Event Recognition in Personal Photo Collections Using Hierarchical Model and Multiple Features

Cong Guo [1], Xinmei Tian [2]

Department of Electronic Engineering and Information Science, University of Science and Technology of China
Hefei, Anhui, China
[1] gcong18@mail.ustc.edu.cn
[2] xinmei@ustc.edu.cn

***Abstract*—With the proliferation of digital cameras and mobile devices, people are taking many more photos than ever before. The explosive growth of personal photos leads to problems of photo organization and management. There is a growing need for tools to automatically manage photo collections. Recognizing events in photo collections is one efficient way to organize photos. The use of textual event labels can allow us to categorize and locate an event without browsing through an entire photo collection. Most existing research on this topic focuses on recognizing events from single photos and only a few studies have examined event recognition in personal photo collections. In this paper, we propose a hierarchical model to recognize events in personal photo collections using multiple features, including time, objects, and scenes. Since some events are more difficult to identify and categorize, ambiguous events require fine event classifiers, while the coarse categories of the events can be sufficiently organized with a coarse event classifier. We evaluate our coarse-to-fine hierarchical model on a real-world dataset consisting of personal photo collections, and our model achieves promising results.**

I. Introduction

The proliferation of digital cameras and mobile devices has changed the world by allowing people to keep beautiful memories of their lives. It is estimated that 1.6 trillion photos are taken annually with smartphones, digital cameras, and other devices. This is a massive increase from the year 2000 when only about 100 billion photos were taken digitally [1]. The explosive growth of the photos leads to problems of photo organization and management. To find a specific event, people must often waste a great deal of time browsing through a huge number of photos. This inconvenience has lead to a growing demand for automatic event recognition in photo collections, which helps people retrieve specific photos through textual event labels instead of browsing all photos.

In recent years, many works in computer vision focus on understanding a single photo, while few focus on recognizing events in photo collections. In a single photo recognition task, the visual contents of a photo are highly relevant to a specific class/event label. Extracting features that precisely represent the specific visual contents will be helpful for understanding the photos. However, those methods do not work well for personal photo collections. Compared with the photos in single photo recognition tasks, photos in personal photo collections have several unique properties: 1) Personal photos are more general and closer to daily life. Not all of them are relevant to the events. 2) Events in collections show a very large variety in their content composition. Different events may share some common contents. 3) Events in personal photo collections are always composed of several sub-events. A single photo may only present part of the visual information from an event. Fig. 1 shows some examples of photos from personal photo collections. In contrast with single photo recognition tasks, personal photo collections contain a wide variety of scenes and objects, as well as many ambiguities. These features render it more difficult to distinguish an event within personal photo collections than within individual photos.

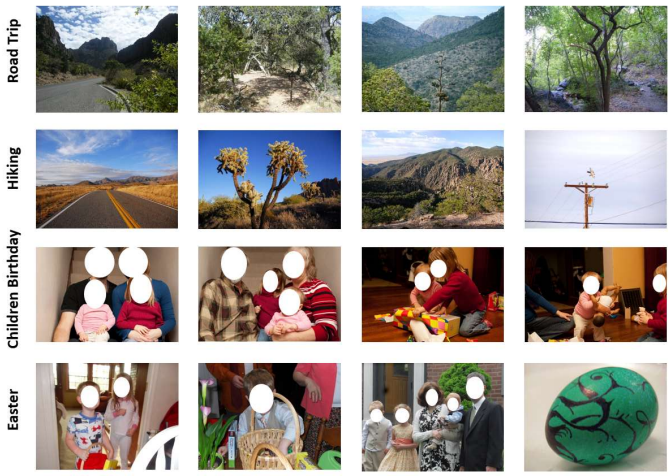


Fig. 1. Examples of photos in personal photo collections. Each row corresponds to a event. Different events may share some common contents. For example, photos of road trips and hiking may share ambiguous contents including trees and mountains, while photos of children's birthdays and Easter parties will present content that is similar to family portraits and baby faces.

To recognize events in personal photo collections, we mainly rely on the visual contents of photos. It is said that when we try to understand the world in a single glance, it takes only a few tens of milliseconds to recognize the category of an

MMSP'15, Oct. 19 - Oct. 21, 2015, Xiamen , China.
978-1-4673-7478-1/15/$31.00

object or environment with our brains [2]. As a result, we can take advantage of the visual contents from two perspectives: objects and scenes. Luckily, the existing deep convolutional neural network (CNN) architectures have already achieved outstanding performance in object and scene recognition for single photos. We take advantage of the features extracted by the CNNs: one is trained on ImageNet [3] and can adequately represent the features of objects, while the other is trained on the Place database [2] to describe the scenes. Some events often occur at certain times, such as hiking on weekends and concerts at night, so the time the photos were taken is an additional important feature for recognition.

In personal photo collections, events commonly share similar contents, which makes them difficult to separate from each other. To tackle this problem, we propose a hierarchical model that takes advantage of the coarse-to-fine method for event recognition in personal photo collections. 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. Here, we take advantage of the CNN features based on the Places database [2] to train the coarse classifier for coarse event recognition. After the events have been assigned to coarse clusters, information from the scenes is still insufficient to build the fine classifiers. We introduce two more features: CNN features for objects and time features. We train fine classifiers with the three features respectively and late fusion is used to get the fine predictions. A probabilities averaging method is adopted to combine the predictions of the coarse and fine classifiers to form the final predictions.

In summary, this paper introduces the following contributions:

- We build a hierarchical model for event recognition that performs well on a real-world personal photo collection dataset.
- We build a coarse classifier by using scene information from photos to separate distinct events. Multiple features are exploited for the fine classifiers to handle the complex and ambiguous contents.
- Collection-level visual features can more accurately predict the events from personal photo collections than the aggregated photo-level ones.

The rest of the paper is organized as follows. Section II briefly introduces related works. Section III presents our hierarchical model and the multiple features we used. Experimental settings and results are given in Section IV, followed by the conclusion in Section V.

II. RELATED WORK

Recently, a large body of research has focused on single photo recognition and has achieved satisfactory performance with the help of existing deep convolutional neural network(CNN) architectures. The major research on photo recognition can be divided into two categories: object recognition and scene recognition. For object recognition, the CNN trained on the ImageNet has proved to be a fine architecture [3]. And for scene recognition, the CNN trained on the Places database has shown promising performance [4]. Alternatively, we focus on recognizing the events in personal photo collections instead of merely detecting one object or scene from a photo.

Event classification in single photos has also been considered for years. Salvador et al. combined visual features extracted from convolutional neural networks with time information to automatically classify images from 50 different cultural events [5]. Imran et al. presented a novel approach to discover the most informative features for event recognition from each event category [6]. Researchers have also tried to classify the event in the image and provide a number of semantic labels to the objects and scene environment by integrating scene and object categorization [7].

There is also a considerable amount of research in the area of video event recognition. Key-frames extracted from videos can be viewed as collections of photos. In [8][9], researchers tried to select the most suitable number of frames to help recognize the event in the videos. Raptis and Sigal developed the latent key frames action model for recognizing human actions [10]. They modelled the videos as action stories-contextual temporal orderings of discriminant partial poses.

Event recognition within personal photo collections exhibits similar characteristics to both videos and single photos, but it is still more complex. In contrast to single photos and videos, personal photo collections always contain many ambiguous photos. Photos that clearly represent an event may only make up a small proportion of most collections. Therefore, it is difficult to recognize a collection's event from a single photo.

Recognizing the events of photo collections is a new challenge, though researchers have tried some methods to address the problem. Papadopoulos et al. presented a novel scheme that used the visual and tag similarity graphs for automatically detecting landmarks and events in tagged image collections [11]. A Stopwatch Hidden Markov Model, which took account of the time gap between photos, was also introduced for event recognition in photo collections [12]. In [13], a transfer learning method was adopted to obtain typical objects in events and then a classifier was trained for event recognition. Tang et al. proposed a probabilistic fusion framework to obtain a collection level prediction based on a classifier trained by the manually selected photos [14]. Cao et al. introduced a multi-level annotation hierarchy to address the problem of annotating consumer photo collections with additional meta information, such as GPS tracks [15].

III. HIERARCHICAL MODEL FOR PHOTO COLLECTION EVENT RECOGNITION

In this section, we introduce our hierarchical model and the multiple features we choose to use. For personal photo collection event recognition, each collection is labelled with a fine event label y and all photos in this collection share the same event label. However, not all photos in the collection are highly related to the event. This makes it hard to identify the events from a single photo. For this reason, we try to consider the photos in a collection as an integrated whole. We average the features of collections and build coarse-to-fine classifiers.

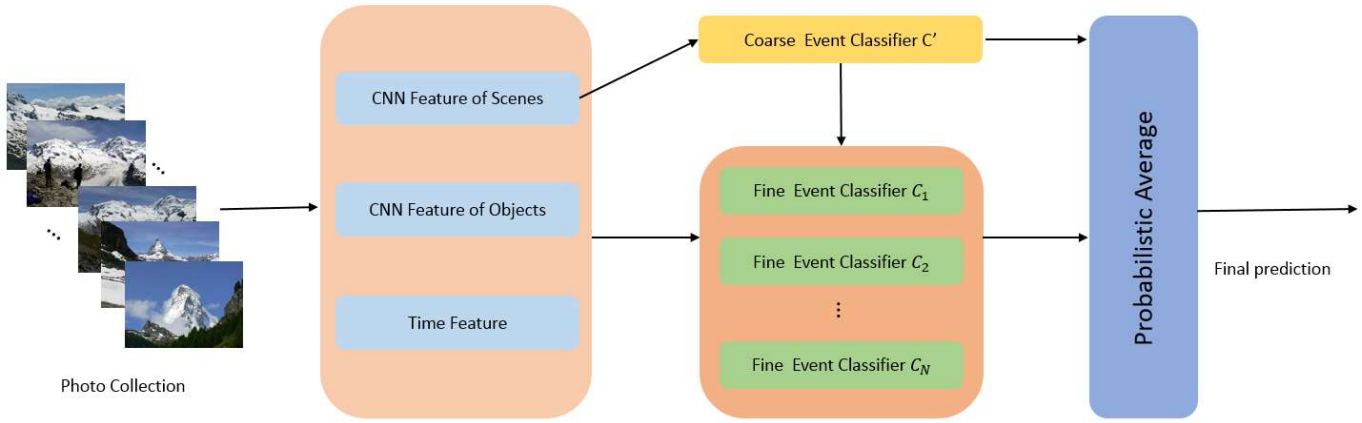


Fig. 2. The hierarchical structure of our model.

Specifically, we build a coarse classifier to discern separate events with the help of the scene information, and then we use multiple features to obtain the fine label of the collections. The architecture of our method is illustrated in Figure 2. We divide the training collections into two parts: $train\_train$ part and $train\_val$ part. We train our classifiers on the $train\_train$ part and evaluate the performance on the $train\_val$ part to determine the parameters of the classifiers.

A. CNN of Objects

Object recognition is one of the most important research topics in the computer vision field. Finding some typical objects or features is meaningful for event recognition. In personal photo collections, some typical objects are highly relevant to certain events, such as Christmas trees for Christmas, Easter eggs for Easter, and mountains for hiking. If we could find the right objects, it would be helpful for recognizing events in the photo collections. Hence, we try to find a descriptor that can represent the object information in photos.

Deep models have recently been applied to large-scale visual recognition tasks and have achieved promising results in single photo recognition. The CNN trained on ImageNet have proved to be a fine descriptor for object recognition. However, the 1,000 categories in ImageNet are not enough to describe the various contents in personal photos. So we choose to extract a 4096-dimensional feature vector using the ImageNet-trained CaffeNet [16] from the fc6 layer for each photo.

B. CNN of Scenes

To recognize events in a photo collection, we can start from another point of view. In addition to recognizing the typical objects, we can also try to recognize the events from the scenes. By ignoring the people and specific objects, we can also guess what happened in photos from the background information. For example, beaches may appear in travel events and indoor or outdoor religious services may appear in weddings.

Traditional low-level features are not sufficient to describe the scenes in various personal photos. Luckily, a convolutional neural network (CNN) trained on the Places database [2] has shown positive performance on scene recognition [4]. We take advantage of this CNN and extract the visual features from each photo.

C. Time Feature

The time at which photos are taken can also help us to recognize events since their occurrence is frequently associated with a certain time of day. For example, hiking is usually arranged on weekends and concerts are often held at night.

We extract the time information from the EXIF data of photos. For the photo-level time feature, we transfer the timestamps and get the time features: year, month, date, and day of the week. For the collection-level time feature, we compute the duration for each photo collection and all the photos in the collection share this data, allowing us to obtain a 5-dimensional feature vector for each photo.

D. Hierarchical Structure

In this section, we introduce our hierarchical model. Our hierarchical model consists of three parts:1) a coarse classifier that separates the easily distinguishable events. 2) many fine classifiers that separate the ambiguous events within coarse events. 3) a probabilistic averaging component that combines the probabilistic predictions of the coarse classifier and the fine classifiers to obtain the final predictions.

1) Pretraining the Coarse Event Classifier: Here, we introduce the first part of our hierarchical model, which tries to separate the easily discernible events. Humans usually use visual information from objects and environments to accomplish the recognition process for a single photo. We mimic this natural process to try to find a visual descriptor for our coarse event classifier. In our model, we take advantage of the CNN features to describe the scenes. After averaging the feature vectors within each collection, we train a standard SVM classifier using the $train\_train$ part and evaluate it on

the $train\_val$ part to determine the parameters of the coarse classifier.

2) Identifying Coarse Events: After analyzing the classification results of the coarse classifier on the $train\_val$ part, we can obtain the confusion matrix $F \in R^{N\times N}$ in which N is the number of events. We first simply make F symmetric by computing $F = \frac{1}{2}(F + F^T)$. The element $F_{i,j}$ estimates how confused $event_i$ and $event_j$ are. Then, Affinity Propagation [17] is employed to the symmetric matrix F to cluster the N events into N_c coarse clusters. Furthermore, we obtain the mapping $P : y \longmapsto y'$ which presents the mapping relationship of a fine event label to a coarse cluster label. The probability that collection A_i is predicted into a coarse cluster C_j is calculated by:

$$B_{i,j} = \sum_{E_k \in C_j} P^c_{E_k}(A_i). \tag{1}$$

where $P^c_{E_k}(A_i)$ is the probability that collection A_i is predicted to be event E_k by the coarse classifier.

Affinity Propagation is applied because it does not require a certain number of clusters before running the algorithm. In addition, the clusters the clusters are more balanced in size than other clustering algorithms. The damping factor λ in the Affinity Propagation is set to be 0.5 throughout the experiments.

3) Training the Fine Event Classifiers: The fine event classifiers are trained independently within each coarse cluster. As it is not enough to build the fine classifiers using only the information from the scenes, we introduce two other kinds of features to help us recognize fine events: a CNN feature of objects and a time feature. Similar to pretraining the coarse event classifier, we train fine classifiers within each coarse cluster with different features, respectively. Then, we produce a weighted average to combine the predictions of the coarse and fine classifiers.

$$P(A_i) = \sum_{j}^{C} B_{i,j} P_j(A_i). \tag{2}$$

where $B_{i,j}$ is the probability that collection 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 Eqn 2. Next, we adopt late fusion to combine the different predictions by

$$\begin{aligned} P_{final}(A_i) = \alpha \times P_{scene}(A_i) + \beta \times P_{object}(A_i) \\ +(1-\alpha-\beta) \times P_{time}(A_i). \end{aligned} \tag{3}$$

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.

4) Fine-tune the parameters of the Coarse Event Classifier: By combining the classifiers together, we preliminarily build our hierarchical model for personal photo collection event recognition. While keeping the mapping relationship P unchanged, we fine-tune the parameters of the coarse event classifier on the $train\_val$ part.

TABLE I
STATISTICS OF THE DATASET [12]

Event	Collections	#Photo
Birthday	60	3227
Children Birthday	64	3714
Christmas	75	4118
Concert	43	2565
Boat Cruise	45	4983
Easter	84	3962
Exhibition	70	3032
Graduation	51	2532
Halloween	40	2403
Hiking	49	2812
Road Trip	55	10469
St. Patricks Day	55	5082
Skiing	44	2512
Wedding	69	9953
Total	807	61364

IV. EXPERIMENT

In this section, we first introduce a personal photo collection dataset for event recognition, and then present the experimental settings as well as comparison methods, followed by the performance of different approaches and analyses.

A. Data Set

We use the personal photo collection dataset released in [12] for event recognition. All the photos are real life photos from Flickr. The contents of the events in the dataset are chosen from the most popular tags on Flickr, Picasa and Wikipedia from categories that correspond to social events. The dataset contains 14 event classes and 807 collections. Each collection has 76 photos on average. The statistics of the dataset are shown in Table I. Collections for training and testing have already been defined in [12]. We randomly divide the training set into 5 partitions and utilize the cross-validation method to determine model parameters in the following experiments. We use average accuracy, recall, and the F1-score to evaluate the performance of different recognition methods.

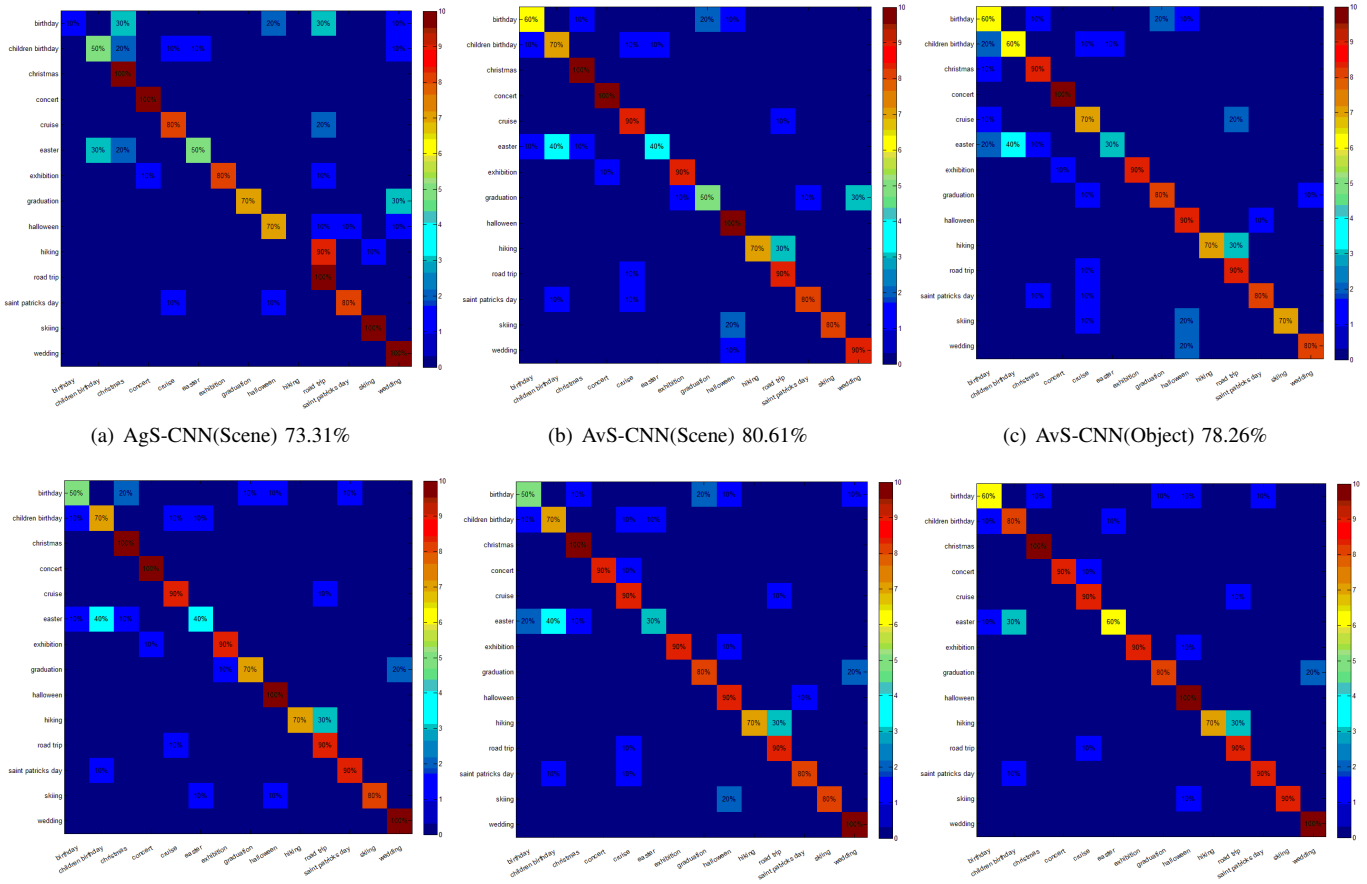
B. Experimental Settings

We take advantage of LibSVM [18] to fulfil our experiments. For CNN features of scenes and CNN features of objects, we perform $L2$ normalization on each feature vector. For time features, Min-Max normalization is adopted for each photo-level time feature, while for the collection-level time feature, we scale it to the size of day.

When training the classifiers, the linear kernel is assigned to the high dimension features: the CNN feature of scenes and the CNN feature of objects. The RBF kernel is assigned to the low dimension feature: the time feature.

TABLE II
PERFORMANCE OF DIFFERENT METHODS

Method	Avg. Acc. (%)	Recall (%)	F1-Score
AgS [12]	41.43	-	0.3887
ShMM [12]	55.71	-	0.5616
AgS-CNN(Scene)	73.31	70.71	0.6705
AvS-CNN(Scene)	80.61	79.29	0.7852
AvS-CNN(Object)	78.26	75.71	0.7543
MFAS	82.11	81.43	0.8068
HASFS-CNN(Scene)	80.91	79.29	0.7854
HAS	**86.32**	**85.00**	**0.8485**



(a) AgS-CNN(Scene) 73.31% (b) AvS-CNN(Scene) 80.61% (c) AvS-CNN(Object) 78.26%

(d) MFAS 82.11% (e) HASFS-CNN(Scene) 80.91% (f) HAS 86.32%

Fig. 3. Confusion matrices for different approaches. We also show the average accuracy for each confusion matrix.

C. Approaches for Event Recognition

In this section, we present the methods for comparison.

1) Aggregated SVM (AgS): We follow the baseline mentioned in [12] as one of our baseline methods. We train a linear multi-class SVM in the photo-level recognition. Each photo inherits the label of the collection it belongs to. We sum up the confidence scores of the photos in the collections and choose the events with the highest scores as the final predictions.

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

3) Multi-Feature Average SVM (MFAS): We use the three kinds of features mentioned above and train SVMs by approach 2), respectively. Next, we adopt late fusion to combine confidence scores on different views.

4) Hierarchical Average Single Feature SVM (HASFS): We simplify the coarse-to-fine hierarchical model proposed in this paper by only using one kind of feature throughout the coarse and fine classifiers.

5) Hierarchical Average SVM (HAS): We use the full hierarchical model mentioned above. Scene features are used for coarse classifiers and multiple features are used for fine classifiers.

D. Experimental Results

We present the performance of different methods for personal photo collection event recognition in Table II and display

the fusion matrices in Fig 3.

When comparing the baseline of aggregated SVM with different features, the CNN feature of scenes achieves an average accuracy of 73.31%, which is 31.88% higher than the low-level visual features in [12]. It proves that high-level features extracted by CNN are much more powerful in event recognition than low-level features.

Average SVM achieves a better average accuracy(80.61%) than aggregated SVM(73.31%) when CNN scene features are applied in both methods. This is because personal photo collections are always composed of several sub-events. Usually the contents in a single photo can only describe part of the event, and they are not enough to define it. Also, photos in different collections may share some common characteristics. For example, a baby's birthday party and a Christmas party may share similar contents, such as people dining. Thus when a collection has many ambiguous photos, equivocal predictions will lead to unexpected misclassification. Luckily, when people take photos of events, they often try to record the complete contents. That's why the collection-level averaged visual features can more accurately describe the contents in the collections and perform better than aggregated predictions of single photos.

Another kind of visual feature we use is the CNN feature of objects. We also compare the two different CNN features using the average SVM method. The CNN feature of objects obtains an average accuracy of 78.26%, which is 2.35% worse than the CNN feature of scenes. Though the performance of these features are similar, we can see that the confusion matrices are different; this inspires us to combine the results of the CNN features. By combining the two kinds of CNN features and the time feature, we obtain much better results than with single features.

Finally, we present the performance of our hierarchical model. When only the CNN feature of scenes is applied, our hierarchical model and the AvS method perform similarly. This is because the limited descriptive ability of the scene feature cannot handle the fine classifiers for event recognition. After adding the CNN feature of objects and the time feature, our full hierarchical model obtains the best average of accuracy, 86.32%, the best recall of 85.00%, and the best F1-score of 0.8485 among all methods.

V. CONCLUSION

In this paper, we propose an event recognition method with a hierarchical structure for personal photo collections. Based on the assumption that not all photos are equally difficult to recognize, we first sort easily discernible events into coarse clusters, and then finely classify them to obtain our final predictions. Multiple features, including time, objects and scenes are introduced to help us better recognize the events in photo collections. We find that the scenes from easily identifiable events are quite different and well-suited for the coarse classifier. Another useful finding is that the prediction of averaged visual features performs better than aggregating the predictions of single ones. We have evaluated our coarse-to-fine hierarchical model on a real-world personal photo collection dataset and our method has proved to be a promising solution for event recognition.

VI. ACKNOWLEDGMENT

This work is supported by the 973 project under the contract No.2015CB351803, the NSFC under the contract No.61390514 and No.61201413, the Fundamental Research Funds for the Central Universities No. WK2100060011 and No.WK2100100021, the Specialized Research Fund for the Doctoral Program of Higher Education No. WJ2100060003.

REFERENCES

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

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

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

[5] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Giro-i Nieto, "Cultural event recognition with visual convnets and temporal models," *arXiv preprint arXiv:1504.06567*, 2015.

[6] N. Imran, J. Liu, J. Luo, and M. Shah, "Event recognition from photo collections via pagerank," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 621–624.

[7] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[8] Y.-G. Jiang, "Super: towards real-time event recognition in internet videos," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 2012, p. 7.

[9] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2650–2657.

[11] 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, 2011.

[12] L. Bossard, M. Guillaumin, and L. Van, "Event recognition in photo collections with a stopwatch hmm," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1193–1200.

[13] S.-F. Tsai, T. S. Huang, and F. Tang, "Album-based object-centric event recognition," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[14] F. Tang, D. R. Tretter, and C. Willis, "Event classification for personal photo collections," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 877–880.

[15] L. Cao, J. Luo, H. Kautz, and T. S. Huang, "Annotating collections of photos using hierarchical event and scene models," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[17] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[18] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.