Event-specific Image Importance

Yufei Wang[1] Zhe Lin[2] Xiaohui Shen[2] Radomír Měch[2] Gavin Miller[2] Garrison W. Cottrell[1]
[1]University of California, San Diego [2]Adobe Research
{yuw176, gary}@ucsd.edu {zlin, xshen, rmech, gmiller}@adobe.com

Abstract

When creating a photo album of an event, people typically select a few important images to keep or share. There is some consistency in the process of choosing the important images, and discarding the unimportant ones. Modeling this selection process will assist automatic photo selection and album summarization. In this paper, we show that the selection of important images is consistent among different viewers, and that this selection process is related to the event type of the album. We introduce the concept of event-specific image importance. We collected a new event album dataset with human annotation of the relative image importance with each event album. We also propose a Convolutional Neural Network (CNN) based method to predict the image importance score of a given event album, using a novel rank loss function and a progressive training scheme. Results demonstrate that our method significantly outperforms various baseline methods.

1. Introduction

With the proliferation of cameras (in cell phones and other portable cameras), taking photographs is practically effortless, and happens frequently in everyday life. When attending an event, for instance, a Thanksgiving holiday, participants often take many photos recording every interesting moment during the event. This leads to an oversized album at the end of the event. When we need to simplify the album before saving to a device, or if we want to make a photo collage or a photo book to share our important moment with others, we have to go through the tedious and time-consuming work of selecting important images from a large album. Therefore, it is desirable to perform this task automatically.

Automatic photo selection or album summarization has been studied by some researchers [29, 20, 23, 3, 30]. They aim at personal event albums, and visual content information as well as diversity and coverage is often considered jointly to obtain a summarization. However, these works ignored the role of the event type in the selection process. Intuitively, the event type of the album is an important criterion when we select important images. For example, if we need to select important photos from a vacation to Hawaii, the photo of the volcano on the Big Island is definitely important to keep, whereas if the album is a wedding ceremony, beautiful scenes are only background and are not likely to be more important than the shot of the bride and groom.

In this paper, we introduce the concept of event-specific image importance. It is different from general image interestingness or aesthetics, in that it is contextual, and is based on the album the image is in. We focus on the event-specific importance score of a single image, and do not consider summarization problems where diversity and coverage are also important: Image importance prediction is the most challenging and crucial part of the event curation/album summarization process; Moreover, the importance score can be directly applied to to any album summarization algorithm. We collect an event-specific image importance dataset from human annotators, and we show that the event-specific importance is subjective yet predictable. Finally, we provide a method for predicting event-specific image importance using Convolutional Neural Network (CNN). We propose a new loss function and training procedure, and our CNN method greatly outperforms different baselines.

2. Related Work

Image properties. Importance of an image is a complex image property, and is related to many other image properties. Many image properties can be viewed as cues when selecting important images, such as memorability [13, 12], specificity [14], popularity [16], aesthetics and interestingness [7, 10]. Those image properties are correlated to image contents, such as high level features: object and scene categories [13, 12, 14, 16, 7], and low level features: texture, edge distribution, etc. [10, 16]. In this work, rather than the general image properties mentioned above, we study event-specific image importance, which summarizes human preferences related to images within the context of an album, where the album is of a known event type.

Convolutional Neural Networks(CNNs). The development of methods for training deep CNNs has led to rapid progress in many computer vision tasks in recent years. Substantial improvements have been made in basic computer vision problems such as image classification [17, 26], object detection [9, 4] and scene recognition [33, 8]. Now, there is a greater focus on learning higher-level image properties. One example closely related to our project is Xiong *et al*.'s work on event recognition from static images [31]. In this work, the network is divided into different channels, creating human and object maps that are then fused with the original images to jointly train a deep architecture that predicts the event type from a single image. Our model also uses deep representations to capture event features, but our focus is on event curation rather than event recognition. In fact, our model assumes that the event type is known. Event curation then requires choosing the most important images for the event in question.

Album summarization and photo selection. The most closely related work to our project is on summarization and selection from an album or several albums.

Event summarization of public photo/video collections involves selecting the most important moments of a social event from a variety sources on the web [6, 21]. Here, the goal is to retrieve all of the important moments (diversity), while covering the whole event (coverage). More relevant to this project is work that attempts to summarize a single album [29, 20, 23]. Again, coverage and diversity of the albums are considered, and single image importance is used as a cue [20, 23]. Sinha *et al*. aim at summarization of personal photo collections taken over a long time span, taking the event type as one photo descriptor to calculate diversity and coverage of the photo subset [23].

For the photo selection problem, Yeh *et al*. proposed a ranking system for photographs based on a set of aesthetic rules and personal preferences [32]. Walber *et al*. use gaze information from user's photo viewing process to assist the automatic photo selection algorithm, so this work requires eyetracking [30]. The work by Ceroni *et al*. [3] is probably most relevant to our work. It focuses on selection of important photos from a single event album, and different factors are considered: image quality, presence of faces, concept features, and collection based features such as album size. However, each album used for training and testing in this work is collected from a single participant, and the important subset is picked by the same person: it does not focus on common human preferences. Moreover, the prediction algorithm is tested on unseen images in the same album used for training, and it does not focus on new album prediction.

Our work differs from all the above in that we focus on: i) whether humans have common preferences for image importance/preference scores, ii) whether image importance can be predicted for unseen albums with widely varying content, and iii) whether event type information is important for the prediction. To summarize, we are introducing a subjective but predictable image property: event-specific image importance, and we propose a method to predict this property.

3. The Curation of Flickr Events Dataset

Are people's ratings for images in albums representing particular events predictable? Our intuition is that in an album of a certain event type, there will be a consistent subset of images that will be preferred by most people. However, there is no available dataset to verify this intuition, or to test the degree of people's agreement on this highly subjective task. In this section, we describe the collection of the CUration of Flickr Events Dataset (CUFED), and measure the consistency of human subjects' preferences on this dataset. CUFED provides a ground truth dataset that allows us to measure the predictability of human rated image importance scores, and to develop our prediction model. CUFED is publicly available.[1]

3.1. Album Collection

In order to collect a dataset of albums of different event types, we segmented albums from the Yahoo Flickr Creative Commons 100M Dataset (YFCC100M) [28]. The YFCC100M Dataset has 100 million images and videos from Flickr. In this collection, each image has the following metadata: the user ID who uploaded this photo; the time the image was taken; and often there are user tags. We took advantage of the metadata to segment dataset into albums: For each photo uploader, events are segmented based on timestamps and tags: images taken within short time interval (3 hours) and with more than 1/3 common tags belong to one event. Using tags to filter the data was inspired by the observation that users tend to give the same tags to an event album instead of individually tagging every single image in it. Using this approach, we segmented 1.8 million albums from the YFCC100M dataset. Here, we randomly selected 20,000 albums to work with.

To get the event type of those albums, we presented the albums to workers on Amazon Merchanical Turk (AMT) and asked them to classify the albums into 23 event types. Aside from these event types, the workers could choose "Other events", "Not an event", "More than a single event" or "Cannot decide" instead of available event types. We chose our 23 event types so that they cover the most common events in our lives, ranging from weddings to sports games. All 23 event types are shown in Table 1. Each album was labeled by 3 workers. Over 82% of the 20k albums received the same labels from at least 2 of the 3 workers. We

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

Categories	**Important Personal Event**	**Personal Activity**	**Personal Trip**	**Holiday**
Event types and # albums	Wedding:198 (98%) Birthday:180 (91%) Graduation:178 (88%)	Protest:50 (92%) Personal Music Activity:25 (92%) Religious Activity:50 (90%) Casual Family Gather:50 (84%) Group Activity:50 (82%) Personal Sports:100 (78%) Business Activity:50 (76%) Personal Art Activity:54 (70%)	Architecture/Art:50 (92%) Urban Trip:100 (89%) Cruise Trip:50 (88%) Nature Trip:50 (86%) Theme Park:100 (86%) Zoo:99 (85%) Museum:50 (84%) Beach Trip:50 (82%) Show:100 (82%) Sports Game:50 (58%)	Christmas:100 (87%) Halloween:99 (86%)

Table 1: 23 Event types, their corresponding number of albums, and percentage of significant albums at level $q = 0.05$ using Kendall's W statistics. The event types fall into four categories.

kept the albums which were given the same label by 2 or more workers, and this label was given to the album. This resulted in 16,489 albums.

We further randomly selected 50-200 albums from each of the event types (except for *Personal Music Activity*, which has 25 albums), resulting in a dataset of 1883 albums. The number of events of each type is shown in Table 1. The size of the albums varies between 30 and 100 images. We chose these parameters by hand to emphasize our intuition that some event types will have more consistent ratings, and hence more predictability, than others. Therefore, in this dataset, we emphasized those events in hope of learning more from them.

3.2. Data Annotation

In order to get the rating for each image in an album, we presented an album together with its event type to AMT workers and let them rate each image in that album as very important, important, neutral, or irrelevant. The four ratings are mapped to scores {2,1,0,-2} when creating ground truth. We intentionally did not give specific criteria for the rating levels, to encourage the workers to rate based on their intuition. In our pilot study, workers on AMT tended to mark a large proportion of the images as very important/important. This is understandable, since most of the albums are of high quality, but it leads to a ceiling effect on the ratings. To control the size of images marked as important, we forced the workers to label 5%-30% of the images as very important, and 10%-50% as important. The average time to rate each image was 7.7 seconds. Each album was annotated by 5 distinct workers. 292 workers participated in the tasks. Over 90% of our data was annotated by 93 workers.

The image ratings collected from AMT differed in quality among different AMT workers. To avoid low quality work, only workers who passed an event recognition test using albums could proceed to the real task. In addition, we added two distractor images per album which were clearly not related to the event in order to screen workers who were not paying attention. However, it is not possible to assure the quality of an individual submission because of the subjective nature of the image importance rating task. Therefore, in order to filter "bad" submissions, we found workers who consistently gave scores far from others and filtered out their submissions. If more than 30% of his/her submissions had a euclidean distance from the average of other workers' submissions greater than a threshold, that worker's submissions were filtered out. Only two workers were filtered out in this way.

3.3. Consistency Analysis

To examine the consistency of the human ratings of images, we split our subjects into two independent groups of two and three raters for each album, and used Spearman's rank correlation (ρ) to evaluate their consistency. ρ ranges from -1 (perfectly inverse correlation) to 1 (perfect correlation), while 0 indicates no correlation. For each album, we averaged the correlation scores of all possible random splits. The average correlation over all albums was 0.40.

We further evaluated the annotation consistency with Kendall's W, which directly calculates the agreement among multiple raters, and accounts for tied ranks. Kendall's W ranges from 0 (no agreement) to 1 (complete agreement). Note that in our workers' rating of one album, tied ranks are very frequent, since there are only 4 possible ratings, and the average album size is 52. Coincidentally, the average Kendall's W over all albums was also 0.40. Both Spearman's rank correlation ρ and Kendall's W showed significant consistency across subjects despite the high subjectivity of this problem.

To test the statistical significance of Kendall's W score, we did a permutation test over W to obtain the distribution of W under the null hypothesis, and for each event type, we used the Benjamini-Hochberg procedure to control the false discovery rate (FDR) for multiple comparisons [2]. At level $q = 0.05$, 86% of albums had significant agreement on average. Table 1 shows the percentage of albums with significant agreement for each event type. The different percentages of significant albums in different event types confirmed our intuition that some event types would be more consistently rated than others. The wedding event was the most consistently rated, with 98% of albums being significantly consistent, while for the sports game category, only 58% of the albums received significant consistency scores, the lowest among the 23 events. In the supplementary material, we include examples of albums that received high and low consistency ratings.

4. Approach

In this section, we propose a Convolutional Neural Network (CNN) based method for estimating an event-specific image importance score in an album, given the event type of this album. We use a siamese network [24] with a novel rank loss function to take two images at a time and rank them relative to one another based on their scores.

4.1. CNN Structure

The design of our siamese CNN architecture is shown in Fig. 1. It has several properties described in the following subsections.

4.1.1 Feature sharing among event types

We train a single siamese network with albums from all event types. The last layer, however, has separate outputs for each event type. The reasons are as follows. First, there exists strong visual similarity among different event types in terms of image importance, therefore for a specific event type, labeled data from other event types will help as implicit data augmentation. Second, feature sharing will significantly reduce the number of parameters in the network and regularize the network training. Especially for our problem, high variance among albums within each event type and relatively small datasets make this even more necessary. Therefore, in our network, all event types share the features, while the output level has event-specific ratings. During the training process, only the output corresponding to the event type of an image pair receives an error signal, and we assume that we know the event type at test time.

4.1.2 2-stage progressive training

Due to the large variation among albums and the relatively small scale of the dataset (especially for some event types such as *casual family/friends gathering*), directly training a CNN for separate event types as in Section 4.1.1 may lead to over-fitting for some event types with less training data. Therefore, we use a 2-stage progressive learning method: we train all images with one output for the whole network; and then switch to training with a separate output for each event type. Initialization of the second stage is done using the network from the first stage. This helps in that i) the features that are useful for all event types are learned first, using all of the data; ii) the individual event-type output units are initialized with the weights from the one-output unit, so they already have some knowledge of what makes an important image; and iii) the discrimination is then refined based on the properties of individual event types. Some pictures are just excellent no matter what the occasion; our two-stage learning system leverages that intuition[2].

4.1.3 Siamese architecture

There is large variation in the quality of the albums within an event type, which might bias the judgment of participants in our AMT task. Therefore it is difficult to learn a reliable absolute image importance score that is suitable for different albums. Meanwhile, the relative importance ranking of images within the same album is more meaningful and more practical in applications. Hence, rather than training on an absolute image score, we use the average score difference between a pair of images from the same album to train the network. This is the motivation for using the siamese network architecture [24], which processes pairs of images. In the siamese network, the two pathways share weights, so a common representation is learned (see Fig. 1).

4.1.4 Piecewise ranking loss

For each input image pair to the network $(\mathrm{I}_1, \mathrm{I}_2)$, $G(\mathrm{I}_i)$ is the ground truth score of image I_i, and $P(\mathrm{I}_i)$ is its predicted score from the network. We use a piecewise ranking loss (PR loss) to train the network:

$$\mathrm{PR} = \begin{cases} \frac{1}{2}\max(0, |D_\mathrm{p}| - \mathrm{m_s})^2 & \text{if } D_\mathrm{g} < \mathrm{m_s} \\ \frac{1}{2}\left\{\max(0, \mathrm{m_s} - D_\mathrm{p})^2 + \max(0, D_\mathrm{p} - \mathrm{m_d})^2\right\} & \\ & \text{if } \mathrm{m_s} \le D_\mathrm{g} \le \mathrm{m_d} \\ \frac{1}{2}\max(0, \mathrm{m_d} - D_\mathrm{p})^2 & \text{if } D_\mathrm{g} > \mathrm{m_d} \end{cases} \tag{1}$$

where $D_\mathrm{g} = G(\mathrm{I}_1) - G(\mathrm{I}_2)$ is the ground truth score difference between the input image pair, and $D_\mathrm{p} = P(\mathrm{I}_1) - P(\mathrm{I}_2)$ is the predicted score difference. $\mathrm{m_s}$ and $\mathrm{m_d}$ are predefined values for similar and different margins. In Equation 1, several conditions are considered:

- When $D_\mathrm{g} > \mathrm{m_d}$, the loss function reduces to a variation of ranking SVM hinge loss [5]. We use L-2 loss which penalizes high errors more heavily than traditional hinge loss [27]. This is similar to contrastive loss function when the input pair of images are deemed dissimilar [11], but we are not using the euclidean distance of the output of the network, since the sign of D_p is important here.

[2] We also tried to cluster the event types into k "superclasses" according to their similarity, and to use the superclass information for the first stage training. However, that didn't lead to a better result. One possible reason is that our event type clustering algorithm does not perform well.

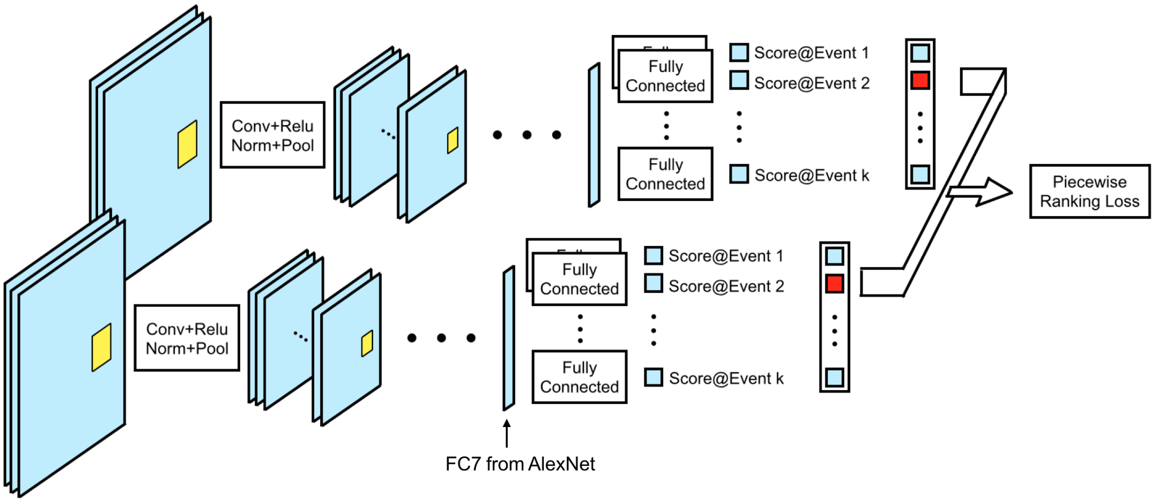

Figure 1: A siamese CNN architecture for joint training over events. A pair of images from the same album is the input to the two pathways. Intermediate layers are omitted here for simplicity. The network computes an importance score for its input image; only the units corresponding to the correct event type are activated and back-propagated through (the red square represents a mask).

- When $D_{\mathrm{g}} < \mathrm{m_s}$, the loss function reduces to a variation of contrastive loss when the input pair is deemed similar [11]. In addition to the contrastive loss in [11], we introduce a margin: $\mathrm{m_s}$. The margin serves as a slack term. The reason to have it is that the ground truth importance score is acquired from a group of humans, and the variance is relatively high among the humans, as shown in Section 3.3. The introduction of relaxation with $\mathrm{m_s}$ makes the network less sensitive to this variance in our ground truth.
- When $\mathrm{m_s} < D_{\mathrm{g}} < \mathrm{m_d}$, the loss function will only penalize the D_{p} not being in the same range with D_{g}. This pulls D_{p} towards D_{g} when the image pair is similar in rating, reducing the loss function's vulnerability to the variance in our ground truth.

The PR objective loss function has the following advantages: Rather than training only on images with different ratings, it provides an error signal even when image pairs have the same rating, moving them closer together in representational space. This makes full use of the training dataset. Our piecewise version also introduces relaxation in the ground truth score, thus making the network more stable, which is beneficial when the ratings are subjective.

4.2. Incorporating Face Heatmaps

Images with faces tend to be more interesting than images without them [23]. Moreover, our intuition is that in an event album, important people will appear more frequently. This across-album feature cannot be captured by a CNN trained with image pairs. In order to incorporate face information, we generate face heatmaps, and use them to train a shallow CNN to independently predict the importance score of the photos. A separate face heatmap-based score enables flexible tuning of the relative strength of the two scores from original images and face heatmaps.

To generate the face heatmaps, we use a state-of-the-art face detection network [18]. In order modulate the heatmaps according to face frequency, we need facial identity information. We train 18 CNN models for different face parts and concatenate the final fully-connected layers as the final face descriptor, following a similar pipeline as [25]. We then do agglomerative identity clustering to obtain the frequency of faces in an album. In the face heatmap, faces are represented with Gaussian kernels, and the two most frequent faces are emphasized by doubling their peak values. These are used as input to a shallow siamese CNN trained from scratch, with one convolutional layer and two fully connected hidden layers, in the same manner as the image network. Details of the architecture are described in the supplementary material.

Examples of face heatmaps are shown in Fig. 2. In the testing stage, the prediction from the original image and the face heatmap network are combined according to the following formula:

$$P = P_{\mathrm{I}} + \lambda \cdot \min\left\{\max\left\{P_{\mathrm{f}}, \beta\right\}, \alpha\right\} \qquad (2)$$

where $(\mathrm{P_I}, \mathrm{P_f})$ are predicted scores from the original photo network and face heatmap network respectively. The face heatmap contains a limited information, therefore we constrain the effect of the face heatmap for the final prediction with (α, β), so that extreme predictions from the face heatmap are eliminated; λ is also used to further control the effect of the face heatmap-based prediction. These parameters are set using cross-validation and a grid search.

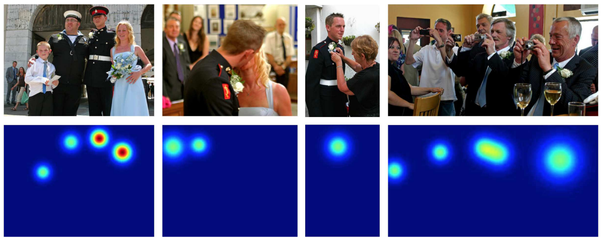

Figure 2: Face heatmaps from a wedding event album. First row: original images; Second row: face heatmaps. Faces of the two most important people have higher peak values (red dots). The second column shows that face detection is not ideal; the third column shows that identity clustering is not perfect, as the groom is not emphasized.

5. Experimental Results

In this section, we compare our results with several baseline methods.

5.1. Experimental Settings

Dataset For training and testing, we randomly split the Curation of Flickr Events Dataset into 3:1 albums for every event type. The training set consists of 1404 albums, and the test set has 479 albums.

Parameter setting We use Alexnet to initialize the CNN architecture and then fine-tune it [17, 15]. In Fig. 1, FC7 is from Alexnet, driving the event-specific sigmoidal score prediction layer. We assume we know the event type, and the teaching signal is masked by the correct event. For PR loss, we set $\mathrm{m_s} = 0.1$ and $\mathrm{m_d} = 0.3$. For training parameters, we use the default settings for pre-training in Caffe [15], but we start from a smaller learning rate of 0.001 [9].

We follow [17]'s data augmentation approach: Input images are resized to 256×256. During the training stage, images are randomly cropped to 227×227 crops, and there is a 50% probability that input images are horizontally flipped. In the test stage, predictions are averaged on five crops (four corners and the center) and their horizontal reflections. We train five different CNNs with 5-fold cross validation, and use an ensemble of the five networks for the final prediction.

Evaluation metrics We use two evaluation methods to compare the different approaches. For both evaluation methods, we assume that given an event album, we view the top $t\%$ images as relevant images, and measure the metric at various values of t.

First, we use mean average precision (MAP) to evaluate our models. MAP is a common evaluation method for information retrieval [1]. It is the averaged area under the precision-recall curve over all albums. Given the collection of albums, and top $t\%$ of the images as being relevant images, MAP@($t\%$) can be calculated:

$$\text{AP(S)@}t\% = \int_0^1 p(r)d(r) \approx \frac{\sum_{k=1}^{n} p(k) \times \text{rel}(k)}{\lceil n \cdot t\% \rceil} \quad (3)$$

$$\text{MAP(U)@}t\% = \frac{1}{N} \sum_{i=1}^{N} \text{AP(S}_i\text{)@}t\% \quad (4)$$

where S_i is the ith album, and U is the collection of all albums. n is the size of album S, $p(k)$ is the precision at rank k, and rel is an indicator of whether the kth ranked image from our algorithm is a relevant image, i.e. among the top $t\%$ ground truth.

Second, we calculate the precision (P), the ratio between the number of relevant photos in the retrieved images over the total number of relevant images at each level of t. Unlike MAP, P cares entirely about how many important images can be retrieved at a cut-off level, and does not care about the position they are in the retrieval list, or where the rest of important images are in the ranking system. Although less informative than MAP, P is also an intuitive way to demonstrate the effectiveness of our predicted image ranking result. Since we are solving an image selection problem, we care more about MAP and P for small $t\%$, so we only present results for $t \leq 30$.

5.2. Results and Analysis

In this section, we compare our method, Piecewise Ranking-CNN trained progressively (PR-CNN(Progressive)), on all event types to various baselines, and demonstrate the advantages of our method. Figure 3 is an example to show how our algorithm performs intuitively. Our result clearly learns meaningful concepts for the wedding event. More examples are shown in the supplementary material. The quantitative comparison of the methods, broken out into each of the 23 event types, is also shown in the supplementary material. In the following sections, we describe the various baseline methods we benchmark our system against. To make a long story short, we achieve our best result using an ensemble of the five PR-CNN(Progressive) networks (See Table 2). To show that the analysis in this section holds for more powerful network architectures, we also provide a comparison of several key methods using VGG network [22], by fine-tuning the fully connected layers, as shown in Table 3.

5.2.1 Does aesthetics play an important role?

In a user study, Walber *et al*. show that that humans use the visual appeal of an image as a criterion for selecting important images in an album [30]. In order to quantify the role attractiveness plays, we use an aesthetic score prediction method instead of the importance score. We train a CNN classifier similar to [19], but using a Siamese architecture with the ranking loss, which we found outperforms the classification loss [19] on aesthetics ranking.

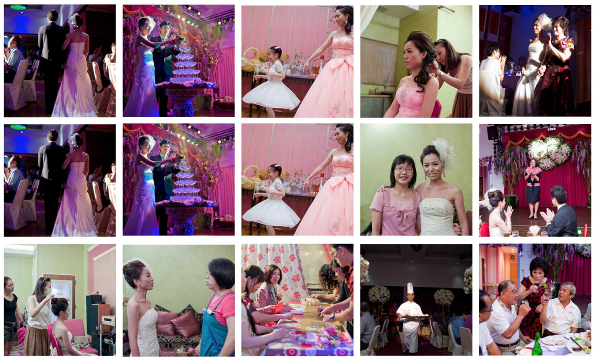

Figure 3: Example results for one wedding album. Top 5 images of the album from different methods are shown here. First row: Ground truth acquired from AMT workers; Second row: Our prediction using Ensemble-CNN; Third row: Random selection.

$t\%$	5	10	15	20
NoEvent-CNN	0.266	0.330	0.383	0.437
SVM-CNN	0.281	0.345	0.401	0.456
PR-CNN(Direct)	0.292	0.354	0.413	0.467
PR-CNN(Progressive)	0.298	0.365	0.420	0.473

Table 3: MAP@$t\%$ for different methods using VGG network architecture.

Table 2 shows that the aesthetic score of images is only slightly better than random. We conclude that aesthetics, at least using this method, is not a very important criterion for human selection of important images in event albums. In the supplementary material, we observe that the aesthetic score is more predictive for some events than others, e.g. *Nature trip*, *Personal art activity* (in which many photos are portrait shots). This is consistent with our intuition: aesthetics is an important criterion for human selection in events without strong narrative structure.

5.2.2 Are pre-trained CNN features useful?

Pre-trained CNN features have been shown to have a high generalization ability to new tasks [4, 9]. Using the FC7 layer of Alexnet [15, 17] as our feature vector, we apply a K-NN classifier and a Ranking-SVM classifier. We also provide unsupervised k-means method for comparison.

For the KNN approach, we perform a 10-nearest neighbors search against all training images in the same event type, and use the weighted average of the 10 images' ground truth importance score, where the weight is the image's similarity score to the query test image. We denote this method as Pre-KNN. We also train 23 Ranking-SVMs (one for each event type) on pairs of the 4096-d feature vectors. This method is denoted Pre-SVM. For the k-means based unsupervised method, we use k-means to partition the photos in a test album into k clusters using pre-trained FC7 features. Here we set k be the 1/10 of the album size. The photos closest to big cluster centers are considered most representative for this album, and are assigned high score. The importance score of an image is proportional to the size of cluster it is in, and is inversely proportional to the distance it is to the cluster center. The result is denoted K-Means.

Table 2 shows the results of using pre-trained CNN features. The unsupervised K-Means slightly outperforms Random by 4% MAP. The KNN method significantly outperforms the aesthetic score and random ranking. However, it is still much lower than our proposed method. This shows that the high variation of albums makes the direct score prediction using images in other albums with similar visual appearance unreliable. The Pre-SVM method performs better than the KNN method, but the improvement is limited.

The results of the above two experiments verify that the pre-trained CNN features can generalize to some extent to the event-based image importance prediction problem.

5.2.3 Is Piecewise Ranking loss necessary?

In order to show the advantage of PR loss, we compare our results with the results trained from a conventional ranking SVM hinge loss. For the SVM ranking loss, the network architecture is exactly the same as our proposed method except for the loss function:

$$\mathrm{L}(\mathrm{I}_1, \mathrm{I}_2) = \max(0, 1 - D_\mathrm{p}) \tag{5}$$

where $D_\mathrm{p} = P(\mathrm{I}_1) - P(\mathrm{I}_2)$ is the predicted score difference between the image pair.

This method is denoted as SVM-CNN. As shown in Table 2, PR loss (PR-CNN(direct)) outperforms Ranking SVM hinge loss (SVM-CNN) especially when $t < 20$. Ranking SVM uses 87% of image pairs as the training data compared to PR loss, because it does not use the image pairs with the same score. The reason for PR's better performance may be due to differences in the loss function or because it has 15% more training data.

We also tried a single network with Euclidean Loss to directly predict the importance of a single image. The results are presented in the supplementary material, but they are consistently worse than SVM-CNN by about 0.6%.

5.2.4 Is event information useful?

In the previous work on album summarization or photo selection, a common approach is to use general image interestingness/quality to represent the image importance score irrespective of the event type of the album [3, 20, 23]. We propose that event type information is an important factor in determining the image importance score, and that using 2-stage learning will help with the prediction. In this section, we verify our proposal by comparing the performance of CNNs trained i) without the event type information, ii) with 2-stage learning, and iii) with only the second stage learning on 23 event types.

We train a CNN with exactly the same architecture and training parameters except that the last layer of each of the

	MAP@$t\%$						P@$t\%$					
$t\%$	5	10	15	20	25	30	5	10	15	20	25	30
Random	0.122	0.164	0.211	0.260	0.305	0.350	0.058	0.093	0.141	0.195	0.251	0.298
Worker	0.328	0.410	0.476	0.531	0.580	0.624	0.242	0.371	0.448	0.505	0.552	0.591
Aesthetic	0.139	0.191	0.242	0.290	0.338	0.384	0.060	0.121	0.176	0.228	0.284	0.335
K-Means	0.154	0.207	0.253	0.302	0.349	0.394	0.067	0.123	0.183	0.240	0.293	0.340
Pre-KNN	0.220	0.276	0.326	0.373	0.419	0.465	0.138	0.216	0.275	0.326	0.372	0.419
Pre-SVM	0.252	0.320	0.370	0.420	0.466	0.512	0.169	0.262	0.318	0.363	0.410	0.458
SVM-CNN	0.266	0.337	0.396	0.451	0.500	0.546	0.172	0.280	0.345	0.402	0.447	0.491
NoEvent-CNN	0.261	0.318	0.369	0.422	0.474	0.520	0.167	0.247	0.310	0.372	0.425	0.468
PR-CNN(Direct)	0.296	0.358	0.410	0.462	0.511	0.557	0.199	0.293	0.352	0.403	0.454	0.498
PR-CNN(Progressive)	0.302	0.361	0.415	0.469	0.517	0.563	0.214	0.296	0.356	0.410	0.458	0.502
Ensemble-CNN	**0.305**	**0.364**	**0.417**	**0.471**	**0.519**	**0.563**	**0.216**	**0.301**	**0.360**	**0.411**	**0.459**	**0.504**

Table 2: Comparison of predictions using different methods. Evaluation metric here is MAP@$t\%$ and P@$t\%$. Random ranking score is also shown as a lower bound.

halves of the siamese network in Fig 1 is one unit, so there is essentially one "superclass" event type. This method is denoted as No Event CNN (NoEvent-CNN). As shown in Table 2, although trained with the same loss, without event type information, the network performs worse than PR-CNN(Progressive) by a large margin of 4% over the MAP scores. In addition, the difference of P@$t\%$ is especially large for smaller t, which is the region of most importance.

We also train a CNN with only the second stage directly on 23 event types, as PR-CNN(Direct). Table 2 shows the performance gain using 2-stage learning is about 0.6% on MAP score. This difference is consistent across our experiments. Again, our best result is with an ensemble of the PR-CNN(Progressive) networks (Ensemble-CNN).

5.2.5 Incorporation of face information

In order to incorporate the face information, we use 5-fold cross validation on the training set to set the parameters $\{\alpha, \beta, \lambda\}$ in Equation 2 using a grid search.

Among 23 event types, only 10 event types show a performance gain after face information is incorporated in the validation set, and thus the face information is used for only these 10 event types on the test set. Table 4 shows the effect of adding face information for some example event types. As shown, for some event types, face information substantially helps performance, while for other event types, face information has little impact, or even harms performance. We present the result for all 10 event types as well as the overall average result on 23 event types in the supplementary material. In summary, counter to our expectation, our method for incorporating face information has little effect on performance, increasing it by about 0.1%, which is not likely to be significant.

$t\%$	5	15	25
Beach Trip	0.353(+0.051)	0.455(+0.022)	0.555(+0.011)
Nature Trip	0.167(+0.008)	0.272(+0.008)	0.369(+0.07)
Group Activity	0.315(+0.003)	0.489(+0.001)	0.586(+0.03)
Halloween	0.315(+0.000)	0.424(+0.001)	0.529(+0.02)
Museum	0.293(-0.010)	0.367(-0.010)	0.453(-0.06)

Table 4: MAP@$t\%$ for the Ensemble-CNN after the using of face information on five event types. Performance gain is shown in parentheses.

6. Conclusion

In this work, we introduce a new image property: event-specific image importance. We provide a new dataset consisting of common personal life events, and we provide human generated image importance score ground truth for the dataset. We provide evidence that although the event-specific image importance score is subjective, it is a well-defined and predictable property: there is consistency among different subjects. We develop a CNN-based system to predict event-specific image importance. We show that although aesthetics is usually considered in an image selection system, it is not the most important criterion for people. More importantly, we also show that the event information is an important criterion when people select important images in an album. In our prediction system, we design a Piecewise Ranking Loss for a dataset with subjective or high variance ground truth, and we use a 2-stage progressive training process to train the network. We show that our system is advantageous over the conventional Ranking SVM loss and training procedure.

This work is the first attempt to predict event-specific image importance. This image property is especially useful in album summarization and image selection from an album. In future work, it will be interesting to further investigate the relationship between event types, and to deal with albums with multiple/ambiguous event types. Also, we plan to develop a curation system based on the image importance score, taking diversity and coverage into consideration.

References

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999. 6

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001. 3

[3] A. Ceroni, V. Solachidis, C. Niederée, O. Papadopoulou, N. Kanhabua, and V. Mezaris. To keep or not to keep: An expectation-oriented photo selection method for personal photo collections. In *Proceedings of the 5th International Conference on Multimedia Retrieval (ICMR)*, 2015. 1, 2, 7

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014. 2, 7

[5] W. Chen, T. yan Liu, Y. Lan, Z. Ma, and H. Li. Ranking measures and loss functions in learning to rank. In *Advances in Neural Information Processing Systems 22 (NIPS)*, 2009. 4

[6] M. Del Fabro, A. Sobe, and L. Böszörmenyi. Summarization of real-life events based on community-contributed content. In *Proceedings of the Fourth International Conferences on Advances in Multimedia (MMEDIA)*, 2012. 2

[7] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[8] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 6, 7

[10] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. J. V. Gool. The interestingness of images. In *IEEE International Conference on Computer Vision, ICCV*, 2013. 1

[11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *In Proc. Computer Vision and Pattern Recognition Conference (CVPR)*, 2006. 4, 5

[12] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011. 1

[13] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[14] M. Jas and D. Parikh. Image specificity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6, 7

[16] A. Khosla, A. D. Sarma, and R. Hamid. What makes an image popular? In *International World Wide Web Conference (WWW)*, 2014. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012. 2, 6, 7

[18] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[19] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, 2014. 6

[20] F. Sadeghi, J. Tena, A. Farhadi, and L. Sigal. Learning to select and order vacation photographs. In *Winter Conference on Applications of Computer Vision (WACV)*, 2015. 1, 2, 7

[21] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *IEEE International Conference on Computer Vision, ICCV*, 2007. 2

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[23] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *ICMR*, 2011. 1, 2, 5, 7

[24] R. H. Sumit Chopra and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. of Computer Vision and Pattern Recognition Conference, IEEE Press*, 2005. 4

[25] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[27] Y. Tang. Deep learning using linear support vector machines. In *Workshop on Representational Learning, ICML*, 2013. 4

[28] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 2

[29] S. Tschiatschek, R. Iyer, H. Wei, and J. Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Neural Information Processing Systems (NIPS)*, 2014. 1, 2

[30] T. C. Walber, A. Scherp, and S. Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014. 1, 2, 6

[31] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[32] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the International Conference on Multimedia*, 2010. 2

[33] 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 27 (NIPS)*, 2014. 2