowdsourcing for image classification in Section 2. We provide detailed descriptions for the proposed semantic expansion and classification techniques in Section 3. We will detail the experimental settings and results in Section 4. We conclude and discuss the future extensions in Section 5.

2. Related work

In this work, rather than on sophisticated learning models or complex features that researchers had exploited, we are to focus on crowdsourcing from auxiliary image collection for more training images. Especially, there is a need for large-scale semantic understanding for images and videos [13,1], where hundreds or thousands supervised learning are required. The need for efficiently collecting training images or reducing the number of training images is vital. The most intuitive solution is to grab the (training) images directly from the search engine by keywords. However, the text (or tag) associations with images are noisy [14]. Manual cleanup for the retrieved images is formidable for large-scale image classification.

Generally, two directions are promising for leveraging Internet image collections. The first is to improve annotation quality or filter out spurious results returned by keyword queries as acquiring a new collection of images that can be used for training. The second is to derive alternative representations for the training images – reflecting the image semantics. The former is to automatically identify the correct association between the labels (textual tokens or tags) and images without any human intervention; for example, researchers select effective images from web image search engines (by keyword-based queries) as training data [10–12,15]; other researches have used the captions associated with the news photos to learn face recognition models without manual supervision [16]. The latter for leveraging web resources for alternative (semantic) representations is proposed in [17,18], where the tags from similar images in the image collection are used as another semantic feature representation. However, automatic acquisition of images has not been taken into consideration. Our approach deals with the both directions discussed above and significantly improves classification results. Meanwhile, we will compare the proposed method with [17] in the experiments (cf. Section 4.4.1).

Leveraging crowdsourcing for image classification, an alternative way is to conduct semi-supervised learning among labeled and unlabeled images. It benefits the learning process using a limited amount of labeled images [19–21]. In [21], image tags associated with labeled and unlabeled images were used to improve the classifier. The classifier was learned using both content and keywords by labeled images to score and choose the unlabeled ones in an auxiliary image collection. The purpose is very similar to our proposed method in obtaining additional examples to train classifiers. However, we observe that the textual (tag) information is mostly noisy and missing. Assuming the text availability for both training and test images is unrealistic. In contrast, by utilizing multimodal (visual and textual) visual graphs, we can expand the semantically related (and diverse) training images and automatically generate text cues as additional features for classification even though we are not provided any keywords as training input. Meanwhile, it is worth noting that most of the existing semi-supervised approaches heavily rely on the images labeled by human. As a result, a sufficient amount of labeled images are needed to ensure that the pre-constructed model is stable enough to select accurate examples and further refine them. We will show that our method can tolerate very few images provided (e.g., 5) where the other methods require more (e.g., 30 or more), experimented in an image benchmark.

3. Proposed framework

The goal of this work is to enable automatic expansion of training examples to improve the image classification or even easing the annotation efforts for supervised learning. Fig. 2 shows the expansion and classification procedure. With an *auxiliary image collection* from social media, given a few training images, we do content-based image retrieval first, which maps our training examples to the similar images within the auxiliary image collection by visual features (described in Section 3.1). Due to the semantic gap problem, most of CBIR methods are not satisfying and suffer from the common problem of high precision and low recall in the retrieval results. Dominated by visually similar images, the diversity issue, most ignored in prior crowdsourcing methods, is usually poor.

Our approach – semantic expansion – is to improve the search quality for common content-based image retrieval (CBIR) methods and tackle the low diversity problem. As illustrated in Fig. 2, in the offline stage, we model images in the auxiliary image collection as visual and textual (tag) graphs (introduced in Section 3.2.1). Such graphs model the multimodal semantic correlations from the noisy images and can be later used to mine more semantically related images or tags for learning image classification. We will introduce the expansion procedure in Section 3.2.2. Also, other semantic representations can be derived in Section 3.3. Such expanded images, with visual diversity and complementary semantic representations, have great potential to benefit classification problems. Finally, the classification is conducted in Section 3.4.

3.1. Initial image list generation

The first step in our expansion scheme is mapping the training examples to visually similar images in the auxiliary image collection as seeds for expansion over the semantic graph. Given few training images, we derive the initial image list by CBIR methods [22]. Note that the initial images are retrieved automatically and they are not restricted to any categories. To construct a robust CBIR baseline, we investigate variant visual features and fusion strategies. The effectiveness are then demonstrated in the experiments (cf. Section 4.2).

Visual words (*VW*): Scale-invariant feature transform (SIFT) feature [23] is widely used in image retrieval and recognition. We adopt Hessian-affine detector to extract salient patches from each image that contains rich local information. The SIFT features are then quantized into bag-of-visual words. A vocabulary of 10K VW is used. Each image is then represented by a histogram of VW occurrences.

Gist: Gist feature [24] is proven as an effective descriptor in scene representation. Gist describes visual information which was caught by a glance at the scene. The process of extracting the gist of an image uses features from several domains, calculates holistic characteristics, and still takes into account coarse spatial information. Each image is represented by a 960 dimensional Gist descriptor.

PHOG: PHOG [25] captures the shape characteristics of images. For each image, the edge contours are extracted using Canny edge detector and the orientation gradients are then computed using Sobel mask. A histogram of gradient vector is computed for each grid cell at each pyramid level. The final descriptor is a concatenation of all HOG vectors. We take 20 bins for the gradients and adopt three pyramid levels (1, 4, 16 cells, respectively). The dimension of the final PHOG feature is 420.

Color: The color features are extracted in RGB domain. For each pixel, we quantize each channel to eight bins and assign pixels to the corresponding color bins, ranging from 1 to 512. The bins are then normalized. Each image is represented as a 512 dimensional histogram.

Unified: Unified feature is the concatenation of the above four visual features. Each weight is proportional to the classification performance for each feature modality alone, measured by cross-validation.

3.2. Semantic expansion

We formulate the semantic expansion as a random walk [26,27] problem over image graphs, where nodes are images and the link between each node is the pairwise similarity. We then derive more semantically related images by the random walk process from the initial CBIR list, also illustrated in Fig. 2. Note that since we target to operate in large-scale image graphs, we also consider effective methods for visual and text graph construction. We also consider the noisy and missing text tokens (tags) problems commonly observed in images from social media.

3.2.1. Graph construction

The semantic expansion is first based on graph construction. We construct two complementary types of graphs: visual and textual. For visual representation, the images are represented by 10K VWs. For textual graph, we collect textual information associated with images from social media website (Flickr). The raw texts are not directly used as feature representation. They are further expanded by Google snippets from their associated (noisy) tags. Through the expansion process, semantically related user-provided tags (e.g., "Eiffel" versus "Eiffeltower") can still have high similarities. Each image then can be represented with 50K text tokens. In our experiments, the photo information "title," "description," and "tag" collected from Flickr are used as textual information to construct the graph.

The graphs are constructed with visual and textual similarity of images in auxiliary image collection in the offline stage. However, it is very challenging to construct image graphs from large-scale datasets. To tackle the scalability problem, we construct image graphs using MapReduce model [28], a scalable framework that simplifies distributed computations. The visual and textual graphs are constructed separately. We take the advantage of the sparseness and use cosine measure as the similarity measure.

We observe that the textual and visual features are sparse for each image and the correlation between images are sparse as well. Our algorithm extends the method proposed in [29]. The algorithm uses a two-phase MapReduce model: indexing phase and calculation phase, to calculate pairwise similarities. At indexing phase, each input image feature vector is mapped to a list of feature-image-value tuples, and then all the tuples are organized into inverted lists according to their features. At calculation phase, the inverted lists are used to compute a similarity value for each pair of images in each inverted list. Then the values for each pair of images are aggregated into a single value, which is the pairwise similarity value of corresponding feature pair of images. Unlike df-cut for documents in [29], we found that tf-idf is a better choice to reduce computation for images.

It takes around 42 min to construct a graph of 550K images on 18-node Hadoop servers. Through the evaluation on a small-scale dataset, we found the quality to be close to the brute-force algorithm.

Fusing the visual and text image graphs (G^v and G^t, respectively), we build a final combinational (multimodal) image graph G, where we conduct the random walk process for image expansion. Assuming we have N images in the auxiliary image dataset, the graphs (i.e., G^v, G^t, G) are then $N \times N$. However, we observe that they are quite sparse and most of the entries are zero. To simplify, G can be generated directly by a linear combination as follows:

$$G = \alpha G^t + (1-\alpha) G^v \tag{1}$$

$\alpha \in [0,1]$ and $(1-\alpha)$ are the weights for the two image graphs. The sensitivity for α is investigated in Section 4.5. We will show how each image graph improves the overall results in the experiments. In the following section, we will leverage the graphs for example-based semantic expansion.

3.2.2. Expansion on semantic graph

Give the few training images, we initially identify those visually similar images (assuming their semantic similarities to the query images) in the auxiliary image collection by CBIR methods described in Section 3.1. The goal is to link the initial training images to the rich image collections through these visually similar images. We then try to expand and identify more semantically related images from the semantic graph G. It is observed that those connected images in the graph are with certain (visual and textual) similarities. The higher the weights the more semantically similar they are. Based on the observations, we can utilize the semantic graph G, constructed in the offline stage, to mine more semantically related images from the auxiliary image collections, which are potential for augmenting the training images for image classification, as illustrated in Fig. 1.

The problem can actually be modeled as a random walk problem [30], which had been shown effective in certain applications such as image/video reranking [31] and product search by multiple similarities [27]. Given the semantic graph G, let us assume a surfer randomly surfing the image graph. The surfer starts from certain initial nodes in the graph and then jumps to different nodes iteratively. The probability for jumping from one to another is proportional to their pairwise similarity, also called the transition probability, G', where ensuring the row summarization is 1 by normalizing G. When converging, the stationary probabilities over the nodes are set as the final image list ordered as the probability a random surfer stays. We set the top K CBIR-retrieved images (from the auxiliary image collection) as the initial nodes for random walk.[1] Let $\mathbf{v} \in [0,1]^{N \times 1}$ be the initial state vector for the images in the auxiliary collection. We set $\mathbf{v}$ for the K initially retrieved images by their similarities to the (multiple) queries (normalized by the *sigmoid* function) and 0 for the others. Note that we also normalize $\mathbf{v}$ and ensure $|\mathbf{v}| = 1$. Here we investigate two alternative methods for random walk.

Random walk (*RW*): Besides G', usually, there is also a possibility that the surfer does not follow the probabilistic transition matrix induced from the image graph G' but jumps to another image uniformly from the collection. It is used to make graph connected and ensures the existence of stationary probability. In RW the transition matrix can be modified as $P = \epsilon G' + (1-\epsilon) U$,

[1] Through the sensitivity test we found that the quality for the expanded results are quite similar as varying K from 10 to 50. Here we set K=20.

where U is a uniform matrix with the same size of G and ϵ is used to modulate the weights on the normalized semantic graph and the uniform jump.

Random walk with restart (*RWR*): During each RW iteration, the random surfer would restart from the initial states (specified by $\mathbf{v}$) with probability $1-\epsilon$, and jump between states following the transition probability of G' with probability ϵ. For RWR, the transition matrix can be modified as $P=\epsilon G'+(1-\epsilon)V$, where $V=\mathbf{e}\mathbf{v}^T$ ($\mathbf{e}$ is an N-dimensional column vector with all 1). The intuition is that the random surfer tends to persevere his preferences, specified by $\mathbf{v}$, during each random surfing.

For RW and RWR, we then adopt power method for solving the random walk problem – deriving the dominant eigenvector from the transition matrix (P) for the stationary probability $\boldsymbol{\pi} \in [0,1]^{N\times 1}$ [30]. The top L images in the auxiliary image collection with the highest stationary probability $\boldsymbol{\pi}$ are selected later for training for image classification. Note that we will vary L and investigate how they impact the classification performance in the experiments (cf. Section 4). More technical details regarding random walk can be found in [30,31], which also suggests setting $\epsilon=0.85$ for the experiments.

3.3. Semantic similarity from auxiliary image collection

Besides using the image graph to do training example expansion, we also leverage it to obtain alternative semantic representation for our training (or testing) examples. Having an auxiliary image collection from the social media, for each image, the semantic similarity can be discovered by mining the related text features.

Text feature (*text*): The textual feature of each training (or test) image is derived by propagating the associated textual information of its K nearest images in the auxiliary dataset determined by the similarity in the visual graph (Section 3.2); the associated textual information includes “title,” “description,” and “tags” along with each image, if available. By summing the tag (text) counts in the expanded image collection, we can use the (normalized) occurrence frequencies as another semantic feature representation for the training (testing) images. The intuition is to locate the semantically related images in the auxiliary image collection and propagate their associated tags (or text tokens) to the (training or testing) image. See the illustration in Fig. 1(c). Such method is similar to that in [17]. In the experiments (Section 4.4), we will show that the expansion process does help classification (even using unified visual features only) and the accuracy can be further improved by fusing visual and textual features.

3.4. Classification

Various classifiers could be applied to such expansion scheme we proposed. For our experiments of image classification, we use the chi-square kernel for our features in a SVM classifier. In the same manner as [25], multi-way classification is achieved using a one-versus-all SVM: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest classification margin. However, single feature alone is not sufficient to distinguish all types of images. Leveraging the text and visual features, the fusion is conducted by combining two kernel functions (as a late fusion method). The experiments for image classification baseline and the impacts by varying expanded images will be shown in Section 4.

4. Experiments

In this section, we first introduce the datasets used in our experiments including training, test, and auxiliary image sets. Since the expansion result would be affected by the quality of the initial CBIR list, illustrated in Fig. 2(a), we will also brief the CBIR performance by different features in Section 4.2. Then, we measure the quality of the expanded images (cf. Fig. 2(c)) in Section 4.3. Finally, in Section 4.4, we discuss the impacts for image classification by the proposed image expansion methods (cf. Fig. 2(d)).

4.1. Datasets

4.1.1. Training and test set—Caltech-6

To compare with the traditional approaches and existing works that use web resources for classification, we first conduct experiments on six selected categories from the Caltech 256 dataset [7], which are widely used benchmark in object classification. The categories include three landmark (The Eiffel Tower, The Golden Gate Bridge, The Tower of Pisa), two animals (Giraffe, Penguin), and one equipment category (Basketball Hoop).

4.1.2. Auxiliary image collection from social media – Flickr13K

Flickr13K is a large dataset consisting of 13,381 medium resolution (500×360) images. Most of them are the subset of the Flickr550 dataset [32] downloaded from Flickr. We manually labeled 873 images as our ground truth across three categories including the Tower of Pisa (139), the Eiffel tower (544), and the Golden Gate Bridge (190). We further collected 1535 positive images for the other three categories: giraffe (428), penguin (579) and basketball hoop (528) from Flickr. We also randomly selected 10K images among 550K images as background images. We merge the ground truth and background images as our auxiliary image collection with the size of 13K.

4.2. CBIR for multiple modalities and images

The first step of our expansion scheme is content-based image retrieval, which maps our training examples to the images in the auxiliary image collection; the latter will be the seeds for semantic expansion to further discover the semantically related images (for training) in the multimodal image graph. We observe that the quality of the initial list might be essential for expansion results. We first evaluate the retrieval results by varying features and fusion strategies. Since only top K images are set as the initial states during the semantic expansion procedure, we just evaluate the performance of top ranked images in the initial list by P@10 (precision at the top 10 retrieved images).

From the evaluations in Tables 1 and 2, landmark categories (e.g., pisa tower and golden gate bridge) are much easier to get good retrieval results by visual words. Still, some query images are with sparse visual words but benefit from the global contextual features (e.g., Gist and PHOG). As the result, we fuse the four visual features mentioned in Section 3.1, as “Unified” and achieve 0.393 in averaged P@10 (across six categories). Note that the queries are evaluated in a leave-one-out manner in order to utilize the whole ground truth images.

Table 1
The overall retrieval performance (P@10 in %) for the initial CBIR image list by varying different features such as VW (V), GIST (G), HOG (H), and Unified (U). “Mq_” stands for utilizing multiple queries from the training images. The best setting is to use multiple queries combined with unified features. The setup is used for following experiments.

VW	Gist	HOG	Color	Unified	Mq_V	Mq_G	Mq_H	Mq_C	Mq_U
38.3	24.6	29.7	9.7	39.3	40.7	49.3	41.6	18.0	**57.7**

Table 2
The retrieval performance (P@10) of the initial CBIR image list over six categories with single query and multiple queries (Mq_U) from the training images. Note that both use "unified" (multimodal) features.

	Basketball	Eiffel	Giraffe	Golden	Penguin	Pisa
Single	0.16	0.41	0.43	0.25	0.47	0.48
Mq_U	0.36	0.72	0.51	0.42	0.40	0.78

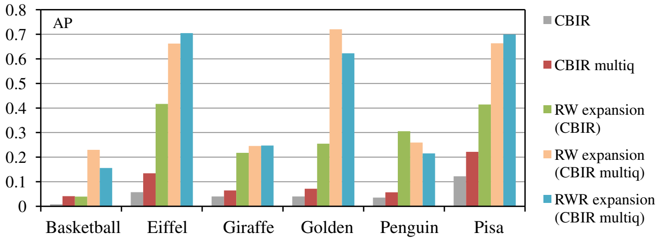

Fig. 3. Retrieval performance (in AP) over six categories by the variant expansion methods, which further improve the initial CBIR results significantly. See more discussions in Section 4.3.

Since there might be multiple training images provided for image classification, we are keen to know the effectiveness as aggregating multiple training images (i.e., 5) and multiple queries (multq or Mq_). The results look reasonable across different features as using multiple queries as showing constant improvements. With multiple features, multiple queries (Mq_U) can improve the single query (Unified) from 0.39 to 0.58 (P@10). See details in Tables 1 and 2. Note that the multiple queries are randomly selected and repeated many times for each query.

4.3. Quality for expanded images

Given the few training images, our purpose is to expand more semantically related and diverse images later for augmenting the training images for classification. Note that we take the CBIR retrieved images in the auxiliary image collection by multiple (i.e., 5) queries and multiple modalities (Unified) as the seeds for expansion. We will evaluate the expansion quality first and then evaluate their impacts in classification in the next section.

In Fig. 3, we can clearly see the significant improvements (in MAP[2]) by the proposed semantic expansion methods. Landmark-related categories (e.g., the Tower of Pisa, the Eiffel Tower, the Golden Gate Bridge) all get higher than 0.6 in AP after the expansion. In general, through semantic expansion, we can improve the performance by more than 100% for each category. In other words, the noises of expanded images have been eliminated through our semantic expansion procedure. Comparing Table 2 and Fig. 3 at once, we can see that the expansion is sensitive to the quality of the initial CBIR list. Initial lists by multiple queries result in much better expansion performance than those by single query only. For topic "penguin," since the P@10 by multiple query (0.4) is less than that by single query (0.47), the expansion MAP is lower as well. In topic "basketball hoop", due to low precision by single query (0.16), the MAP is not improved after expansion. However, Most CBIR results by multiple queries (multiq) are significantly boosted by the expansion process.

We also compare expansion using different graph refinement methods mentioned in Section 3.2.2. In our experiments, simple random walk (RW) is superior to random walk with restart (RWR) in certain query categories. On the average, RW performs better than RWR. However, both of them have small differences in accuracy. Definitely the expanded results are better than the initial list only (CBIR or CBIR multiq). See more details in Fig. 3.

Note that in our experiments later, the expansion accuracy is not proportional to the classification accuracy. We will show that the diversity for the expanded images pose another importance factor. We will show that the proposed semantic expansion method by multimodal image graphs can actually improve the diversity rather than retrieval accuracy only. See the discussions in Section 4.6.

4.4. Classification results by semantic expansion

The expanded images from the expansion framework are combined with the few initially annotated images (fixed at five in our experiments) for further training image classification models for the separated test set. To assess the effectiveness of our proposed expansion scheme, the evaluations focus on: (1) Comparing the image classification performance with different expansion methods. (2) Comparing the expansion methods with human annotations for the training images only. (3) Comparing our method with one of the state-of-the-art [17], also exploiting auxiliary image collection to derive additional (textual) features for each image. See the descriptions in Section 3.3. All experiments are also evaluated based on different modalities.

4.4.1. Expansion versus non-expansion

We first evaluate if the proposed expansion method is effective for bringing more semantically related image from the auxiliary image collection. The top K expanded images (cf. Fig. 1(c)) from the auxiliary image collection are used to train image classifiers for the six categories from Clatech-256 test set. In this experiment, if expanding too many images, we observe that few false positives in the expanded images might occur and degrade the classification performance; it is due to the limited number of the controlled experimental dataset. We hypothesize that the problem can be solved by enlarging the auxiliary image collection to provide sufficient true positives for training image expansion. We compare that with the top K CBIR images only (cf. Fig. 1(a)) without the semantic expansion. Note that we have $K=70$ and the CBIR images are by multiple (i.e., 5) query images from the Caltech-256 training images. The results are compared in Table 3. Even without semantic expansion, the auxiliary textual features (Text) (cf. Section 3.3 and [17]) still shows the effectiveness (0.65) over the visual features only, 0.58 (visual words) and

Table 3
Semantic expansion significantly improves image classification on the test set of the six categories selected from Caltech-256. (a) Image classification accuracy by learning from the top K CBIR images by the provided five query images from the Caltech-256 training set (cf. Fig. 1(a)). (b) Learning by the top K images by semantic expansion (cf. Section 3). Note that we have $K=70$ and evaluate at different features and fusion strategies for image classification. See more discussions in Section 4.4.1.

	(a) No expansion				(b) Semantic expansion			
	VW	Unified	Text	Fusion	VW	Unified	Text	Fusion
Basketball	0.45	0.47	0.42	0.49	**0.78**	0.70	0.40	0.71
Eiffel	0.30	0.33	0.67	0.63	0.72	0.76	0.82	**0.90**
Giraffe	0.63	0.64	0.57	0.68	0.66	0.82	0.86	**0.93**
Golden	0.72	0.75	0.73	0.82	0.90	**0.96**	0.70	0.95
Penguin	0.52	0.62	0.67	0.72	0.74	0.78	0.86	**0.87**
Pisa	0.83	0.74	0.84	0.83	0.86	0.80	0.80	**0.94**
Accuracy	0.58	0.59	0.65	0.70	0.78	0.80	0.74	**0.88**

[2] MAP: mean average precision across all the queries to evaluate the overall system performance. Average precision is to approximate the area under a non-interpolated precision–recall curve for a query. See more introductions in [32].

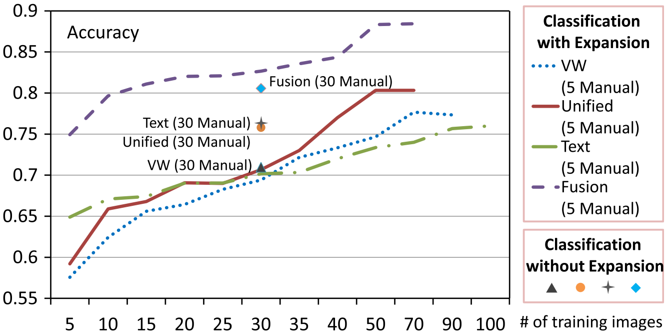

Fig. 4. The image classification accuracy by varying expanded image numbers (based on five initially provided annotated images). We also show the performance as using 30 manually annotated images from the benchmark. The experiment shows that the automatically expanded images significantly benefit the classification and even outperform human annotations.

0.59 (unified) in classification accuracy. The fusion of visual and textual features can even improve the accuracy to 0.70, which matches the observation reported in [17].

For our expansion method, the performance is listed in the four columns to the right in Table 3. We can observe that our expansion method can improve the classification (averaged) accuracy from 0.70 to 0.88 (27% relatively) by the fusion feature. Apparently the proposed expansion method is complementary to the textual feature proposed by [17]. Besides, it is worth mentioning that classification with semantic expansion using text feature does not outperform the one using visual features. The reason might be that the (semantically) expanded images are derived from the image graph which consists of both visual and textual similarities.

4.4.2. Expansion versus human annotation

Furthermore, we are interested in the impacts of increasing the number of expanded images. We also compare our method to the traditional supervised approaches with 30 human-annotated images. Fig. 4 shows the performance with different number of expanded (training) images for classification on the test set. In '*M*-Manual' labels, *M* denotes how many images are manually annotated and then used as seeds to expand more images for training. We evaluate at varying numbers (i.e., 5–100) of expanded images by the multimodal image graph. We also compare classification performances by different features and fusion strategies give the expanded training images. As we increase more expanded images automatically, we found a notable increase in classification accuracy. The accuracy of our results with 30 training images (5 annotated plus 25 expanded by our method with feature "unified") yields 70.7%, which is close to 75.8% by 30 manually annotated images. A superior performance is obtained when the number of expanded images (from the auxiliary image collection) exceeds 35. Moreover, further using 50 expanded, we achieve similar accuracy with visual features only against the best performance of using 30 manually annotated training images (80.3% versus 80.5%). Note again that our system uses only 5 annotated examples to expand training images automatically.[3] Finally, our fusion results are well improved and are much better than non-expansion (manually annotated) methods because of the high diversity in training data. Also, the semantic correlations discovered through expansion is highly complementary to visual features. The best performance we obtained is 88.4% (learning from the automatically expanded training images by five annotated ones only).

Comparing with the state-of-the-art approach with the similar goal for mining tags form the similar images from the auxiliary image collection as the additional (text) feature [17] (i.e., 80.5% by 30 manually annotated training images), the proposed learning by expansion method is significantly better. Meanwhile, our method requires very few annotated images and demonstrates its superiority to its counterpart with manually annotated training images.

4.5. Significance of weights on image graph combination

To analyze the impact of weights on image graph combination, we compare the quality of the expansion results (in MAP) by varying α in Eq. (1), which denotes the weight of text graph. We plot the result in Fig. 5. It states that in the best case, the weight for G^t is around 0.7. The performances of α from 1 to 0.3 are in the plateau. However, when the text graph weight is less than 0.3, the performance decreases dramatically. Since the initial list is generated by CBIR with visual features only, expansion by image graphs, parameterized with text cues for expansion, is really beneficial. On the contrary, heavily relying on the visual graph for expansion causes bias and degrades the expansion quality (in MAP). With the combination of image graphs, we can leverage both visual and textual features and improve expansion quality. Note that though the expanded images are with similar (retrieval) MAPs, they will have different impacts on learning for image classification—majorly due to the diversity from the expanded images. We will look into the issues in Section 4.6.

4.6. Diversity on expanded images for classification

We evaluated the significance of weights on fusing visual and textual image graphs in Section 4.5. From Fig. 5, we observe that the expansion results with α ranging from 1 to 0.3 are flat with quite closed (retrieval) MAPs (rather than classification accuracy). For investigation, in Fig. 6, we further list the top 10 expansion results of the two categories ("Penguin"and "Basketball") with different α, 0.8 and 0.4, respectively, i.e., incurring different visual diversities for the expanded images. Note that they have similar performances in (ranking or retrieval) MAP (see Fig. 5) but different in visual diversities. If we weight more on visual graph (i.e., $\alpha = 0.4$), the random surfer tends to rely on visual cues more for the expansion, more visually similar images will be expanded, i.e., low diversity. On the contrary, as weighting more on text graph (i.e., $\alpha = 0.8$), more semantically similar but visually diverse images are yielded.

We further take the expended images (with the two diversities, respectively) as training images for image classification. The

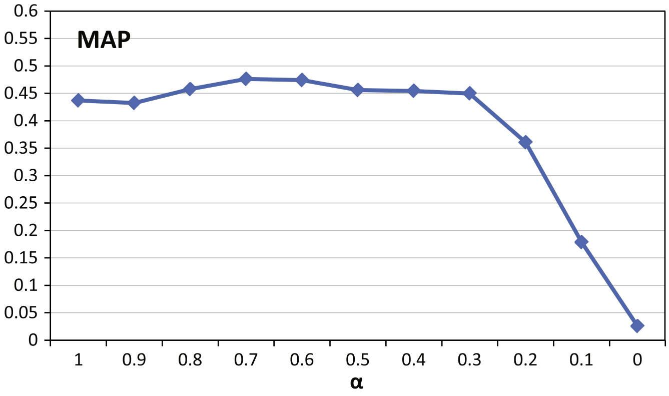

Fig. 5. The quality (in MAP) for the expanded images over the multimodal image graph by varying α in Eq. (1) for visual and textual fusion. The best MAP is achieved as α=0.7. See more discussions in Section 4.5. We show its impacts on image classification (for training images) in Section 4.6.

[3] The results are the averaged performances among the repetitive experiments by randomly sampling five training images from the training set.

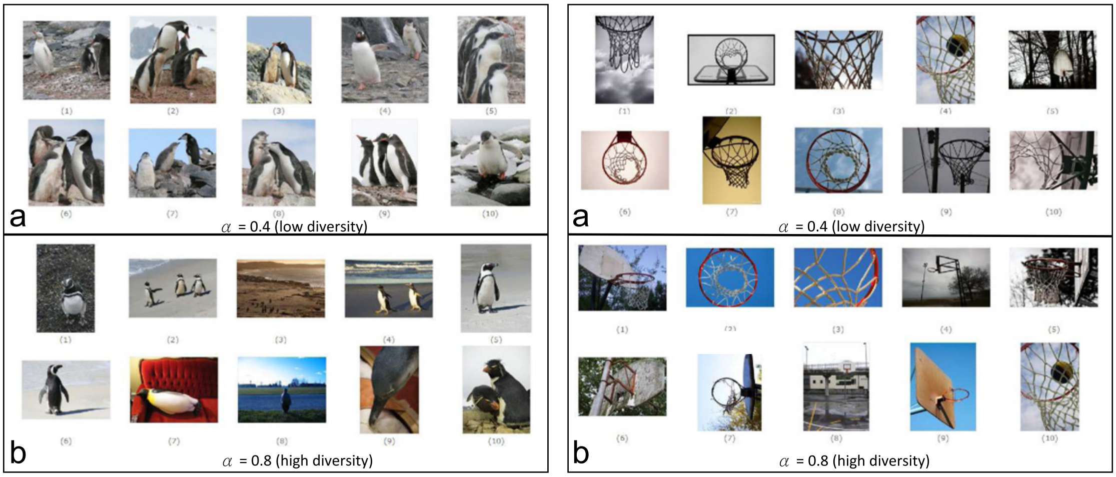

Fig. 6. Expanded images for "penguin" and "basketball" with different α for controlling the image diversity. The similar performance (in retrieval ranking by MAP) are achieved, but the results are with varying degrees of visual diversity. Higher α, weighting more on textual cues, bring more diversity. See more discussions in Section 4.6.

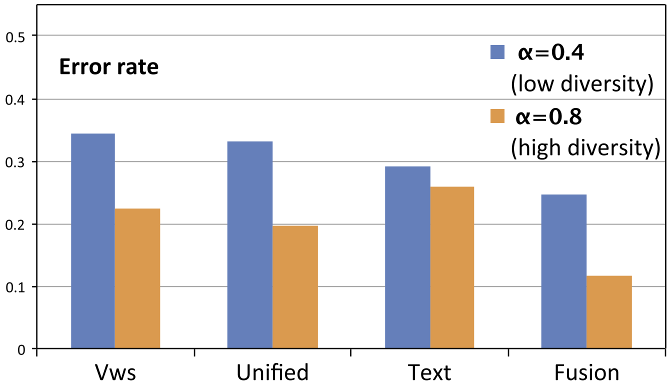

Fig. 7. The classification performance (in error rate) by varying image diversity over difference classification features. High diversity helps image classification a lot across classification features.

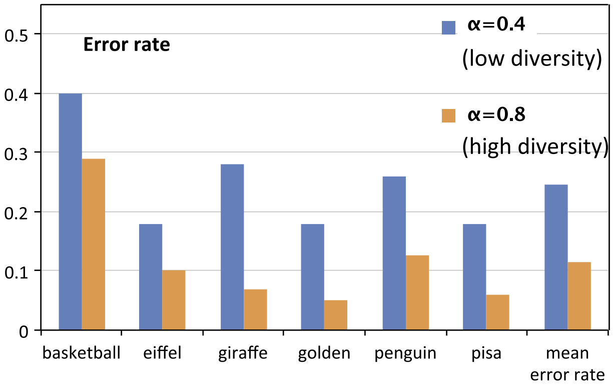

Fig. 8. The classification performance (in error rate) by varying image diversity across different image categories. High diversity helps image classification consistently across image categories.

classification over different classification features are shown in Fig. 7. Note that their MAPs for the ranking of expanded images are about the same but with different classification accuracies consistently across classification features. It states clearly that more diverse (but semantically consistent) training images help image classification more since they might cover more on the possible feature space for the positives. We also report the classification accuracies across image categories in Fig. 8, the diverse images still help image classification (with low error rates) even across image categories.

The experiments show that the expansion over the multimodal image graphs bring more semantically consistent but visually diverse images from the auxiliary image collection (i.e., Flickr) and can further improve the image classification quality as learning on more diverse training images. Note that such images are expanded in an automatic fashion by few training images only and can further save the efforts required for image annotation, as discussed in Section 4.4.2.

5. Conclusions and future work

We have proposed a novel semantic expansion framework to mine more semantically related images as supplemental training images from the explosive user-contributed images and their associated tags. Such mined images are freely available and very helpful for image classification as reducing the number of manual annotations and even improving the detection accuracy by including more visually diverse images for learning. The scalability issue is also considered as constructing the image graph for expansion by distributed computation (i.e., MapReduce). Experimenting in certain benchmarks, we show that the proposed method outperforms the state-of-the-art crowdsourcing method by more than 27%. For our future work, we are extending the classification benchmark for further generic evaluations and investigating scalable methods for constructing large-scale image graphs for effective expansion. We are also increasing the size of auxiliary image collection and investigating the impacts for training image expansion.

References

[1] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia, IEEE MultiMedia 13 (3) (2006) 86–91.
[2] A. Hegerath, T. Deselaers, H. Ney, Patch-based object recognition using discriminatively trained gaussian mixtures, in: CVPR, 2006.
[3] I. Laptev, Improvements of object detection using boosted histograms, in: BMVC, 2006, pp. 949–958.
[4] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006, pp. 2169–2178.
[5] H. Zhang, A.C. Berg, M. Maire, J. Malik, Svm-knn: Discriminative nearest neighbor classification for visual category recognition, in: CVPR, 2006, pp. 2126–2136.

[6] J. Zhang, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, Int. J. Comput. Vision 73 (2007) 2007.

[7] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology, 2007.

[8] A. Torralba, R. Fergus, W.T. Freeman, 80 million tiny images: a large data set for nonparametric object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1958–1970.

[9] T. Mei, W.H. Hsu, J. Luo, Knowledge discovery from community-contributed multimedia, IEEE MultiMedia 17 (4) (2010) 16–17.

[10] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from Google's image search, in: ICCV, 2005.

[11] T.L. Berg, D.A. Forsyth, Animals on the web, in: CVPR, 2006, pp. 1463–1470.

[12] F. Schroff, A. Criminisi, A. Zisserman, Harvesting image databases from the Web, in: ICCV, 2007.

[13] J. Deng, W. Dong, R. Socher, L. Jia Li, K. Li, L. Fei-fei, Imagenet: a large-scale hierarchical image database, in: CVPR, 2009.

[14] L. Kennedy, M. Naaman, S. Ahern, R. Nair, T. Rattenbury, How Flickr helps us make sense of the world: context and content in community-contributed media collections, in: ACM Multimedia, 2007.

[15] S. Vijayanarasimhan, K. Grauman, Keywords to visual categories: multiple-instance learning for weakly supervised object categorization, in: CVPR, 2008.

[16] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E.G. Learned-Miller, D.A. Forsyth, Names and faces in the news, in: CVPR, 2004, pp. 848–854.

[17] G. Wang, D. Hoiem, D. Forsyth, Building text features for object image classifications, in: CVPR, 2009.

[18] G. Wang, D. Hoiem, D. Forsyth, Learning image similarity from Flickr groups using stochastic intersection kernel machines, in: ICCV, 2009, pp. 428–435.

[19] O. Delalleau, Y. Bengio, N. Le Roux, Large-scale algorithms, in: O. Chapelle, B. Schölkopf, A. Zien (Eds.), Semi-Supervised Learning, MIT Press, 2006, pp. 333–341.

[20] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: COLT, 1998, pp. 92–100.

[21] M. Guillaumin, J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: CVPR, 2010.

[22] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: ICCV, 2003.

[23] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.

[24] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.

[25] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: CIVR, 2007, pp. 401–408.

[26] W.H. Hsu, L.S. Kennedy, S.-F. Chang, Reranking methods for visual search, IEEE MultiMedia 14 (3) (2007) 14–22.

[27] Y. Jing, S. Baluja, Pagerank for product image search, in: WWW, 2008.

[28] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.

[29] T. Elsayed, et al., Pairwise document similarity in large collections with mapreduce, in: Proceedings of ACL-08: HLT, Short Papers, 2008, pp. 265–268.

[30] A.N. Langville, C.D. Meyer, A survey of eigenvector methods for web information retrieval, SIAM Rev. 47 (1) (2005) 135–161.

[31] W.H. Hsu, L. Kennedy, S.-F. Chang, Video search reranking through random walk over document-level context graph, in: ACM Multimedia, 2007.

[32] Y.-H. Yang, P.-T. Wu, C.-W. Lee, K.-H. Lin, W.H. Hsu, Contextseer: Context search and recommendation at query time for shared consumer photos, in: ACM Multimedia, 2008.

Sheng-Yuan Wang is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan.
E-mail: libby@cmlab.csie.ntu.edu.tw.

Wei-Shing Liao is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan.
E-mail: peegoo77@gmail.com.

Liang-Chi Hsieh is with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan. E-mail: viirya@gmail.com.

Yan-Ying Chen is with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan.
E-mail:yanying@cmlab.csie.ntu.edu.tw

Winston H. Hsu received the Ph.D. degree from the Department of Electrical Engineering, Columbia University, New York, NY. He is an Associate Professor in the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, since September 2011. Prior to this, he was in the multimedia software industry for years. He is also with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. His research interests include multimedia content analysis, image/video indexing and retrieval, machine learning and mining over large-scale databases.