# Joint Event Recognition and Image Selection

Anonymous ECCV submission

Paper ID 1644

**Abstract.** Automatic image organization of personal photos is a problem with many real world applications, and can be divided into two main tasks: recognizing event types of the photo collections, and selecting interesting images from the collections. The two tasks are both challenging, in that the photos from the same event type are highly varied, and there is a great deal of overlap in photos from different event types. In this paper, we looked into the possibility of simultaneously solving both tasks: album-wise event recognition and image-wise importance prediction conditioned on the recognized event types. We collected an event album dataset with both event type labels and image importance labels, refined from the existing CUFED dataset. We propose a hybrid system consisting of three parts: Siamese network based event-specific image importance prediction, Convolutional Neural Network(CNN) based event recognition, and Long Short-Term Memory(LSTM) based sequence level event recognition, and we propose an iterative updating procedure for event type and image importance score prediction. We show with experiments that the proposed method outperforms the classical approach based on single static image classification, and more importantly, we verified that image importance score prediction and event type recognition can in turn help the performance of each other.

**Keywords:** Event Recognition, Image Importance, Convolutional Neural Network, Long Short-Term Memory

## 1 Introduction

With the advent of cheap cameras in nearly all of our devices, automated uploading to the cloud, and practically unlimited storage, it becomes painless to take photos frequently in daily life, resulting in the 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. This can be used to pick photos for an album cover, and to make a photo book. The two stages assist users in keeping the photo collections organized and free of irrelevant images.

Both event recognition and image importance prediction have been studied independently in previous literature. Studies of event recognition can be categorized into several types. The most popular branches of study use videos as

input [1–3], so spatiotemporal feastures are used for the task. At the other end of the spectrum, event recognition for single images has also been attempted [6–8]. In contrast to video-based event recognition, there is no temporal information to exploit, and there is no need to consider relevant frame importance or contribution to the event, since there is only one "frame." This problem can be viewed as a special case of scene recognition, and both object level features and scene level features are used [6].

Album-wise event recognition lies in the middle between single-image-based and video-based event recognition, and is the most related branch to our work. Images in an album can be thought of as very sparse samples from the event video. Photo albums differ from videos in that consecutive images from the photo album are no longer continuous, and can have very different visual and semantic information. However, there is still sequential information in time-stamped albums. In [9], the success of an HMM-based model for album event recognition indicates that the sequential information in an album is still very helpful for album-wise event recognition.

Image importance is a complex image property which is related to various factors, such as aesthetics [10], interestingness [11] and image memorability [12]. It has been shown [13] that image importance is modulated by the context of the image, i.e., image importance is album-dependent, or event-specific. Event-specific image importance is a highly subjective judgment, and learning it is a very challenging task, due to the very high intra-class variability, and the underlying uncertainty caused by the subjectiveness of the property. Nevertheless, recent work [13] based on Convolutional Neural Networks (CNNs) shows that it can be reasonably predicted.

Event recognition and image importance prediction are inherently related to one other: 1) importance prediction is event-dependent, so we need to know the event to better predict importance; 2) albums often contain "outlier" photos that aren't directly related to the event, which affects event prediction. So if we know the important/key images in an album, we better recognize the event type. Therefore, returning to the task of photo collection organization, we ask the question: can we simultaneously recognize the event type for an album, and discover important images in the event album?

To answer this question, this paper makes the following contributions: 1) We develop a joint event recognition and image importance prediction algorithm, which has not been addressed before. We use a CNN for image level event recognition, and a Siamese Network for event-specific image importance prediction. An iterative update scheme is conducted during the test stage, and we find that event recognition and image importance prediction can improve each other's performance; 2) We further boost the performance of event recognition with an LSTM network that leverages sequence 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, making explicit the ambiguity between event

types. This provides us with more training information, and allows for a fairer evaluation at the testing stage.

## 2 Related Work

Our work is partly inspired by [13], who proposed a novel image property: event-specific image importance. In that work, it was claimed that image importance or "interestingness" is contextual and is related to the album it is in. 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. A Siamese network is used to predict the relative importance score difference between an input image pair, and is jointly trained on all event types.

However, in [13], it is assumed that the event type of the album is already known. This is unacceptable if we want to build an end-to-end photo organization algorithm. In this work, we relax this assumption and train a system simultaneously for event recognition and image curation, so that additional user input of the event type is not required.

Our work is also closely related to the study of event recognition for a personal album. The model in [14] classifies a personal album into 8 social events and 10 sports events simply by aggregating SVM classification results from single images in the album. In [15], Tsai *et al.* exploit object level patterns for event type recognition. Object patterns are learnt from single images, and then an album-wise SVM is trained on the frequency distribution of different object patterns appearing in an album. Similarly, Imran *et al.* [16] use the Pagerank technique to mine the most useful features for an event, and an album-wise SVM classifier is used for recognition. The above work treats albums as unordered collections of images. On the other hand, in [9], Bossard *et al.* exploit the sequential nature of personal albums and use an HMM based sub-event approach for event recognition. They use temporal sequence of the images, and model an album with successive latent sub-events to boost the recognition performance. It has been shown that the temporally-sensitive HMM outperforms the simple aggregation of predictions from all the images in an album. They also collected a 14 class dataset consisting of 807 albums for the task as a benchmark for album-wise event recognition.

Event recognition for single photos has also been studied. Li *et al.* [6] use a generative graphical model to recognize event types of a database with 8 sports events. Their model integrates cues from scene and object categorization to classify the sports events. Deep Neural Networks, especially Convolutional Neural Networks(CNNs) provide us with outstanding visual representations that can be used for the single static image event recognition task [7, 8]. Salvador *et al.* [8] apply CNNs to cultural event recognition. They integrate cues from visual features extracted by a CNN with the time-stamp of a photo, inspired by the fact that photos of a cultural event are mostly taken in the same period of time. However, in personal photo collections, the relevance of an image within an event

album varies a great deal. These approaches for single images are useful, but not sufficient for album-wise event recognition.

Convolutional Neural Network(CNN) methods have greatly boosted performance in image understanding tasks, such as image classification, object detection and scene recognition [17–20]. Now a many researchers have switched their focus to higher-level image properties, such as event recognition [21], semantic segmentation [22], multilabel image annotation [23], and image captioning [24].

Long Short-Term Memory(LSTM) networks [4] have been proposed for sequence prediction and sequence labeling. LSTM networks have advantages over traditional Recurrent Neural Networks (RNN) in that they can maintain contextual information across an extended sequence of inputs. LSTMs have achieved success for tasks such as handwritten text recognition [25] and speech recognition [26]. The success of LSTMs for sequence prediction tasks have also been extended to video-based event recognition. In [4], the LSTM network and visual feature extraction network is stacked, and the network can deal with both event recognition and description. Reiter *et al.* [5] also combine LSTM and HMMs for video meeting analysis. Relevant to our work, the Long-term Recurrent Convolutional Network (LRCN) model [24] has been proposed to stack CNN feature extractors and LSTM networks for sequential learning of videos or images.

## 3 The ML-CUFED Dataset

In order to train and evaluate the joint curation-recognition model, we use the Curation of Flickr Events Dataset (CUFED) [13], and refine it by collecting more 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 will be made available to the public.

### 3.1 The CUFED Dataset

The CUFED Dataset [13] is an image curation dataset extracted from the Yahoo Flickr Creative Commons 100M Dataset (YFCC100M). It contains 1883 albums over 23 common event types from our daily life, ranging from 50 to 200 albums for each event type. The event type of each album was decided by three workers' annotations through Amazon Mechanical Turk (AMT). Meanwhile, within each album, the event-specific importance score of each image is obtained by averaging 5 AMT workers' votes when the event type is given to them.

One problem the CUFED Dataset has is that the event type of an album is decided by only three workers, who were constrained to give a single label to each album. However, some of the event types identified in that dataset are related (e.g., architecture and urban trip). For an album with ambiguous event types or with multiple event types, such a constraint is overly restrictive. For example, the two albums in Fig 1 are both birthday events, but they can also fall into the category of casual family/friends gathering. These two event types

are not mutually exclusive. Moreover, intuitively, we would consider the album on the right to be a more typical birthday event, with distinguishable elements such as birthday hats and cakes, while the album on the left is more of a casual family/friends gathering rather than an obvious birthday event. 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.

Fig. 1: Example of two birthday albums (both have the photo uploader's tag "birthday").

### 3.2 Data collection

In addition to the three votes the dataset already includes, we collected 9 more workers' opinions, and allowed for multiple choices for each album. We allowed workers to select up to three event types for an album. 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 an album event recognition test (which is very similar to the actual task) were allowed to proceed to work on the actual task. During the tasks, there was another round of quality control. After workers submitted the tasks, the results they 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 very rare votes: for each album, all the event types with only one vote were discarded.

For the final ground-truth event types and their proportion of one album, we use the the proportion of votes among all the votes. For workers who give votes to more than one label, each of their votes is normalized so that all the votes sum to one.

### 3.3 Dataset Analysis

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. This suggests the high coherence of those albums. 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 was consistent with the other half. For each album, we examined whether the top event types suggested by these two independent groups were the same. We repeated the random split 100 times, and on average, for 89.6% of the albums, the event type receiving most votes were the same for both groups. This suggests that despite the ambiguity of some album types, the opinions of different AMT workers were consistent.

## 4 Joint Event Recognition and Image Curation

In this section, we describe our approach to jointly attain image importance prediction and album event-type 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. Also, the identity of event is needed for accurate individual image importance prediction, as shown in [13]. Moreover, it has been shown that sequential information in an album is useful for the prediction. Therefore, we build a joint system that can simultaneously predict the event type and image importance ratings for an album. The system is shown in Figure 2. We elaborate on the different parts of the system in the rest of this section.

Fig. 2: The joint album recognition-curation system. The system consists of three parts: a Siamese network for image importance score prediction, a CNN for single image event-type recognition, and an LSTM network for album-wise event-type recognition. During the test stage, followed by a iterative updating rule, the three components interact and jointly produce the prediction of album event type and image importance score.

### 4.1 Event curation network

For event curation purposes, we followed the approach in [13], using Piecewise Ranking (PR) loss to jointly train a Siamese network to predict the importance score difference between an image pair given the ground-truth event type of the input image pair. The architecture can predict the relative event-specific importance score for a set of images, and is found to perform better than a traditional CNN that directly predicts the absolute important score of an image. One difference in our implementation compared to [13] is that the ground-truth event for the input image pair is not a one-hot representation any more; instead, it is represented as a soft distribution. In [13], the event type label is used to gate the output and gradient of the Siamese network: for an input image pair of a certain event type $c \in$ only the part of the network which corresponds to event type $c$ is back-propagated. Here, to train the Siamese network on the multi-event type label, we change the 0/1 gating to a soft gating, thus for an input image pair with multiple event type labels, the error signals from all the possible labels are back-propagated, but with a weight the ground-truth probability of that event type. Another scheme is to still use gating, but for an input image pair with multiple event type labels, ~~there is~~ equal probability for the network to view the input image pair as from one of the possible labels. This gives more relaxation on the ground-truth event type label.

### 4.2 Event recognition network

One of the properties of an "event album" that makes it different 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, and its ability to do long-range context memorization is suitable for our task of album-wise event recognition. 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 [27] [17], and fine-tune it on the ML-CUFED Dataset to recognize a single image's event type. We then extract the high level CNN features for each image from the adapted network, 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, a mean pooling layer, and a softmax prediction layer.

The target for both the fine-tuned CNN and LSTM network is the soft distribution of the ground-truth event labels. Again, another scheme is to use the one-hot target, but treat each training example as from one of the possible event types with equal probabilities.

~~During testing stage, the prediction of an album type is achieved by simply averaging the two predictions from CNN and LSTM.~~

### 4.3 The iterative curation-recognition procedure

For an "event album", more interesting images or important images give us more information about the event type of this album. For example, although a candle blowing image may only appear in an album once, it is ~~very helpful for deciding~~ the event type of the album. However, as shown in [13], the importance of an image is related to the context it is in, and is event-type dependent. Therefore, we propose that the image importance score can help with event recognition of an album, while event recognition of an album will in turn improve the image importance score prediction.

We denote an $N$-image album as $\mathbf{S} = \{I^1, ..., I^N\}$. The outputs of the algorithm are the $N$-dimensional column vector $V = [v^1, ..., v^N]^T$, which is the prediction for all images' importance score in album $\mathbf{S}$, and the $C$-dimensional column vector $P = [p_1, p_2, ..., p_C]^T$ is the prediction for the album's event type. Here, $C$ is the number of event types.

We also need to define the intermediate result ~~we need to perform~~ the iterative update. From the event curation Siamese Network in Section 4.1, we obtain the importance score of an image $I^n \in \mathbf{S}$ given its event type: $W^n = [w_1^n, ..., w_C^n]^T$ ~~From~~ the event recognition CNN, we can ~~get~~ the ~~confidence~~ distribution of the event type for a single image $Q^n = [q_1^n, ..., q_C^n]^T$. Moreover, ~~from~~ the LSTM network, we can ~~get~~ the ~~confidence~~ distribution of the event type for the album $\hat{P} = [\hat{p}_1, ..., \hat{p}_C]^T$.

The iterative curation-recognition procedure ~~can be described as follows:~~

1. ~~**Update**~~ **album event type prediction** from CNN with predicted image importance score $V(k)$.

$$\check{p}_c(k+1) = V^T(k)^\alpha \cdot Q^n \tag{1}$$

   where the $N$-dimensional column vector $V(k) = [v^1, ..., v^N]^T$ is the $k$-th step prediction for all images' importance score in album $S$. $\check{P} = [\check{p}_1, ...\check{p}_C]$ denotes the album event type prediction from CNN. Thus, the (k+1)-step album-wise prediction of event type is the average of each image's prediction averaged by the predicted image importance score. $\alpha$ is a parameter to control the strength importance score has for the update.

2. ~~**Correct**~~ **event type prediction** with $\hat{P}$.

$$p_c(k+1) = \frac{1}{2}(\check{p}_c(k+1) + \hat{p}_c) \tag{2}$$

   Thus, the prediction of LSTM network is involved to ~~make correction of~~ the event type prediction.

3. **Update image importance score prediction** with the event type prediction $P(k+1)$.

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

where $\circ$ is denotes the element-wise matrix multiplication. Thus, the update image importance is acquired from the average of the importance score given different event types, weighted by the event type possibility.
$\mathrm{I}\{p_c^n \geq m \cdot \max_{c'}(p_{c'}^n)\}_{n,c}$ denotes the binary mask that eliminates the event type predictions with low confidence, and makes sure only event types with high probability contribute to the image importance update.

The algorithm is initialized with equal image importance score for each image in an album. By iteratively conducting the procedures in Equation 1-3, we obtain the album-wise event prediction $P$ and image importance score prediction $V$.

Note that procedure is not guaranteed to converge. In the case of oscillation between states, we set a maximum number of iterations, and when the iteration number hits the threshold, predictions for $Q$ and $V$ are obtained by averaging over the three previous steps.

The entire system is illustrated in Figure 2.

## 5 Experiments

In this section, we evaluate our approach for both event recognition and image importance prediction on the ML-CUFED Dataset, and we compare our event recognition result with [9] on the PEC Dataset.

### 5.1 Baselines

Our joint recognition-curation method produces two outputs: album event type prediction, and image event-importance prediction. For event recognition, we compare our result with the baseline by training a single CNN to classify each input image into event types. We denote this method CNN-recognition. 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 a single CNN to predict the event type for a single image. To predict the event type of one album, predictions of all the images are averaged.
- **CNN-Iterative**: Use the event type prediction from the CNN-recognition network and the image importance prediction result from the Siamese Network to jointly obtain the modified recognition prediction. Here, no LSTM network result is involved.
- **CNN-LSTM**: Use the feature extracted from CNN-recognition as input to an LSTM network, and directly predict the album event type given the input image sequence.
- **CNN-LSTM-Iterative**: proposed method as described in Section 4, using CNN-recognition and LSTM to predict event type label, Siamese network to predict the importance of images, and use iterative update rule to obtain the final prediction for the album-wise event type prediction.

To evaluate our image importance prediction result, we provide a comparison with several baselines:

- **CNN-Noevent**: Train a Siamese Network to predict the importance score difference of an input image pair. No event type information is given to the network while training or testing. All albums are considered having the same "hyper" event type.
- **CNN-Noevent(test)**: When training the Siamese Network, use the ground-truth event type information to gate the output error and back-propagation signal. In this way, features for each event type are learned separately. When testing, given an image, average the predicted importance score for all possible event types.
- **CNN-LSTM-Iterative**: Our proposed joint curation-recognition method as in Sectoin 4.

### 5.2 Experimental Details

**Dataset** For ML-CUFED Dataset, we split the albums to training:test 4:1. The test set has 368 albums. To decide the hyper-parameters $(\alpha, m)$ in our iterative model, a validation set with 119 albums is separated from the training set. For the PEC Event Recognition Dataset [9], we use directly the test set consisting of 10 albums for each event type as described in [9], so that we can directly compare it with our methods.

**Parameter setting** For both the CNN-recognition network and the Siamese network, we use the 8-layer AlexNet [17] pretrained on ImageNet as the starting point, and fine-tune it on the ML-CUFED Dataset. We use a similar training scheme to [27], but with a lower learning rate of 0.001. For the LSTM network, we use the 7th layer feature from the trained CNN-recognition as the input. The 7th layer feature is 4096-dimensional, and we reduce it to 128-dimensional with PCA. For the LSTM network, we use AdaDalta as optimization method [28–30]. For the Siamese network, we follow the settings in [13] and choose the two thresholds as $m_s = 0.1$ and $m_d = 0.3$. We set the nubmer of iteration of our joint recognition curation algorithm to 10.

**Evaluation** For event recognition, we use two metrics to evaluate different methods: average accuracy and $\mathrm{F}_1$ Score. $\mathrm{F}_1$ Score is the harmonic mean of precision and recall, and can account for multi-label groundtruth. For ML-CUFED where multi-label ground-truth is allowed, both accuracy and $\mathrm{F}_1$ are calculated with top-1 prediction. For image importance prediction, we follow the evaluation metrics in [13], using MAP@($t$%) and Precision P@($t$%). Precision is calculated as the ratio between the number of retrieved relevant images and the number of all relevant images. MAP is the averaged area under the precision-recall curve, and can be calculated:

$$\mathrm{AP(S)}@t\% = \int_0^1 p(r)d(r) \approx \frac{\sum_{k=1}^{n} p(k) \times \mathrm{rel}(k)}{\lceil n \cdot t\% \rceil} \tag{4}$$

$$\mathrm{MAP(U)@}t\% = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP(S}_i\mathrm{)@}t\% \tag{5}$$

### 5.3 Results on the ML-CUFED Dataset

**Event-specific image importance**

For the event-specific image importance score prediction task, we provide the comparison of our methods with several baselines in Table 1. To show the upper-bound of our method, we also show the result for CNN-Cheating: where we assume the ground truth even type is known for each test album, and use the predicted importance score for that ground-truth event type as the final importance prediction. Note that in our setting we assume only albums are given, and no additional information is added. Therefore, the result of CNN-Cheating here just serves as the best result we can get when the event recognition stage is perfect. The CNN-Cheating result is very similar to that achieved in [13]. As shown in Table 1, CNN-Noevent performs a little better than CNN-Noevent (test). This suggests the divergence of the importance prediction for different event types. CNN-Iterative outperforms CNN-Noevent greatly, with a steady 3% MAP increment. There is also a steady 3% increment for P at $t < 20\%$.

CNN-Iterative greatly reduces the performance loss from knowing the event types (CNN-Cheating) to not knowing the event types (CNN-Noevent), and approaches closely the upper bound (CNN-Cheating).

| | **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) | 0.251 | 0.303 | 0.358 | 0.414 | 0.462 | 0.508 | 0.142 | 0.211 | 0.284 | 0.335 | 0.384 | 0.436 |
| CNN-Noevent | 0.258 | 0.316 | 0.369 | 0.425 | 0.475 | 0.519 | 0.168 | 0.245 | 0.307 | 0.373 | 0.422 | 0.468 |
| CNN-Iterative | 0.278 | 0.347 | 0.400 | 0.453 | 0.502 | 0.547 | 0.191 | 0.280 | 0.340 | 0.394 | 0.450 | 0.491 |
| CNN-Cheating | 0.305 | 0.372 | 0.424 | 0.476 | 0.522 | 0.565 | 0.218 | 0.304 | 0.361 | 0.417 | 0.461 | 0.504 |

Table 1: Comparison of event-specific image importance predictions using different methods. Evaluation metric here is MAP@$t$% and $P$@$t$%. We also show Random ranking score as a lower bound. We also provide a CNN-cheating result which uses ground-truth event type information when testing.

**Event recognition**

Table 2 shows the results of different methods for event recognition. Note that ML-CUFED allows multiple labels, therefore the accuracy here is different from the conventional accuracy: for an album, we only evaluate the top 1 prediction. If the predicted event type is among the ground-truth event label, we deem this album correctly predicted. As shown, there is more than 4% performance gain on the average accuracy, and 4% gain on $F_1$ Score, from the baseline CNN-recognition method to our proposed method. 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.

| Method | Avg. Acc. | $F_1$-Score |
|---|---|---|
| CNN-recognition | 75.0% | 0.698 |
| CNN-Iterative | 78.0% | 0.729 |
| CNN-LSTM | 76.6% | 0.713 |
| CNN-LSTM-Iterative | 79.1% | 0.737 |

Table 2: Comparison of different methods on ML-CUFED Dataset.

Figure 3 provides the confusion matrix of the baseline CNN-recognition and proposed method CNN-LSTM-Iterative. To allow for multi-label, each pair of (ground truth label, prediction) counts for one point on the confusion matrix. We can observe overall performance gain from baseline method to CNN-LSTM-Iterative, and the overall reduction of confusions from baseline to proposed method. Note though there are two categories with 0% accuracy: Personal Art Activity, and Personal Musical Activity. This is caused by the intrinsic high ambiguity of these two event types with other event types, and the lack of training data for these two even types (there are only 23 albums with Personal Music Activity in the entire dataset, and all of them have multiple votes).

In Figure **??**, we show our event curation and recognition result with a few examples in the ML-CUFED Dataset.

Fig. 3: Confusion matrices for the baseline method(left) and our proposed method CNN-LSTM-Interative (right).

## 5.4 Results on the PEC Dataset

To show the generalizability of our algorithm, we compare our result with [9] on their PEC Dataset.

The PEC Dataset is an 807-album event dataset, with 14 social event classes. There is no ground-truth importance score labels 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 PEC test set containing 10 albums each class, and compare with the result achieved in [9] to make a fair comparison.

In PEC Dataset, there are 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 event in the ML-CUFED dataset. Therefore, we provide the mapping from PEC label to ML-CUFED label in Table 3. Note that the mapping is not perfect. For example, although most albums in Road Trip(PEC) fall into Nature Trip(ML-CU), some of them should fall into Urban Trip(ML-CU), but our mapping ignore this to avoid any bias it might bring to the evaluation. 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) [9] on the test data. Merged labels are straightforward to deal with, just by merging the two corresponding rows in the confusion matrix. However, for disappeared labels, there are three possible approaches: 1. Assume false positive on those labels will be correct predictions if those labels disappear; 2. Assume false positive on those labels will still be incorrect; 3 Ignore false positive on those labels, shrinking the size of the test set. 1 and 2 are two extreme ends, being very loose and very strict respectively, while 3 lies in the middle.

It is also easy to calculate the accuracy of our model: We still take the maximum of predicted probabilities from our model, but ignore those that don't belong in Table 3.

The comparison of different methods is shown in Table 4. For the result of SHMM, we provide all the three criteria mentioned above. It is worth noting that our networks are fully trained on ML-CUFED, and despite the difference between the two datasets (collected from different sources) and the inevitable label mis-matching, our baseline method (CNN-recognition) already outperforms SHMM. Similar to the result on ML-CUFED, we observe the performance gain from both LSTM network and iterative updates.

| | | | | | |
|---|---|---|---|---|---|
| **PEC** | (Children's) Birthday | Christmas | Concert | Cruise | Exhibition |
| **ML-CUFED** | Birthday | Christmas | Show | Cruise | Museum |
| **PEC** | Halloween | Hiking, Road Trip | Wedding | Graduation | |
| **ML-CUFED** | Halloween | Nature Trip | Wedding | Graduation | |

Table 3: Event type matching from PEC Dataset to ML-CUFED Dataset.

We show some of the examples of our recognition and event-specific image importance prediction result in Figure 4. Although there is no ground-truth labeling for the event-specific importance score for the images, we can look at the sample results qualitatively. From the first row in Figure 4, we can see that the model does not simply assign a high importance score to the characteristic Christmas tree, but predicts a very high score when there are people sitting in

| Method | Avg. Acc. | $F_1$-Score |
|---|---|---|
| SHMM | 76.3(1) / 59.1(2) / 72.2(3)% | - |
| CNN-recognition | 80.9% | 0.805 |
| CNN-Iterative | 81.8% | 0.810 |
| CNN-LSTM | 82.7% | 0.824 |
| CNN-LSTM-Iterative | 84.5% | 0.842 |

Table 4: Comparison of different methods on PEC Dataset. For the SHMM result, accuracies under three criteria are provided. See Section 5.4 for details.

front of the Christmas tree. From the second row, we can see the property of the event modulated importance score: although the first image of the beautiful architecture should have a high importance in a trip album, in the birthday album, it is much less important even than the image of the table of celebrative food.

Fig. 4: Some results from the PEC Dataset. For first two rows, the prediction from CNN-Direct is wrong as shown in the label above, whereas the sequential information helps the algorithm to modify the final prediction to give the correct answer. For the third row, a case where the model fails to recognize a Christmas event is shown. The predicted importance score of each image is shown below the corresponding image. Only a subset of images is shown here for each album due to limited space.

## 6 Conclusion

# References


1. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., Quéenot, G., Ordelman, R.: Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015, NIST, USA (2015)
2. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: CVPR. (2012)
3. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative cnn video representation for event detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1798–1807
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
5. Reiter, S., Schuller, B.W., Rigoll, G.: A combined LSTM-RNN - HMM - approach for meeting event segmentation and recognition. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006. (2006) 393–396
6. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (2007)
7. Park, S., Kwak, N.: Cultural event recognition by subregion classification with convolutional neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2015)
8. Salvador, A., Zeppelzauer, M., Manchon-Vizuete, D., Calafell-Orós, A., Giró-i Nieto, X.: Cultural event recognition with visual convnets and temporal models. In: CVPR ChaLearn Looking at People Workshop 2015. (06/2015 2015)
9. Bossard, L., Guillaumin, M., Van, L.: Event recognition in photo collections with a stopwatch hmm. In: Computer Vision (ICCV), 2013 IEEE International Conference on. (2013)
10. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: Rating pictorial aesthetics using deep learning. In: Proceedings of the ACM International Conference on Multimedia. (2014)
11. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
12. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
13. Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G, W.: Event-specific image importance. In: In Proc. Computer Vision and Pattern Recognition Conference (CVPR). (2016)
14. R.MattiviandJ.R.R.UijlingsandF.deNataleandN.Sebe: Exploitation of time constraints for (sub-) event recognition. In: ACM Workshop on Modeling and Representing Events (J-MRE 11). (2011)
15. Tsai, S., Cao, L., Tang, F., Huang, T.S.: Compositional object pattern: a new model for album event recognition. In: Proceedings of the 19th International Conference on Multimedia 2011. (2011)
16. Imran, N., Liu, J., Luo, J., Shah, M.: Event recognition from photo collections via pagerank. In: Proceedings of the 17th ACM International Conference on Multimedia. MM '09, New York, NY, USA, ACM (2009) 621–624
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. (2012)

18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
20. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems 27 (NIPS). (2014)
21. Xiong, Y., Zhu, K., Lin, D., Tang, X.: Recognize complex events from static images by fusing deep channels. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CVPR (to appear) (November 2015)
23. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. CoRR **abs/1312.4894** (2013)
24. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015)
25. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2009)
26. Sak, H., Senior, A.W., Beaufays, F.: 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
27. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
28. Zeiler, M.D.: Adadelta: An adaptive learning rate method. CoRR **abs/1212.5701** (2012)
29. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference (SciPy). (June 2010) Oral Presentation.
30. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y.: Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop (2012)