Multigranular Event Recognition of Personal Photo Albums

Cong Guo, Xinmei Tian, *Member, IEEE*, and Tao Mei, *Senior Member, IEEE*

***Abstract*—People are taking more photos than ever before in recent years. To effectively organize these personal photos, the photos are usually assigned to albums according to their events. An efficient way to manage our photos would be if we could recognize the events of the albums automatically. In this paper, we study the problem of recognizing events in personal photo albums. Recognizing events in photo albums is a new challenge since the contents of photos in albums are more complicated than in traditional single-photo tasks, since not all photos in an album are relevant to the event and a single photo in an album often fails to convey the meaningful event semantic behind the album. To solve this problem, we introduce an attention network to learn the representations of photo albums. Then, we adopt a hierarchical model to recognize events from coarse to fine using multigranular features. We evaluate our model on two real-world datasets consisting of personal albums; we find that our model achieves promising results.**

***Index Terms*—Photo album, event recognition, attention network, hierarchical structure.**

I. Introduction

With the fast development of cameras and mobile devices, people are taking more photos than ever before. It was reported that there were about 1.6 trillion photos taken annually in 2013 [1]. The explosive growth of digital photos leads to a growing need for tools to automatically manage them. Usually, in consumer photo albums or online social networks, photos are organized in albums according to their events. However, it will cost a lot of time for users to label their photo albums. To solve this problem, automatic event recognition in photo albums is highly demanded.

There are already many works which focus on single-image understanding. However, only a few works pay attention to event recognition in photo albums. In general, photos will represent highly relevant visual content for a specific object or scene in single-photo recognition tasks. These photos are "typical" to the events since we can directly recognize an event with only a single photo. Extracting features that precisely represent specific visual contents can facilitate understanding these photos. Compared with single-photo recognition tasks, photos in albums have several unique properties: 1) Personal albums often consist of "relevant" and "non-relevant" photos. In most cases, the event-level semantics can not be concluded based on a single photo and must be determined based on the whole album. The "relevant" photos can only describe parts of the event. The "non-relevant" photos are the ones which contents are not related to the events. Some examples are given in Fig. 1. 2) Even given the same event label, albums exhibit significant variation in their content composition since photographers have their own photo styles and focus points. 3) Different events may include the same visual contents. 4) Photographers sometimes need to take a few extra photos to ensure that they obtain the perfect focus. Thus, there exists a great deal of information redundancy among the photos of albums. 5) An event is a higher level concept than other subjects such as objects and scenes. Therefore,

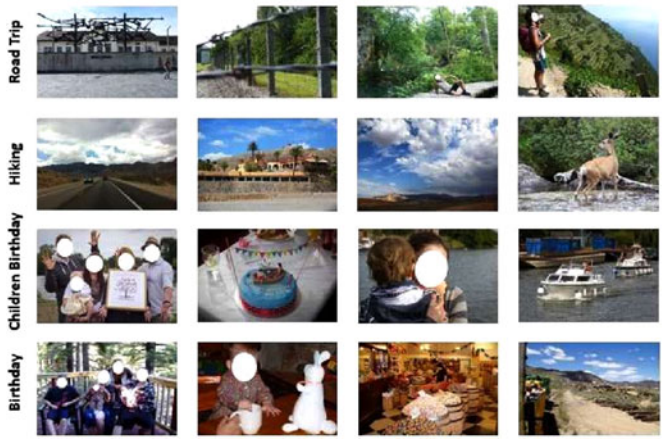


Fig. 1. Examples of photos in personal albums in the PEC dataset [2], where each row corresponds to an event. Photos in the first two rows are from "Road Trip" and "Hiking". The photos in them share some ambiguous contents, including roads and mountains. The Photos in the last two rows are from "Children Birthday" and "Birthday". If only given a family photo with babies, it is difficult to distinguish "Birthday" from "Children Birthday". Personal albums usually consist of "relevant" and "non-relevant" photos. For example, roads in "Road Trip", mountains in "Hiking" and a cake in "Children Birthday" and "Birthday" are "relevant" photos. The yachts in "Children Birthday" and the road in "Birthday" are "non-relevant" photos since their contents are not related to the subject of the events.

Manuscript received December 5, 2017; revised July 7, 2017 and September 2, 2017; accepted November 2, 2017. Date of publication November 24, 2017; date of current version June 15, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002203; in part by NSFC under Grant 61572451, Grant 61390514, and Grant 61632019; in part by the Youth Innovation Promotion Association CAS under Grant CX2100060016; and in part by the Fok Ying Tung Education Foundation under Grant WF2100060004. *(Corresponding author: Xinmei Tian.)*

C. Guo and X. Tian are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, University of Science and Technology of China, Hefei 230027, China (e-mail: gcong18@mail.ustc.edu.cn; xinmei@ustc.edu.cn).

T. Mei is with Microsoft Research, Beijing 100080, China (e-mail: tmei@microsoft.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.2017.2777664

it will be helpful to understand an event from various perspectives. These features render it more difficult to distinguish an event within personal photo albums than within individual photos. Fig. 1 shows some examples of photos from personal photo albums in the PEC dataset [2].

In this paper we study the event recognition problem in personal photo albums from three aspects: photo-level, album-level and event-level. Photographers record their lives by a serious of photos. These photos will be managed to albums according to their events. Understanding the photos within each album is the basic to understand the events. Photo-level features from multi-views can help us understand the photo comprehensively. Here we mainly consider two types of features: the taken times extracted from the meta data and the visual contents extracted from the images. The time feature has proven to be very effective for event recognition [3]. Certain events often occur at certain times such as hiking on weekends and concerts at night. For visual content feature, we extract features from different pre-trained deep convolutional neural networks (CNNs). In recent years, CNNs have achieved outstanding performances in object and scene recognition for a single photo [4]–[6]; it has also been shown that high-level CNN features can be transferred to generic recognition tasks without fine-tuning [7]. We utilize the learned representations extracted by three CNNs: a CNN trained on ImageNet [8] that can adequately represent the features of objects, a CNN trained on the Places database [9] that can describe the scenes, and a CNN trained with user-contributed attributes that can describe the frequently used attributes on Flickr [10].

However, the event-level semantics of albums can not be concluded based on a single photo and must be determined based on the whole album. Using the album-level features is a good way to recognize the events since they contain complete information of the events. In our previous work [11], we proved that the use of global average features is much better than composing the predictions of a single photo. However, this naive average strategy assumes that each photo in an album makes the same contributions to the recognition. In reality, albums usually consist of "relevant" and "non-relevant" photos, as shown in Fig. 1. The "non-relevant" photos may be useless and should make minimal contributions. Thus, if we can better focus on the "relevant" photos and assign them higher weights while paying less attention to "non-relevant" photos and assigning them lower weights, the weighted average features can better describe the albums compared to the previous method. Inspired by the attention model on image question answering (QA) [12], we introduce a new attention network to learn the attentions for albums considering both the semantic meanings of event labels and the global integrity of the albums. This attention network learns to pay more attention to the "relevant" photos and pay less attention to the "non-relevant" ones to generate album-level features for event recognition. Based on the intuition that not all events are equally difficult to recognize, we first build a coarse classifier to classify the easily separable events. In this paper, we use our attention network with the scene-CNN features as the coarse classifier. Then, we use the Affinity Propagation algorithm [13] to generate the coarse event clusters. Within each coarse cluster, we train four fine classifiers. Three of them are our attention network trained using the three CNN features respectively. For the last one, we train a SVM classifier with the time features [14]. Late fusion strategy is adopted to combine the results of these four fine classifiers. To combine predictions of the coarse and fine classifiers, a probabilistic averaging method is proposed to get the final results.

To the best of our knowledge, there is only one large dataset available, known as the PEC dataset [2], for studying the challenge of event recognition in personal photo albums. The limited dataset may be insufficient to evaluate the models. Therefore, we collect another large dataset containing 79,370 photos in 1,210 albums with 22 event classes from Flickr. The photos are all taken from users' daily lives.

In summary, this paper introduces the following contributions:

1) We introduce an attention network to learn album-level feature representation for event recognition.
2) We build a hierarchical model with multi-model features for event recognition, including three CNN features and the time feature.
3) Our model achieves promising performance on two real-world personal photo album datasets.

Compared with our previous work [11], we mainly make extensions from four aspects. First, we replace the ImageNet-trained AlexNet feature with the ImageNet-trained VggNet feature to get better descriptive ability. Second, we add a new attribute-based feature to help us understand the photos in albums. The CNN is trained with user-contributed attributes which describe the frequently used attributes on Flickr. Thus we use features from multi-view and multi-granular. Third, although all photos in the album may contribute to event recognition, they are not equally meaningful or important. Therefore, we introduce a new attention network which can automatically learn to pay more attention to the "relevant" photos and pay less attention to the "non-relevant" photos. Fourth, we collect another large dataset for event recognition in personal photo albums and conduct extensive experiments on it. The experiments show that our model achieves promising results on both datasets.

The remainder of the paper is organized as follows. Section II presents related works. Section III introduces the features, our attention network for album recognition and our coarse-to-fine model. A new dataset and the experimental results are given in Section IV, followed by the conclusion in Section V.

II. Related Work

In recent years, people are taking more photos than ever before. This leads to a growing need for tools to manage them. Recognition [8], [15] and retrieval [16], [17] are two common methods to organize these amount of photos. For personal photos, photos are always grouped to albums according to their events. In this paper, we mainly focus to recognize events in personal albums.

Deep learning has shown satisfactory performance in computer vision and has become the most popular approach to pattern recognition. With the help of the large-scale visual recognition ImageNet dataset, many CNN architectures can

recognize objects in our daily lives [8]. For scene recognition, CNN models trained with the Places database [15] have shown a promising performance. To improve the performance of classification, many methods, such as hierarchical structures [18], [19], [35], instance learning [16] and decision trees [20], have been considered. Similar to object/scene recognition, to recognize events within a single photo, "typical" photos of different events are collected for experiments. In [21], photos from 50 different cultural events were crawled, and visual features extracted from CNNs with time information were used for classification. In [22], eight sporting event categories were collected from the Internet, and the researchers attempted to recognize the events by integrating scene and object categorization. Mattivi *et al.* used time clustering information to improve the sub-event recognition in an efficient bag-of-features classification approach [23]. However, the photos of personal albums are not all "typical" photos, and an event usually cannot be recognized from only a single photo in the album.

Another significant area of research for recognition tasks attempts to recognize the actions in videos. As we know, the key frames extracted from videos can be approximatively viewed as albums of photos. One type of solution requires the contents of the videos to be time continuous [24]–[26], and such solution is not suitable for event recognition for photo albums. Other solutions have attempted to solve the problem based on key frames. In [27], [28], the researchers attempted to find the most suitable number of frames for recognizing the events in videos. Izadinia and Shah [29] modeled the joint relationship between the low-level events in a graph and used this graph to train their classifier with a latent SVM formulation. Recognition from albums is different from recognition from videos. Videos are usually very short, often a few seconds, and the contents are not as diverse as in photo albums.

Album event recognition is more complex than videos and a single photo. Since photographers have numerous styles for taking photos, the photos in albums are much more diverse. It is difficult to find the "typical" photos in albums. In most cases, a single photo in albums can only describe part of the events, and we may need to browse many photos to determine what event occurred in the album.

To tackle the challenging problem of event recognition in albums, researchers have applied various methods. In [30], the tags that users used for annotation were adopted to build a tag similarity graph for detecting events. In [31], typical objects that were highly related to the events are pre-defined to help recognize the events. In [3], GPS location information was utilized. In [32], an album-level classifier was trained by manually selected photos. A Stopwatch Hidden Markov Model, which considered the time gap between photos and sub-events, were introduced for album event recognition [2]. This model treated the sub-events as latent, and each photo was associated with a sub-event. However, it is difficult to assign photos to their correct sub-events because of the varying contents of personal photos. In [33], the authors proposed to learn features from sets of labeled raw images in personal photo albums. They randomly picked several photos from albums, extracted features and summed these features for classification. This method is similar to the average feature method, which assumes that all photos should make equal contribution for the event recognition.

Different from existing methods, in this paper we propose a hierarchical model to recognize events of albums from coarse to fine. To better understand the photos in albums, we extract multimodal features. To obtain album-level feature representation, we propose a new attention model to pay more attentions to "relevant" photos. Coarse and fine classifiers are combined to get the final predictions.

III. Multi-Granular Event Recognition

The overall architecture of our model is shown in Fig. 2. We will introduce the three major components of our model in this section: the multiple features, the attention network and the coarse-to-fine hierarchical structure, which attempt to understand the albums from image, album and event levels, respectively. To recognize personal albums, we first need to understand the photos. We can do this from the multi-view perspective: objects, scene, user-contributed attributes and the taken times of the photos. Then, for the albums, predictions from album-level features are much more better than the aggregated ones from a single photo. However, simple averaged features contain many information from the "non-relevant" photos which may be useless for recognition. To filter out these irrelevant photos, we introduce an attention network. Finally, a hierarchical structure is adopt to help us understand the events from coarse to fine.

A. Feature Representation

1) Image-Based Representation: In personal albums, certain typical objects are highly relevant to certain events such as a cake being relevant for "Birthday", a bachelor's gown being relevant to "Graduation", and Jack-o-lanterns being relevant to "Halloween". If we could find these typical objects, they would be helpful for recognizing events in the albums.

In recent years, deep learning has become the most popular approach to pattern recognition. With the help of the large-scale ImageNet dataset, deep models have been able to achieve promising recognition performance for object classification and can be extended to generic recognition tasks without fine-tuning [7]. Therefore, we adopt the 4096-dimensional fc7 features from the VggNet, which is pre-trained on the ImageNet dataset [5].

2) Scene-Based Representation: To recognize events in an album, the backgrounds in photos are also very important. Sometimes, we can recognize the events simply by browsing the background information from the photos. For example, a "church" may appear in "Wedding" events, and a "gallery" may appear in "Exhibition" events. A CNN model trained on the Places database [9] has shown positive performance in scene recognition [15]. We utilize this CNN and extract the 4096-dimensional features from the fc7 layer for each photo.

3) Attribute-Based Representation: High-level describable attributes of images are useful for people to recognize what is occurring in a photo stream. Users always assign a list of

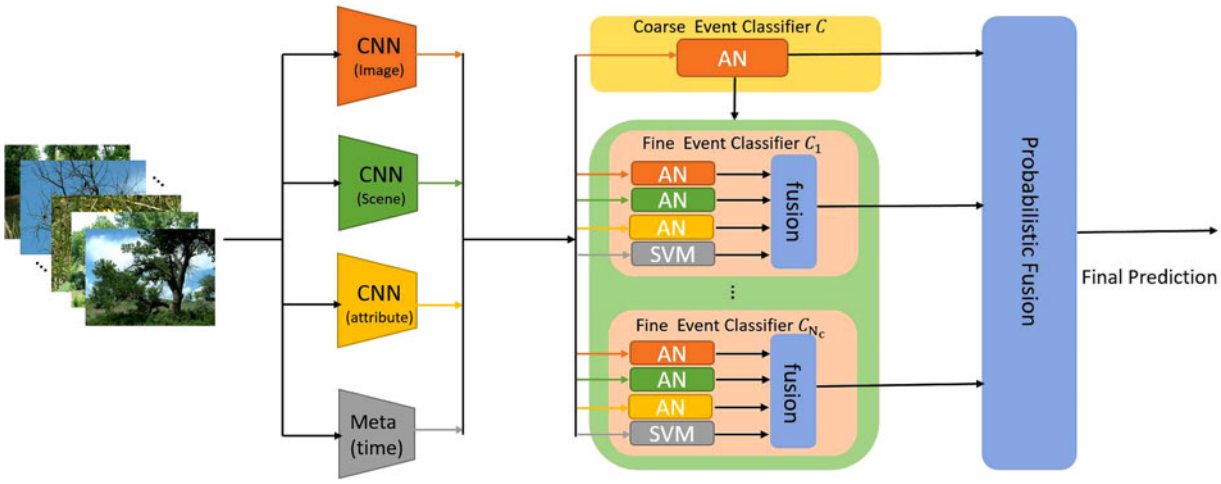


Fig. 2. Our model for album event recognition. For an album, we extract four kinds of features for the photos, including three CNN features from image contents and time features from the meta data. Then, a coarse-to-fine structure is adopted. AN units are the attention networks which try to predict the events with more attention to the "relevant" photos and less attention to the "non-relevant" photos. AN and SVM units work as classifiers. We have one coarse event classifier and many fine event classifiers. The colors of the lines indicate the flow of the data. We use our attention network for the CNN features and use SVM for the time features. We obtain the final results by combining the predictions from the coarse and fine classifiers with a probabilistic averaging method.

attributes (tags) to their photos on Flickr. A CNN model is trained with these user-contributed attributes [10]. We extract these attributes to help us understand the events of albums.

4) Time Feature: The time feature has proven to be very effective for event recognition [3] since events are frequently associated with a certain time of day. For example, "Christmas" is associated with December, and a "Graduation" is often held in summer. We extract the time information from the meta data of the photos. We build the time into a 6-dimensional feature vector for each photo. For the photo-level time feature, we transfer the timestamps to year, month, day, hour, day of the week, and week of the month. For the album-level time feature, we compute the duration for each album.

B. Attention Network

In contrast to event recognition for a single "typical" photo, events in albums are complicated. Personal albums often consist of "relevant" and "non-relevant" photos. The "relevant" photos can only describe part of the events, and we usually need several "relevant" photos to recognize one event. The "non-relevant" photos are the ones that their visual contents are far away from the subjects of the events. In this part, we introduce our attention network to learn the representations of albums, which attempt to better focus on the "relevant" photos while paying less attention to "non-relevant" photos.

With recent advancements in computer vision, attention networks have achieved promising results in image question answering (QA) tasks [12]. Paying close attention to the relative regions instead of the whole image can effectively filter out unimportant information. Inspired by this work, in our album event recognition task, the attention target becomes an album instead of an image. We attempt to focus on the relative photos instead of regions. However, the attention networks need an explicit question as a reference to guide the network find the regions. In our recognition task, we do not have such additional information.

To solve this problem, we introduce a new attention network, as shown in Fig. 3. From our previous work [11], we know that the average features of albums can well represent the albums. Such a global feature contains the visual contents of not only "relevant" photos but also "non-relevant" photos. In addition, we have an event label for each album. The event labels contain high generalizations of semantic meanings to describe albums. Thus, if we could embed the global visual features and the event textual labels to the same latent space and let them have a similar distribution, then the latent visual features can act as references to guide the network in finding the "relevant" photos.

We can extract the CNN features from the CNN networks.

$$v_I = CNN(I) \tag{1}$$

$$v_A = CNN(A) \tag{2}$$

$$v_{avg} = \frac{1}{n_A} \sum_{I \in A} CNN(I) \tag{3}$$

where I denotes a photo, A denotes an album, and n_A is the number of photos in album A. $v_I \in \mathbb{R}^d$ and d is the dimensionality of the CNN feature. $v_{avg} \in \mathbb{R}^d$ is the average feature vector of album A. $v_A \in \mathbb{R}^{d \times m}$ and m is the number of photos selected from A. In this paper, we set m to 64. For albums which contain more than 64 photos, we randomly select 64 photos to extract v_A. For albums which have less than 64 photos, we extract feature vectors for all photos and set 0 to the rest.

For an album event label, we embed the labels in a vector space through a word embedding tool, word2vec [34].

$$v_y = \text{word2vec}(y),$$

where $v_y \in \mathbb{R}^{300}$.

Given the features v_A, the average album feature v_{avg} and the label semantic feature v_y, our attention network first learns

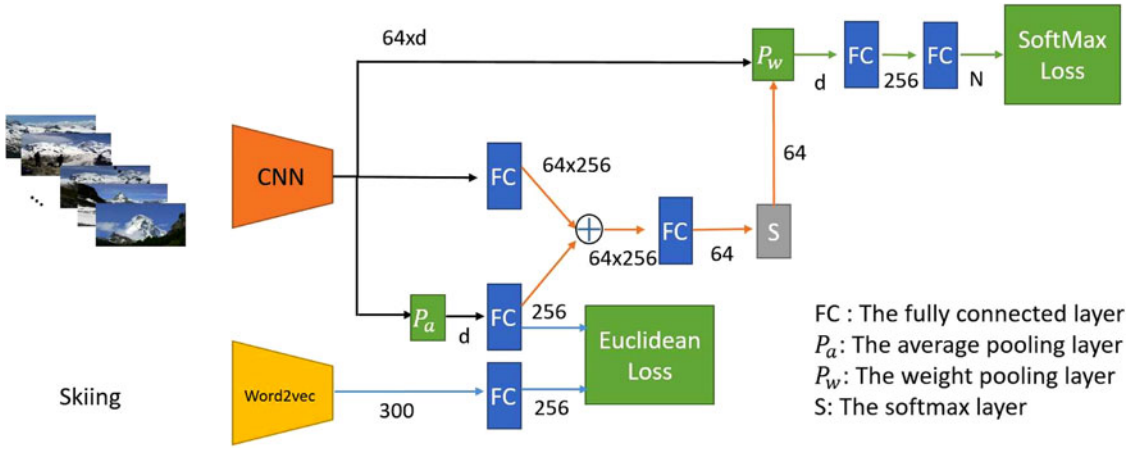


Fig. 3. The attention network in our album event recognition model. d is the dimensionality of the feature vectors extracted from the CNN network. In our experiments, we have three kinds of CNN features. For the image-based and scene-based features, $d = 4096$. For the attribute-based features, $d = 272$.

a single-layer neural network to embed the feature in a latent space. Within this space, we want the latent visual features to have similar distributions as the latent semantic label features. Then, a softmax function is adopted to generate the attention distribution for each photo. We use the Euclidean loss as the first loss term in our attention network.

$$L_C = \| W_A v_{avg} - W_y v_y \|_2^2 \tag{4}$$

$$h_l = \tanh(W_A v_A \oplus W_A v_{avg}) \tag{5}$$

$$p_a = \text{softmax}(W_P h_l + b_P) \tag{6}$$

where $W_A \in \mathbb{R}^{k \times d}$, $W_y \in \mathbb{R}^{k \times 300}$ and $W_P \in \mathbb{R}^{1 \times k}$. $p_a \in \mathbb{R}^m$ is an m-dimensional vector which represent the attention probability of each photos. Here, the symbol $\oplus$ denotes the addition of a matrix and a vector by adding each column of the matrix by the vector.

According to the attention distributions, we can calculate the weighted average features of the album. Then, we feed them through two inner-product layers and a softmax function to obtain the final event predictions. We use the softmax with loss function for classification to be our second loss term.

$$\tilde{v}_A = \sum_{i=1}^{m} p_a(i) v_{I_i} \tag{7}$$

$$P(A) = \text{softmax}(W_2(W_1 \tilde{v}_A + b_1) + b_2) \tag{8}$$

$$L_S = -\sum_{i=1}^{N} log(P_{E_i}(A)) \tag{9}$$

where $p_a(i)$ is the weight for photo I_i in album A and v_{I_i} is the feature vector of image I_i. $\tilde{v}_A \in \mathbb{R}^d$ is the weighted average feature vector of album A. $P(A) \in \mathbb{R}^N$ is the prediction for the N-class events. $P_{E_i}(A)$ is the probability that album A is predicted to be event E_i.

We adopt the joint supervision of the two losses to train our attention network for album event recognition,

$$L = L_S + \mu L_C. \tag{10}$$

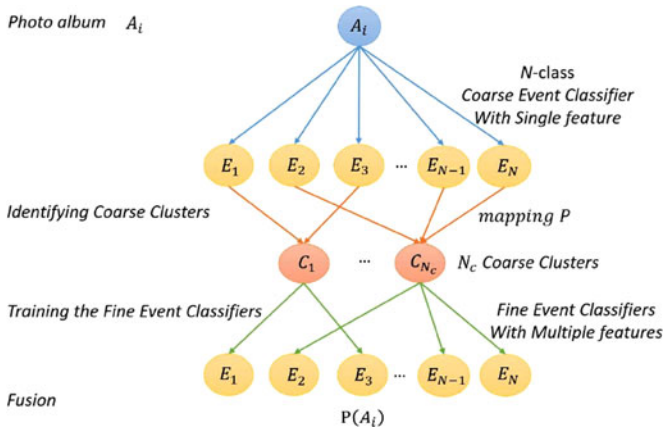


Fig. 4. Coarse-to-fine hierarchical model. The blue lines indicate the prediction of the coarse event classifier. The orange lines represent the mapping P. The green lines are the fusion predictions of the fine classifiers with multiple features.

C. Coarse-to-Fine Hierarchical Structure

Since not all albums are equally difficult to be recognized, recognizing the albums from coarse to fine with multiple features can be an effective way. Our coarse-to-fine hierarchical model has three steps. First, a coarse classifier is trained using our attention network with scene-CNN features, to identify the easily distinguishable event clusters. Second, for each cluster, we train four fine classifiers. Three of them are our attention network trained using the three CNN features respectively. For the last one, we train a SVM classifier with the time features [14]. Third, a probabilistic averaging component is applied to combine the predictions of the coarse and fine classifier and forms the final predictions. We present our hierarchical model in Fig. 4.

1) Coarse Event Classifier Training: Our hierarchical model starts from separating easily discernible events. In fact, humans usually use the visual information of the photos to recognize the coarse event clusters by a simple glimpse, and then

carefully browse the photos to further identify the events. We mimic this natural process to find a visual descriptor for our coarse event classifier. In our model, we utilize the scene-based features extracted from CNN trained on Places dataset and use our attention network for classification. We split our training data into two parts: $train\_train$ part and $train\_val$ part. We use the $train\_val$ part as a validation set to determine the parameters of the coarse event classifier.

2) Coarse Events Clustering: We can obtain the confusion matrix $F \in \mathbb{R}^{N\times N}$ from our coarse classifier. N is the number of events. We first make F symmetric by computing $F = \frac{1}{2}(F + F^T)$, and then adopt the Affinity Propagation algorithm [13] to cluster the N events into N_c coarse clusters. Meanwhile, the mapping $P : y \longmapsto y'$, which records the mapping relationship between fine events and coarse event clusters can be achieved. The probability that album A_i is predicted into a coarse cluster C_j can be calculated by

$$B_{i,j} = \sum_{E_k \in C_j} P^c_{E_k}(A_i), \tag{11}$$

where $P^c_{E_k}(A_i)$ is the probability that album A_i is predicted to be event E_k by the coarse classifier and it can be obtained from (8).

We adopt Affinity Propagation since it does not need to define the number of clusters. And, the clusters are more balanced in size than under other clustering algorithms. The only parameter, damping factor λ, is set to 0.5 by default throughout the experiments.

3) Fine Event Classifiers Training: Within each coarse cluster, fine event classifiers are trained with various features independently. For the multiple CNN features, we follow the same training process of the coarse event classifier. For the time features, we utilize LibSVM [14] with the RBF kernel. Then, we utilize a weighted average function to combine the predictions from the coarse and fine classifiers.

$$P(A_i) = \sum_{j}^{C} B_{i,j} P_j(A_i). \tag{12}$$

where $B_{i,j}$ is the probability that album A_i is predicted into the coarse cluster C_j and $P_j(A_i)$ is the prediction made by the fine classifier trained in the coarse cluster C_j.

We obtain different predictions for the multiple features by (12). Next, we adopt late fusion to combine the different predictions by

$$\begin{aligned} P_{\text{final}}(A_i) = {} & \alpha \times P_{\text{scene}}(A_i) + \beta \times P_{\text{image}}(A_i) \\ & + \gamma \times P_{\text{attribute}} + (1 - \alpha - \beta - \gamma) \times P_{\text{time}}(A_i). \end{aligned} \tag{13}$$

The late integration fusion weights are empirically selected by an exhaustive search and determined when the integrated predictions achieve the best performance on the $train\_val$ part.

IV. Experiment

In this section, we first introduce the personal album datasets for event recognition. Then, we present the experimental settings as well as comparison methods, followed by the performance of different approaches and analyses.

TABLE I
Statistics of the PEC Dataset [2]

Event	#Albums	#Photo
Birthday	60	3,227
Children Birthday	64	3,714
Christmas	75	4,118
Concert	43	2,565
Boat Cruise	45	4,983
Easter	84	3,962
Exhibition	70	3,032
Graduation	51	2,532
Halloween	40	2,403
Hiking	49	2,812
Road Trip	55	10,469
St. Patricks Day	55	5,082
Skiing	44	2,512
Wedding	69	9,953
Total	807	61,364

A. Dataset

To the best of our knowledge, the biggest existing dataset for personal album event recognition is the PEC dataset [2]. The dataset contains 807 albums within 14 event classes. All the photos are crawled from Flickr. The events are defined by the most popular tags on Flickr, Picasa and Wikipedia. The statistics of the dataset are shown in Table I. Albums for training and testing have already been defined in [2].

Moreover, we collected some albums from Flickr to build another dataset for event recognition. We first created a Flickr account and followed a number of users. Then, a breadth-first search method was used based on the users' contact lists, ultimately gathering 300,000 users. We crawled their photosets' titles, split the titles into words, manually chose the 22 most popular events that corresponded to social events according to the frequencies of keywords and downloaded the photo sets containing these keywords. However, some photo sets were irrelevant to their events or consisted of more than one event; we removed these photo sets manually. To make the dataset balanced, we randomly selected photo sets within each event and made every event have the same number of photo sets. Finally, we collected 1,210 albums with 79,370 photos. We randomly chose 45 albums of each event to be the training part, and the remaining albums were the testing part. We named this dataset Flickr-22. The event classes and the statistics of the dataset are shown in Table II.

For both datasets, we randomly choose 20% training set as $train\_val$ set and the rest as $train\_train$ set to determine model parameters in the following experiments. We use average precision, average recall and average F_1-score to evaluate the performance of different recognition methods. The average recall is the same to the average accuracy in [2]. The average F_1-score is obtained by average the F_1-scores of all event classes.

TABLE II
STATISTICS OF THE FLICKR-22 DATASET

Event	#Albums	#Photo
baseball	55	3,882
basketball	55	4,117
birthday	55	3,754
Christmas	55	3,253
concert	55	2,795
cruise	55	5,101
Easter	55	2,986
exhibition	55	3,194
football	55	4,748
graduation	55	3,664
Halloween	55	2,494
hiking	55	3,314
parade	55	3,641
party	55	2,677
skate	55	3,618
skiing	55	3,137
soccer	55	4,091
surfing	55	2,507
Thanksgiving	55	2,254
travel	55	6,614
wedding	55	4,449
zoo	55	3,580
Total	1,210	79,370

B. Experimental Settings

For the scene-CNN features, image-CNN features and attribute-CNN features, the features extracted from the CNN networks are directly used as inputs of our attention network. For the time features, min-max normalization is adopted for each photo-level time feature, whereas we scale the album-level time features, to the size of the day. Then, we average the time features and utilize LibSVM [14] to perform the experiments on the time features with the RBF kernel.

When training the attention networks, we first use the average features to train the last two fully-connected layers for classification with a dropout of 0.5. We start the learning process with the default parameters as $base\_lr = 0.01$, $gamma = 0.1$, $momentum = 0.9$ and $weight\_decay = 0.0005$. Then, we train the whole attention network. We update the weights with a mini-batch size of 64. μ is determined to be 0.1 via the $train\_val$ set. The number of input photos for each album is set to 64 for all the three CNN features. As shown in Fig. 5, we show the accuracies of our attention model for all the 14 categories on the $train\_val$ set with the three CNN features by different m. We can see that as the m increases, all the accuracy curves rises. However when m is greater than 64, the curves will slow down. Considering the size of training batches as well as the performance, we choose $m = 64$ for all three CNN features. We learn the parameters of the our attention network with scene-CNN features and extend them to the other two CNN features. We utilize LibSVM [14] to perform the experiments on the time features with RBF kernel. The parameters of SVM are determined via the $train\_val$ set. After obtaining the predictions of fine classifiers, the late integration fusion weights are empirically selected by an exhaustive search in each coarse event cluster on the $train\_val$ set.

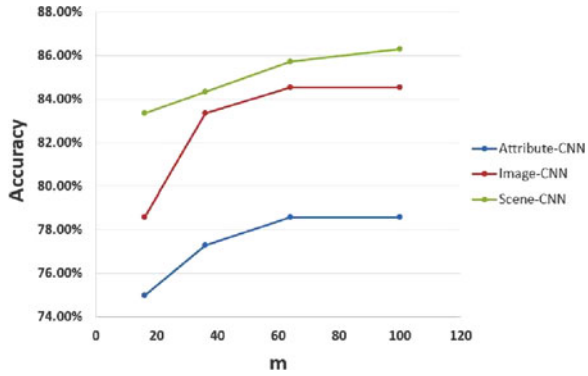


Fig. 5. Accuracies with different m for different features on the train_val set in PEC dataset.

TABLE III
PERFORMANCE OF DIFFERENT METHODS ON THE PEC DATESET

Method	Avg. Precision. (%)	Avg. Recall (%)	Avg. F_1-Score (%)
AgS [2]	-	41.43	38.87
ShMM [2]	-	55.71	56.16
Wu [33]	-	73.43	71.56
H-AvS [11]	86.32	85.00	84.85
AgS-CNN (Attribute)	64.36	54.29	50.25
AgS-CNN (Scene)	73.31	70.71	67.05
AgS-CNN (Image)	69.74	70.00	66.48
AvS-CNN (Attribute)	79.23	77.14	76.94
AvS-CNN (Scene)	82.17	79.29	78.81
AvS-CNN (Image)	82.17	80.00	80.06
LF-AvS	85.95	83.57	83.06
FC (Attribute)	79.41	77.86	77.63
FC (Scene)	82.23	80.00	78.70
FC (Image)	83.59	80.00	79.94
LF-FC	85.08	82.86	82.15
AN (Attribute)	81.59	80.00	79.67
AN (Scene)	86.38	83.57	82.36
AN (Image)	86.64	81.43	81.03
LF-AN	87.49	85.00	84.15
H-AN	90.07	87.86	87.70

When testing, given an album with n_A photos, we first order the photos by time. Then, we split the album to $\lceil \frac{n_A}{64} \rceil$ parts by equidistance sampling to guarantee that each parts can approximately cover the whole album. The label information will not be used. We test them respectively and make an average to generate the final predictions.

C. Approaches for Event Recognition

In this section, we present the representation methods and the classification methods for comparison.

1) Aggregated SVM (AgS): We use the baseline mentioned in [2] as one of our baseline methods. We train a linear multi-class SVM for the photo-level recognition. Each photo inherits the label of the album to which it belongs. We sum the confidence scores of the photos in the albums and choose the events with the highest scores as the final predictions.

Fig. 6. Examples of photos and their weights in personal albums in the PEC dataset [2]. Each row corresponds to an event and the numbers under the photos are their weights assigned by the attention network. Weights in (a) and (b) are obtained with Scene-based CNN features and Image-based CNN features respectively. The first two columns in (a) and (b) display the photos with the two largest weights in each album while the last two columns show the photos with the two smallest weights.

TABLE IV
PERFORMANCE OF DIFFERENT METHODS ON THE FLICKR-22 DATASET

Method	Avg. Precision. (%)	Avg. Recall (%)	Avg. F_1-Score (%)
H-AvS [11]	88.13	86.82	87.03
AvS-CNN (Attribute)	81.29	78.64	79.11
AvS-CNN (Scene)	86.39	85.00	85.16
AvS-CNN (Image)	86.29	85.00	85.12
LF-AvS	87.32	86.36	86.46
FC (Attribute)	80.79	79.09	78.81
FC (Scene)	84.84	84.09	83.86
FC (Image)	86.90	85.91	85.93
AN (Attribute)	80.54	79.55	79.53
AN (Scene)	86.78	85.91	85.95
AN (Image)	87.57	86.36	86.43
LF-AN	88.36	87.27	87.26
H-AN	90.89	89.55	89.60

2) Average SVM (AvS): In this approach, we first average the features within each album and then train a linear multi-class SVM at the album level.

3) Fully Connected Layer (FC): In this approach, we adopt the last two fully connected layers in our attention network with the average album features for classification.

4) Attention Network (AN): In this approach, we adopt our attention network for classification.

5) Late-Fusion (LF): We adopt late fusion to combine the classification confidence scores of different types of features.

6) Hierarchical (H): We use the full hierarchical model mentioned above. Scene-CNN features are used for the coarse classifier, and multiple features are used for fine classifiers.

D. Experimental Results

In this section, we present the performance of our model on two datasets.

TABLE V
THREE COARSE EVENT CLUSTERS AND THEIR FUSION WEIGHTS CORRESPONDING TO FOUR KINDS OF FEATURES IN PEC DATASET

Event Id	Scene-CNN	Image-CNN	Attribute-CNN	Time
C1: 2,3,4,5,6	0.20	0.20	0.15	0.45
C2: 7,10,11	0.35	0.00	0.15	0.50
C3: 1,8,9,12,13,14	0.05	0.05	0.60	0.30

1) PEC Dataset: We present the performance of different methods for personal album event recognition in Table III.

When comparing the baseline of the aggregated SVM with different features, the CNN features achieve higher accuracies than the low-level visual features in [2]. This proves that the high-level features extracted by the CNNs are substantially more powerful in event recognition than are the low-level features. Moreover, this also produces a comparable result with Wu's method [33].

The AvS and FC methods achieve better performance than the AgS method. This is because personal albums often consist of "relevant" and "non-relevant" photos. The "relevant" photos can only describe parts of the event. In most cases, the event-level semantics cannot be concluded based on only a single photo and must utilize the whole album.

However, not all the photos in an album are related to the events. They should not provide equal contributions to the recognition process. To obtain a better representation for albums, we introduce the attention network in Section III-B. The AN method achieves better performance than the FC method in general. It outperforms the FC method by 2.04%, 3.66% and 1.09% in terms of Avg. F_1-score for the attribute-CNN features, scene-CNN features and image-CNN features. This proves that our attention model can find the photos that are relevant to the events while filtering out the non-relevant photos. We show some examples by Scene-based CNN features and Image-based CNN

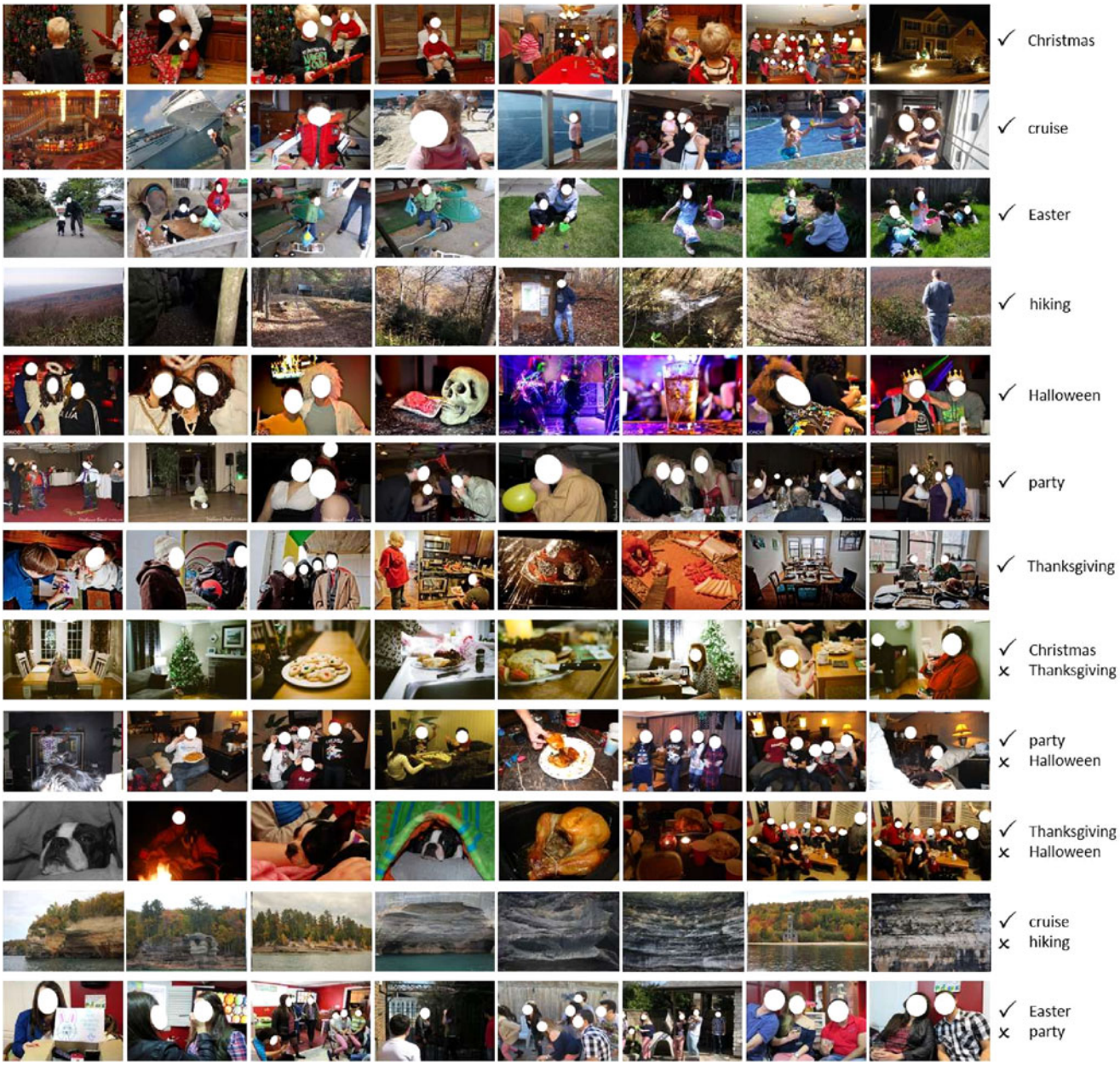


Fig. 7. Some classification examples from the Flickr-22 dataset. Each row corresponds to an event. Rows 1-7 show examples that the events are correctly predicted. The last 5 rows give some incorrectly predicted examples. On the right-hand side, the labels with ticks are the ground truth labels, and the labels with cross marks are incorrect labels predicted by our model.

features in Fig. 6. We can see that the attention network assigns reasonable weights to photos. It is notable that photos get different weights by different features. In the "hiking" event, the photo of branch gets small weight with the Scene-based CNN feature and it gets large weight with the Image-based CNN feature. Different features pay attention to different parts of albums. Therefore, by combining the three types of CNN features and the time feature, we obtain substantially better results than with a single feature.

Finally, we present the performance of our hierarchical model with the prediction scores from our attention network. Our full hierarchical model obtains the best performance among all the other methods and outperforms our previous work [11] by 3.75%, 2.86% and 2.85% in terms of Avg. Precision, Avg. Recall and Avg. F1-score respectively. In our hierarchical model, the scene-CNN features are used for the coarse classifier. Actually, we also have tried to use other two CNN features for coarse classifier training. We find that coarse classifier trained with scene-CNN feature achieves the best performance.

To further investigate how four kinds of features affect the final results, we list the fusion parameters in (13). As shown in Table V, the three coarse event clusters and their fusion weights are given. The events in each cluster are listed in the first column and the "Event Id" follows the order of events in Table I. We can

see that the fusion weights are quite different in three clusters. Time features are important for events in cluster 1 and cluster 2, since many of these events are associated with a certain time, for example "Christmas" in December, "Concert" at night and "Hiking" at weekend. Scene-CNN features play an important role in cluster 2. This is because events (e.g. "Hiking" and "Road Trip") in cluster 2 contain many background photos. Attribute-CNN features have a large weight in cluster 3. The possible reason is that events in this cluster contain highly diverse photos and high-level attribute features can help us to understand their contents.

For the time complexity, we have tested the time cost for four kinds of features extraction. It takes about 120 ms (Tesla K80) to extract all features for one photo. We have about 70 photos in each album. The attention network costs about 3 ms to generate a prediction for the Attribute-CNN features while costs about 12 ms for the Places-CNN features and ImageNet-CNN features (Tesla K80). Our hierarchical structure has one coarse classifier and nine fine classifiers for the CNN features. Thus, it will take about 100 ms for the prediction process. Therefore, it takes about 8.6 s in total to recognize the event of an album (including feature extraction). The time complexity is acceptable.

2) Flickr-22 Dataset: We present the performance of different methods in Table IV. After using our attention network mentioned in Section III-B, the CNN features with the AN method obtain a higher performance than with the FC method.

Other experiments draw and confirm the same conclusions mentioned in the analysis on the PEC dataset. Finally, we obtain the best average precision of 90.89%, the best average recall of 89.55%, and the best average F_1-score of 89.60% among all methods.

We show some event recognition examples from the Flickr-22 dataset in Fig. 7. We present five incorrectly predicted examples. To draw a comparison, we also present some other examples with the ground truth labels and predicted labels from the five incorrectly predicted examples. Next, we list three aspects that make event recognition in personal albums a challenge. First, as shown in Fig. 7, we predict a "Christmas" event as a "Thanksgiving" event in the 8th row. A "Christmas" event always consists of Christmas trees, presents, red hats and families, whereas a "Thanksgiving" event consists of turkeys, pies, long tables and also families, as shown in the 7th row. However, people have their own photo-taking habits. The photos in the 8th row display numerous pies and long tables. This makes our model make an incorrect prediction. Second, the confusion contents among different events represent a difficult problem to solve. We see that the visual contents of cruise and hiking photos can be very similar. Our hierarchical model attempts to address this problem. However, it will fail when the events appear to be highly similar. Finally, different people will do different things even at the same event. As shown in the 3rd and the last rows in Fig. 7, the two rows present two examples of an "Easter" event. A family will spend time with their children, whereas young people will have a get together similar to a party. The various contents of events represent another problem for event recognition in personal albums.

V. Conclusion

In this paper, we propose an attention network to learn the representations of albums and adopt a hierarchical structure for event recognition in personal photo albums. To learn the album-level representations, we borrow the attention networks from the image question answering task and introduce a new attention network for album event recognition considering both the sematic meanings of event labels and global representations of albums. Meanwhile, based on the assumption that not all albums are equally difficult to recognize, we build our coarse-to-fine model by first sorting easily discernible events into coarse clusters and then finely classifying them to obtain our final predictions. Multiple features, including the time feature, Image-CNN feature, Scene-CNN feature and Attribute-CNN feature, are introduced to help us better recognize the events in albums. Through a series of experiments, we find that predictions obtained using our attention network perform better than those obtained using non-weighted average features and are much better than aggregating the predictions obtained using single features. After employing our coarse-to-fine hierarchical model, we achieve the best performance among different methods on two real-world personal album datasets.

References

[1] D. Wakabayashi, "The point-and-shoot camera faces its existential moment," *Technology*, vol. 10, p. 59, 2013.

[2] L. Bossard, M. Guillaumin, and L. Van, "Event recognition in photo collections with a stopwatch HMM," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1193–1200.

[3] L. Cao, J. Luo, and T. S. Huang, "Annotating photo collections by label propagation according to multiple similarity cues," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 121–130.

[4] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," ACM Multimedia, 2014, pp. 675–678.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[6] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 1–9.

[7] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. 27th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[10] J. Fu *et al.*, "Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1985–1993.

[11] C. Guo and X. Tian, "Event recognition in personal photo collections using hierarchical model and multiple features," in *Proc. IEEE 17th Int. Workshop Multimedia Signal Process.*, 2015, pp. 1–6.

[12] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 21–29.

[13] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

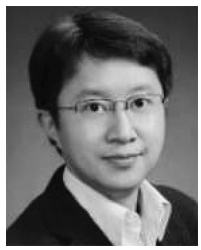
[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," arXiv:1412.6856, 2014.

[16] G.-J. Qi *et al.*, "Concurrent multiple instance learning for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2007, pp. 1–8.
[17] Y. Zhu, X. Huang, Q. Huang, and Q. Tian, "Large-scale video copy retrieval with temporal-concentration sift," *Neurocomputing*, vol. 187, pp. 83–91, 2016.
[18] H. Bannour and C. Hudelot, "Hierarchical image annotation using semantic hierarchies," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 2431–2434.
[19] Z. Yan *et al.*, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2740–2748.
[20] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1467–1475.
[21] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Giro-i Nieto, "Cultural event recognition with visual convnets and temporal models," CVPR Workshops, 2015, pp. 36–44.
[22] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[23] R. Mattivi, J. Uijlings, F. G. De Natale, and N. Sebe, "Exploitation of time constraints for (sub-) event recognition," in *Proc. Joint ACM Workshop Model. Representing Events*, 2011, pp. 7–12.
[24] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," ICCV, 2015, pp. 4489–4497.
[26] Q. Li *et al.*, "Action recognition by learning deep multi-granular spatio-temporal video representation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 159–166.
[27] Y.-G. Jiang, "Super: Towards real-time event recognition in internet videos," in *Proc. 2nd ACM Int. Conf. Multimedia Retrieval*, 2012, Paper 7.
[28] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2008, pp. 1–8.
[29] H. Izadinia and M. Shah, "Recognizing complex events using large margin joint low-level event model," in *Computer Vision–ECCV 2012*. New York, NY, USA: Springer-Verlag, 2012, pp. 430–444.
[30] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, "Cluster-based landmark and event detection for tagged photo collections," *IEEE Multimedia*, vol. 18, no. 1, pp. 52–63, Jan. 2011.
[31] S.-F. Tsai, T. S. Huang, and F. Tang, "Album-based object-centric event recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2011, pp. 1–6.
[32] F. Tang, D. R. Tretter, and C. Willis, "Event classification for personal photo collections," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 877–880.
[33] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1960–1968, Nov. 2015.
[34] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, 2010, pp. 45–50.
[35] Z.-J. Zha, T. Mei, Z. Wang and X.-S. Hua, "Building a comperhensive ontology to refine video concept detection," *Multimedia Inf. Retriev.*, 2007, pp. 227–236.

Cong Guo received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2013. He is currently working toward the PhD. degree at the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. His research interests include multimedia search, computer vision, and machine learning.

Xinmei Tian (M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, in 2005 and 2010, respectively. She is an Associate Professor with the CAS Key Laboratory of Technology, Geospatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. Her current research interests include multimedia information retrieval and machine learning. Prof. Tian was the recipient of the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation Award in 2013.



Tao Mei (M'07–SM'11) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a Senior Researcher with Microsoft Research, Beijing, China. He has authored or coauthored more than 100 papers in journals and conferences, 10 book chapters, and edited four books. He holds more than 15 U.S. granted patents and more than 20 in pending. His current research interests include multimedia analysis and retrieval, and computer vision. Dr. Mei was the recipient of several paper awards from prestigious multimedia journals and conferences, including IEEE Communications Society MMTC Best Journal Paper Award in 2015, IEEE Circuits and Systems Society Circuits and Systems for Video Technology Best Paper Award in 2014, IEEE Transactions on Multimedia Prize Paper Award in 2013, and Best Paper Awards at ACM Multimedia in 2009 and 2007, etc. He was the principle designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He is an Editorial Board member of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications*, *Machine Vision and Applications*, and *Multimedia Systems*, and was an Associate Editor of *Neurocomputing*, and a Guest Editor of eight international journals. Dr. Mei is the General Cochair of ACM ICIMCS 2013; the Program Cochair of ACM Multimedia 2018, IEEE ICME 2015, IEEE MMSP 2015, and MMM 2013; and the Area Chair for a dozen international conferences. He is a Senior Member of ACM and a Fellow of IAPR.