Recognizing and Curating Photo Albums via Event-Specific Image Importance

Yufei Wang[1]
yuw176@ucsd.edu
Zhe Lin[2]
zlin@adobe.com
Xiaohui Shen[2]
xshen@adobe.com
Radomír Měch[2]
rmech@adobe.com
Gavin Miller[2]
gmiller@adobe.com
Garrison W. Cottrell[1]
gary@ucsd.edu

[1] University of California, San Diego
9500 Gilman Drive,
San Diego, USA

[2] Adobe Research
345 Park Avenue,
San Jose, USA

Abstract

Automatic organization of personal photos is a problem with many real world applications, and can be divided into two main tasks: recognizing the event type of the photo collection, and selecting interesting images from the collection. In this paper, we attempt to simultaneously solve both tasks: album-wise event recognition and image-wise importance prediction. We collected an album dataset with both event type labels and image importance labels, refined from an existing CUFED dataset. We propose a hybrid system consisting of three parts: A siamese network-based event-specific image importance prediction, a Convolutional Neural Network (CNN) that recognizes the event type, and a Long Short-Term Memory (LSTM)-based sequence level event recognizer. We propose an iterative updating procedure for event type and image importance score prediction. We experimentally verified that image importance score prediction and event type recognition can each help the performance of the other.

1 Introduction

With the advent of cheap cameras in nearly all of our devices, automated uploading to the cloud, and practically unlimited storage, it has become painless to take photos frequently in daily life, resulting in an explosion of personal photo collections. However, the oversized image collections make it difficult to organize the photos, and thus automatic organization algorithms are highly desirable. The organization of personal photo collections can be decomposed into two stages: recognizing the event type of a photo collection, and suggesting the most interesting/important images in the photo collection to represent the album. The two stages can assist users in keeping the photo collections organized and free of irrelevant images, and can be further used to pick photos for an album cover or to make a photo collage.

Both event recognition and image importance prediction have been studied independently in previous literature. Studies of event recognition fall into three types. The most popular approach uses videos as input [7, 17, 18, 22, 26, 29, 30], and spatiotemporal features are commonly used. Further, event recounting which aims to find the event-specific spatial/temporal discriminative parts of a video is also studied [7]. This is relevant to event-specific image importance, but the image importance of an image is not decided by how discriminative it is. At the other end of the spectrum, event recognition for single images has also been studied [14, 19, 21, 25]. There is no temporal information or relevant frame importance to consider, and both object and scene level features have been used [14].

Album-wise event recognition lies between single-image-based and video-based event recognition, and is most related to our work. Images in an album can be thought of as very sparse samples from an event video, and consecutive images from the photo album are no longer continuous. A common approach is to aggregate evidence from single images to classify the album type [1, 16, 23, 28]. For example, Wu *et al.* [28] fine-tune Alexnet to extract features from single image, and then aggregate the features from each image and train a multi-layer network to recognize the event type of the album. The above work treats albums as unordered collection of images. On the other hand, Bossard *et al.* [4] exploit the sequential nature of personal albums, using an HMM-based sub-event approach (Stopwatch HMM) for event recognition. They use temporal sequence of the images, and model an album with successive latent sub-events to boost the recognition performance, and show that the temporally-sensitive HMM outperforms simply aggregating the predictions from all the images in an album. This indicates that the sequential information in an album is useful for album-wise event recognition.

Image importance is a complex image property which is related to various factors, such as aesthetics [15], interestingness [5] and image memorability [11]. Wang *et al.* [27] show that image importance is modulated by the context it is in, i.e., image importance is event-specific. For example, a photo of a beautiful work of architecture is important in an album of an urban trip, yet not so important in a wedding event. They showed that a siamese-network-based model can reasonably predict this highly subjective image property. However, their work assumed that the event type of the album is already known. This is undesirable if we want to build an end-to-end photo organization algorithm. In this work, we train a system simultaneously for event recognition and image curation, so that user input of the event type is not required.

Event recognition and image importance prediction are inherently related to each other: 1) importance is event-specific, so we need to know the event type to better predict importance; 2) albums often contain "outlier" photos that aren't directly related to the event. If we can reduce effects from the outliers by discovering the important/key images in an album, we can better recognize the event. Therefore, we ask the question: can we simultaneously recognize the event type of an album, and discover important images in it? And more importantly, can we improve the performance of each task by forming a joint solution?

In answering this question, this paper makes the following contributions: 1) We develop a joint event recognition and image importance prediction algorithm.We use a CNN for image level event recognition, and a Siamese Network for event-specific image importance prediction. Then an iterative update scheme is used during the test stage, and we find that event recognition and image importance prediction can indeed improve each other's performance; 2) We further boost the performance of event recognition with an LSTM network that leverages sequential information in labeling the album; 3) We also refine the CUFED dataset by collecting more human annotations for the event types, allowing raters to apply multiple labels to the events. This improves the reliability of the ground-truth, accounting for the ambiguity between event types.

2 The ML-CUFED Dataset

In order to train and evaluate the joint curation-recognition model, we use the Curation of Flickr Events Dataset (CUFED) [27], and refine it by collecting additional human opinions on the event types in the dataset. We call the new dataset MultiLabel-CUFED (ML-CUFED). In this section, we describe the dataset, and provide a consistency analysis of the labels collected from Amazon Mechanical Turk (AMT). The dataset is publicly available[1]. More details of ML-CUFED are in the supplementary material.

2.1 The CUFED dataset

The CUFED dataset [27] is an image curation dataset extracted from the Yahoo Flickr Creative Commons 100M dataset. It contains 1883 albums over 23 common event types, with 50 to 200 albums for each event type. The event type of each album was decided by 3 AMT workers' annotations. Meanwhile, within each album, the event-specific importance of each image is obtained by averaging 5 AMT workers' votes when the event type is given to them.

One problem with CUFED is that the event type of an album is decided by only 3 workers, who were constrained to give a single label to each album. However, some of the event types in that dataset are related (e.g., architecture and urban trip). For an album with ambiguous or multiple event types, such a constraint is overly restrictive. Therefore, collecting the event types and their proportion in one album from more peoples' views is necessary. This results in a multi-label event recognition dataset with richer information. In the supplementary material, we show examples of event labels from CUFED and our refined labels.

2.2 Data collection and Analysis

In addition to the 3 votes the dataset already includes, we collected 9 more workers' opinions for each album, and allowed them to select up to 3 event types. There were 299 distinct workers who participated in the task.

Quality control was performed for each AMT worker in order to collect high quality annotations. Before the real task, only workers who passed a test that was very similar to the actual task were allowed to proceed. During the tasks, the results workers turned in were compared with other workers' submissions, and submissions that highly diverged from others were further manually inspected. If the divergence was unreasonable, the submission was rejected. After all the annotations from workers were collected, we further cleaned the annotations by eliminating the labels with only one vote. To get the final ground-truth event types, we converted the votes to a probability distribution over event types for an album.

To check the validity of the dataset we collected, we analyzed the annotations in several ways. Each album has between 9 and 27 votes (because we allow for multiple choices from one worker). 76% of the albums received votes for two or fewer event types. 95% of the albums received votes for three or fewer event types. To check the consistency among workers, we randomly split the 299 workers into two halves, and for each album we checked whether the annotations from one half were consistent with the other half. We repeated the random split 100 times, and on average, for 89.6% of the albums, the event type receiving

[1] http://acsweb.ucsd.edu/~yuw176/event-curation.html

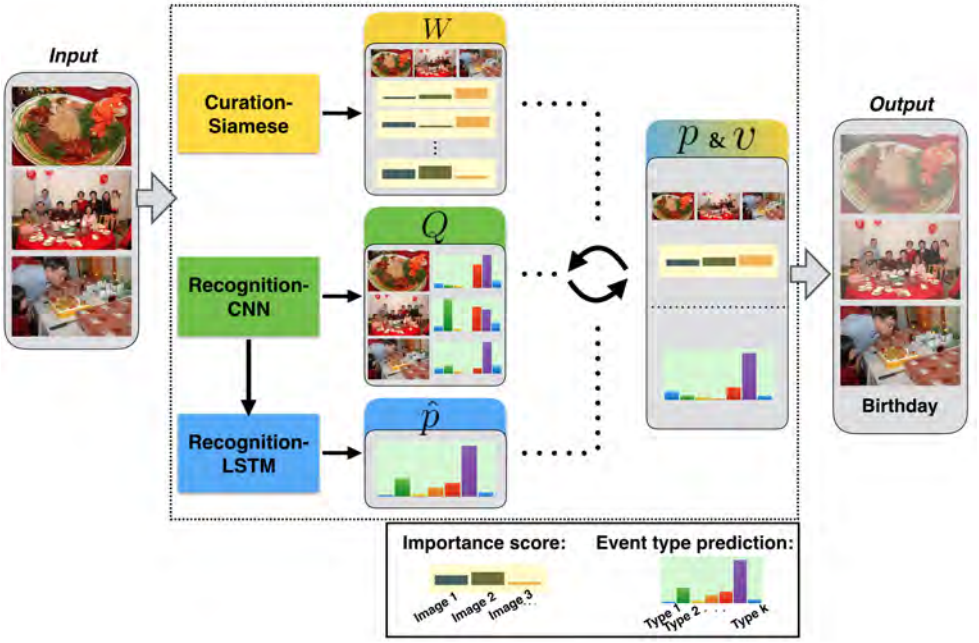

Figure 1: The joint album recognition-curation system. $\{W, Q, \hat{p}, p, v\}$ are described in Section 3.3. W, Q, and $\hat{p}$ are computed once and then used to iteratively update p and v.

the most votes were the same for both groups. This suggests that despite the ambiguity of some album types, the opinions of different AMT workers are consistent.

3 Joint Event Recognition and Image Curation

In this section, we describe our approach to jointly attain image importance prediction and album event recognition. It is intuitive that important images contribute more to the identity of an event, and should be emphasized when deciding the event type of the album from the images. On the other hand, the identity of the event is needed for accurate individual image importance prediction, as shown in [27]. Moreover, it has been shown that sequential information in an album is useful for event prediction [4]. Therefore, we build a joint system that can simultaneously predict the event type and image importance for an album. The system is shown in Figure 1. We elaborate on the different parts of the system in this section.

3.1 Event curation network

For event curation, we use a similar approach as in [27], using Piecewise Ranking Loss to train a Siamese network to predict the importance score difference between an image pair given the ground-truth event type. The Siamese network outperforms a traditional CNN that directly predicts the absolute image importance score. Compared to the architecture in [27], we added a pathway to directly predict the score difference between the image pair, rather than looking at the two images separately. This makes the training process faster and improves the results. The architecture is shown in Figure 2, and the added pathway is ***Pathway2*** in the middle. More details are presented in the supplementary material.

For each training image pair, the "ground-truth" event is sampled from the label distribution, and used to gate the output and gradient of the network. We denote this network as **Curation-Siamese**.

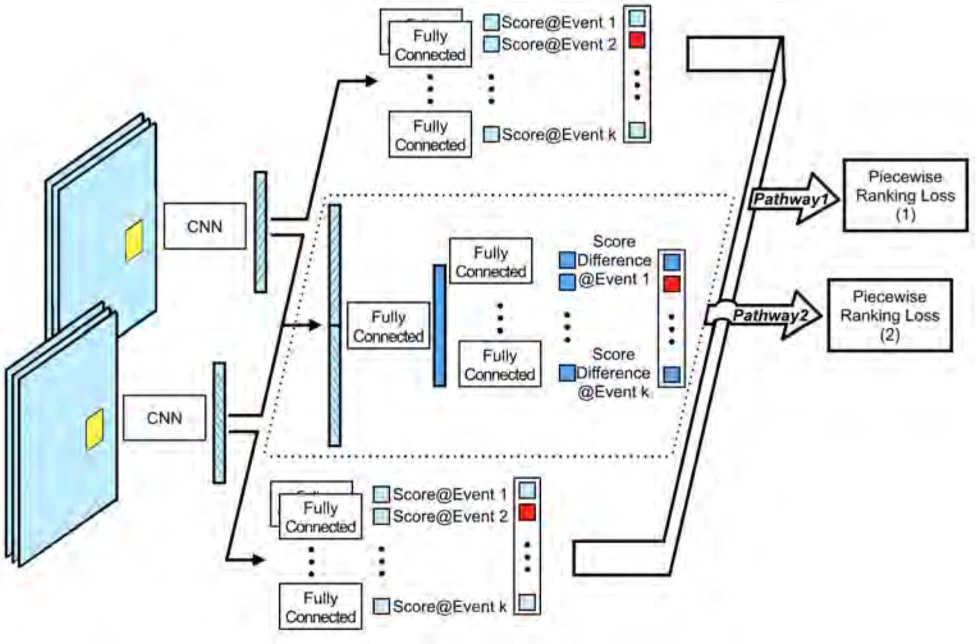

Figure 2: Architecture of the event curation siamese network (Curation-Siamese) during training.

3.2 Event recognition networks

One of the properties of an "event album" that makes it distinct from a simple collection of images is that it is a sequence, and this provides us with the temporal relationship between the images. LSTMs have been successfully applied to sequential tasks [6, 8, 10, 20, 24], and their ability to remember long-range temporal context is suitable for our task. Therefore, we use the LSTM network to capture the sequential information, in addition to a classical CNN that captures the visual features of a single image.

We start with a CNN pre-trained on ImageNet [12, 13], and optionally fine-tune it on ML-CUFED to recognize the event type from a single image. We call this network **Recog-CNN**. We extract the high level CNN features for each image, and use them as the input features to train the LSTM network for album-wise event recognition. The LSTM network consists of a single LSTM layer as in [6], a mean pooling layer, and a softmax layer. Cross entropy loss is used. We denote the LSTM as **Recog-LSTM**. More details about the structure of the LSTM network are presented in the supplementary material. The target for both the CNN and LSTM network is a one-hot encoding, but is sampled from the ground-truth label distribution.

3.3 The iterative curation-recognition procedure

For an "event album", more important images give us more information about the event type. For example, although a candle blowing image may only appear in an album once, it is a critical clue for revealing the event type of the album. However, as shown in [27], the importance of an image is event-type dependent. Therefore, we propose an iterative update procedure to demonstrate that the image importance score and event recognition of an album can be used to improve each other's performance.

We denote an N-image album as $\mathbf{A} = \{I_1, ..., I_N\}$. We assume C different event types. The input to the algorithm is the output of the above three networks: 1) Recog-CNN produces an N-by-C matrix Q, where each row is a probability distribution over event-types, given the image; 2) Recog-LSTM produces a 1-by-C row vector of probabilities of event types, $\hat{p}$, given the image sequence; 3) Curation-Siamese produces a N-by-C matrix W, where each row is the importance score of an image, given the event-type. The output of the algorithm is the N-dimensional column vector $v = [v_1, ..., v_N]^T$, which is the prediction of the importance score for all images in album $\mathbf{A}$, and the C-dimensional row vector $p = [p_1, p_2, ..., p_C]$, the distribution over the possible event types.

The iterative curation-recognition procedure is as follows:

1. **Re-weight Recog-CNN event prediction by image importance.**

$$p'(k+1) \propto (v^T(k))^\alpha \cdot Q \tag{1}$$

   where $v(k)$ is the k-th step prediction for all images' importance scores in album $\mathbf{A}$ (initialized to a uniform distribution) and α is a parameter that controls the strength of the importance score for the update. p' denotes the updated album event type prediction, normalized to a distribution. Thus, p' is a distribution over event types that is the average of each image's event distribution weighted by the image's predicted importance score.

2. **Combine event type predictions** with $\hat{p}$.

$$p(k+1) = \frac{1}{2}(p'(k+1) + \hat{p}) \tag{2}$$

   where $\hat{p}$ is the probability distribution of event types predicted by Recog-LSTM. Thus, $\hat{p}$ serves as an"anchor" for the prediction.

3. **Update image importance score with the updated event type distribution.**

$$v(k+1) \propto \left\{ W \circ \mathrm{I}\left\{ p_c \geq m \cdot \max_{c'}(p_{c'}) \right\}_{(1,c)} \right\} \cdot p(k+1)^T \tag{3}$$

   where W is the importance scores of all the images given the event type from Curation-Siamese, $\circ$ denotes element-wise multiplication of each row, and $\mathbf{I}$ is an indicator that returns 1 if its argument is true and 0 otherwise. Hence, $\mathbf{I}$ forms a binary mask that zeros out the importance scores for columns of W that correspond to low-probability events, computed as a fraction m (a parameter) of the maximum probability event. Thus, the updated image importance is the average of the importance score given different events, weighted by the event type probability. Elements of $v(k+1)$ are normalized to range from 0 to 1.

By iterating Equations 1-3, we obtain the album-wise event prediction p and image importance score prediction v. Note that this procedure is not guaranteed to converge, hence we set a maximum number of iterations, and if this maximum number is reached before convergence, the predictions for p and v are obtained by averaging over last three steps.

4 Experiments

In this section, we evaluate our approach for both event recognition and image importance prediction on ML-CUFED, and on another album-wise event recognition dataset Bossard *et al*. [4] collected called PEC, we compare our event recognition result with Bossard *et al*. [4] and Wu *et al*. [28].

4.1 Baselines

Our joint recognition-curation method produces two outputs: an album event type prediction, and an image importance prediction. For event recognition, we compare our result with the baseline from Recog-CNN. In addition, the intermediate result of our algorithm can also be compared with the final result to validate the necessity of each part of our system. Therefore, we compare our method with the following methods:

- **CNN-recognition**: Use Recog-CNN to predict the event type for each image, and average the results.
- **CNN-LSTM**: The prediction by Recog-LSTM. Note this uses Recog-CNN's feature representation as input.
- **CNN-Iterative**: Use the proposed method as described in Section 3.3, but without step 2. Therefore, Recog-LSTM result is not involved.
- **CNN-LSTM-Iterative**: Our full proposed method as described in Section 3.3.

To evaluate our image importance prediction, we compare with several baselines:

- **CNN-Noevent**: Train a Siamese Network to predict the importance score difference of an input image pair without any event-type information. All albums are considered to be part of the same "uber" event type.
- **CNN-Noevent(test)**: Use Curation-Siamese that is trained using the ground-truth event type information to gate the output error and back-propagation signal, while during testing, average the predicted importance score for all possible event types.
- **CNN-LSTM-Iterative**: As above.

4.2 Experimental details

Dataset For ML-CUFED, we split the albums into training and test in a ratio of 4:1. The test set has 368 albums. To decide the hyper-parameters (α, m) in our iterative model, a validation set with 111 albums is extracted from the training set. For the PEC Event Recognition Dataset [4], we use directly the test set consisting of 10 albums for each event type as described in [4], so that we can directly compare their results with ours.

Parameter setting For both the Recog-CNN and the Curation-Siamese, we use two architectures: 8-layer AlexNet [13] and 101-layer ResNet [9]. Both networks are pre-trained on ImageNet, and we fine-tune AlexNet on ML-CUFED. We use a similar training scheme to [12], but with a lower learning rate of 0.001. For Recog-LSTM, we use high-level features from the Recog-CNN as input. For fine-tuned AlexNet, we use *fc7* layer features, while

	MAP@t%						**P@t%**					
t%	5	10	15	20	25	30	5	10	15	20	25	30
Random	0.113	0.161	0.211	0.256	0.303	0.350	0.044	0.090	0.142	0.193	0.243	0.298
CNN-Noevent(test)	0.272	0.330	0.380	0.434	0.483	0.530	0.167	0.256	0.327	0.379	0.432	0.476
CNN-Noevent	0.280	0.352	0.403	0.455	0.504	0.552	0.178	0.281	0.347	0.404	0.454	0.497
CNN-LSTM-Iterative	**0.302**	**0.371**	**0.419**	**0.470**	**0.520**	**0.568**	**0.205**	**0.300**	**0.360**	**0.413**	**0.459**	**0.507**
CNN-GTEvent	0.309	0.383	0.432	0.482	0.529	0.573	0.205	0.311	0.373	0.428	0.472	0.512

Table 1: Comparison of event-specific image importance predictions with different methods using ResNet features. Evaluation metric here is MAP@$t\%$ and $P@t\%$.

for ResNet, we use the *pool5* layer features. We reduce the feature dimension to 512 with PCA. For the Recog-LSTM, the dimensionality of the LSTM is 512, and we use AdaDelta as the optimization method [2, 3, 31]. For Curation-Siamese, we follow the settings in [27] and choose the two margins as $m_s = 0.1$ and $m_d = 0.3$. We set the number of iterations of our joint recognition curation algorithm to 10. Please find more analysis regarding the performance with respect to the iteration number in the supplementary material.

Evaluation For event recognition on ML-CUFED, we use two metrics to evaluate the models: average accuracy and F_1 Score. F_1 Score is the harmonic mean of precision and recall, and can account for multi-label ground-truth. Both accuracy and F_1 are calculated with top-1 prediction. For event recognition on PEC, only average accuracy is used. For image importance prediction, we follow [27] using MAP@($t\%$) and Precision@($t\%$). Precision is the ratio between the number of retrieved relevant images and the number of retrieved images. MAP is the averaged area under the precision-recall curve.

4.3 Results on the ML-CUFED Dataset

4.3.1 Event-specific image importance

For the image importance score prediction task, we compare our methods to several baselines in Table 4.3.1 using ResNet features. Results for AlexNet features are in the supplementary material. For an upper-bound of our method, we also show the result for CNN-GTEvent, where we assume the ground truth event type is known, and predict the importance score based on that. CNN-GTEvent serves as the best result we can get when the event recognition stage is perfect. As shown in Table 4.3.1, CNN-Noevent performs better than CNN-Noevent (test). This suggests the divergence of the importance prediction for different event types. CNN-LSTM-Iterative greatly outperforms the other two models, filling the gap between CNN-Noevent (test) and CNN-GTEvent by 79%, and between CNN-Noevent and CNN-GTEvent by 62% (averaged over 6 levels of t and between MAP and P).

4.3.2 Event recognition

Table 2 shows the results of different methods for event recognition. For an album with multiple labels, we deem it correctly predicted if the top-1 prediction is among the ground-truth event labels. As shown, ResNet features perform much better than AlexNet features. For both AlexNet and ResNet features, there is a performance gain over all three baselines. We can also observe that both iterative curation-recognition and LSTM method help to improve the final result. This suggests that both these types of information in an event album are helpful in deciding the event type of this album: image importance information, and album sequential information.

Dataset	**ML-CUFED**				**PEC**	
	Avg. Acc.		**F1-Score**		**Avg. Acc.**	
Method	AlexNet	ResNet	AlexNet	ResNet	AlexNet	ResNet
CNN-recognition	75%	82.9%	0.698	0.772	80.9%	84.5%
CNN-LSTM	76.6%	81.5%	0.713	0.759	82.7%	85.5%
CNN-Iterative	78%	83.7%	0.729	0.781	81.8%	86.4%
CNN-LSTM-Iterative	**79.3%**	**84.5%**	**0.737**	**0.786**	**84.5%**	**87.9%**
Wu *et al*. [28]	71.7%	83.4%	0.662	0.773	*84.5%	*89.1%
SHMM [4]	-		-		*76.3%	

Table 2: Comparison of event-recognition models on ML-CUFED and PEC. Note that for the PEC result, our model is trained on ML-CUFED, while Wu *et al*. [28]'s model and SHMM are trained on the PEC training set.

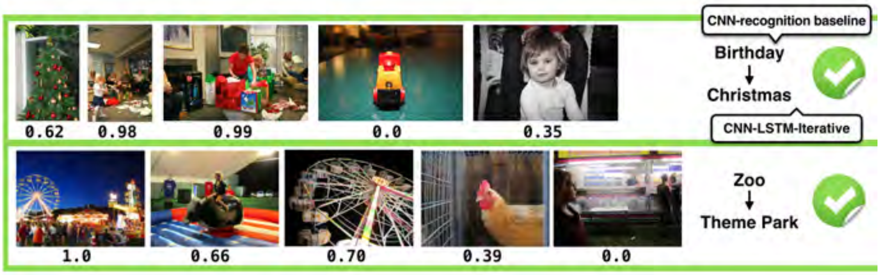

Figure 3: Examples of our model's results on the ML-CUFED Dataset where the event type result is corrected by the iterative model, as shown on the right. The images are in order of ground-truth importance, with the model's importance score below each image.

We compare our results with another CNN-based model in [28]. Wu *et al*. [28] use a fine-tuned AlexNet to extract image features, and aggregate the image features for album-wise prediction of event type. Here, we reimplement their approach for ML-CUFED, using both fine-tuned AlexNet features and ResNet features. Our model substantially outperforms theirs.

In Figure 3, we show our event curation and recognition result using ResNet with two examples in the ML-CUFED Dataset. The images are sorted in ground-truth importance order, and the predicted importance score is labeled below the image. To the right of the album, we show the event recognition of the album with the CNN-recognition method (before arrow), and the CNN-LSTM-Iterative (after arrow). As shown, the event recognition is corrected with the CNN-LSTM-Iterative procedure. We show with more examples in the supplementary material.

4.4 Results on the PEC Dataset

To show the generalizability of our algorithm, we compare our result with [4] and [28] on PEC. The PEC dataset is an 807-album event dataset with 14 social event classes. There is no ground-truth importance score in PEC, thus we cannot train our algorithm on it. Therefore, we use the model we trained on ML-CUFED and test the model on the PEC test set containing 10 albums each class.

PEC has several event types that are not contained in ML-CUFED, such as Saint Patrick's Day, Easter, and Skiing, and there are multiple event types that can map to single event type in ML-CUFED: Children's Birthday and Birthday can be mapped to single Birthday

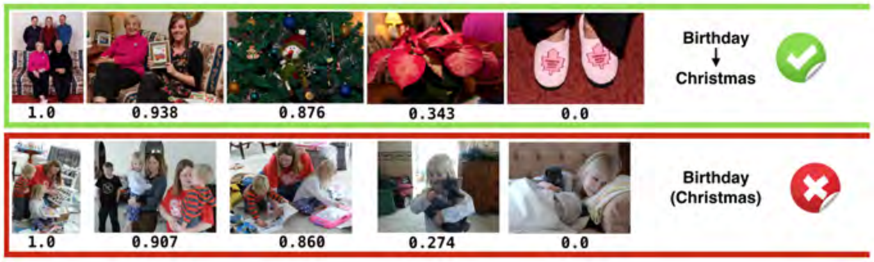

Figure 4: Some results from the PEC dataset. For the first row, the prediction from CNN-recognition is wrong as shown in the label above, but is corrected in the final answer. For the second row, a case where the model fails to recognize a Christmas event is shown. Only a subset of images is shown for each album due to limited space. The images are sorted in predicted importance score order, and the predicted importance score of each image is shown below the corresponding image.

event in ML-CUFED. We provide the mapping from PEC label to ML-CUFED label in the supplementary material. Note that the mapping is not perfect, and the noise in the mapping makes the performance of our method shown here a little poorer than it really is.

Due to the label changes, we recalculate the performance of Stopwatch HMM (SHMM) [4] on the test data based on the confusion matrix they provided in the paper. For merged labels, the corresponding rows in the confusion matrix are merged. For missing labels, there are many possible approaches, and we follow the most loose one which assumes the best possible predictions: Assume false positive on those labels will be correct predictions if those labels disappear.

The comparison of different methods is shown in Table 2. Similar to the result on ML-CUFED, we observe consistent performance gain from both LSTM network and iterative updates. For the result of our reimplementation of [28], it is worth noticing that this model is trained on PEC, and it achieves current state-of-the-art result on PEC. Although our model is trained on ML-CUFED, it achieves very close performance with [28].

We show some examples of our recognition and event-specific image importance prediction result in Figure 4. There is no ground-truth labeling for the event-specific importance score, but we can look at the sample results qualitatively. From the first row, we can see that the model does not simply assign a high importance score to the characteristic Christmas tree which can distinguish the Christmas event , but predicts higher score to the family photo. We show with more examples in the supplementary material.

5 Conclusion

In this work, we explore the problem of automatically recognizing and curating personal event albums. It is the first attempt to solve the following two tasks jointly: recognizing the event type of an album, and finding the important images in this album. Specifically, the result from a CNN for image-wise event recognition, an LSTM Network for album-wise event recognition, and a Siamese Network for image importance prediction are integrated by a unified, iterated updating algorithm. We show that the joint algorithm significantly improves both image importance prediction and event recognition.

References

[1] Siham Bacha, Mohand Said Allili, and Nadjia Benblidia. Event recognition in photo albums using probabilistic graphical model and feature relevance. In *International Conference on Pattern Recognition (ICPR)*, 2016.

[2] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[3] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[4] 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*, 2013.

[5] Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[6] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[7] C. Gan, Naiyan Wang, Y. Yang, Dit-Yan Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2568–2577, June 2015. doi: 10.1109/CVPR.2015.7298872.

[8] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 1997.

[11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.

[14] L. J. Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007.

[15] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[16] R. Mattivi, J.R.R. Uijlings, F. deNatale, and N. Sebe. Exploitation of time constraints for (sub-) event recognition. In *ACM Workshop on Modeling and Representing Events (J-MRE 11)*, 2011.

[17] M. Nagel, T. E. J. Mensink, and C. G. M. Snoek. Event fisher vectors: Robust encoding visual diversity of visual streams. In *British Machine Vision Conference*, 2015. URL https://ivi.fnwi.uva.nl/isis/publications/2015/NagelBMVC2015.

[18] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, and Roeland Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[19] Sungheon Park and Nojun Kwak. Cultural event recognition by subregion classification with convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015.

[20] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*.

[21] Amaia Salvador, Matthias Zeppelzauer, Daniel Manchon-Vizuete, Andrea Calafell-Orós, and X. Giró-i Nieto. Cultural event recognition with visual convnets and temporal models. In *CVPR ChaLearn Looking at People Workshop 2015*, 06/2015 2015. URL http://www.cv-foundation.org/openaccess/content_cvpr_workshops_2015/W09/papers/Salvador_Cultural_Event_Recognition_2015_CVPR_paper.pdf.

[22] Kevin Tang, Li Fei-Fei, and Daphne Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[23] Shen-Fu Tsai, Liangliang Cao, Feng Tang, and Thomas S. Huang. Compositional object pattern: a new model for album event recognition. In *Proceedings of the 19th International Conference on Multimedia 2011*, 2011.

[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, April 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2587640.

[25] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring object-scene convolutional neural networks for event recognition in still images. *CoRR*, abs/1609.00162, 2016. URL http://arxiv.org/abs/1609.00162.

[26] X. Wang and Q. Ji. Hierarchical context modeling for video event recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2616308.

[27] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and W. Cottrell, G. Event-specific image importance. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[28] Zifeng Wu, Yongzhen Huang, and Liang Wang. Learning representative deep features for image set analysis. *IEEE Trans. Multimedia*, 17(11):1960–1968, 2015. doi: 10.1109/TMM.2015.2477681. URL http://dx.doi.org/10.1109/TMM.2015.2477681.

[29] Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. A discriminative cnn video representation for event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2015.

[30] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[31] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. URL http://dblp.uni-trier.de/db/journals/corr/corr1212.html#abs-1212-5701.