



HAL
open science

Image Selection in Photo Albums

Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, Kadi Bouatouch

► **To cite this version:**

Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, et al.. Image Selection in Photo Albums. ICMR '18 - International Conference on Multimedia Retrieval, Jun 2018, Yokohama, Japan. pp.397-404, 10.1145/3206025.3206077 . hal-01934286

HAL Id: hal-01934286

<https://inria.hal.science/hal-01934286v1>

Submitted on 25 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Image Selection in Photo Albums

Dmitry Kuzovkin
Technicolor, IRISA,
University of Rennes 1
Rennes, France
dmitry.kuzovkin@technicolor.com

Tania Pouli
Technicolor
Rennes, France
tania.pouli@technicolor.com

Rémi Cozot
IRISA, University of Rennes 1
Rennes, France
remi.cozot@irisa.fr

Olivier Le Meur
IRISA, University of Rennes 1
Rennes, France
olivier.le_meur@irisa.fr

Jonathan Kervec
Technicolor
Rennes, France
jonathan.kervec@technicolor.com

Kadi Bouatouch
IRISA, University of Rennes 1
Rennes, France
kadi.bouatouch@irisa.fr

ABSTRACT

The selection of the best photos in personal albums is a task that is often faced by photographers. This task can become laborious when the photo collection is large and it contains multiple similar photos. Recent advances on image aesthetics and photo importance evaluation has led to the creation of different metrics for automatically assessing a given image. However, these metrics are intended for the independent assessment of an image, without considering the possible context implicitly present within photo albums. In this work, we perform a user study for assessing how users select photos when provided with a complete photo album—a task that better reflects how users may review their personal photos and collections. Using the data provided by our study, we evaluate how existing state-of-the-art photo assessment methods perform relative to user selection, focusing in particular on deep learning based approaches. Finally, we explore a recent framework for adapting independent image scores to collections and evaluate in which scenarios such an adaptation can prove beneficial.

KEYWORDS

Photo selection, Image aesthetics, Image quality

1 INTRODUCTION

Casual photography has in recent years become an essential part of everyday life, where people tend to document each moment of their life through a photo. Not restricted by storage limitations or camera availability, users often take numerous photos of the same life moment. This is further exaggerated by functionalities such as the burst mode available on most modern smartphones, which produce dozens of nearly identical images to ensure that the perfect moment or expression is optimally captured. As a result, users often evade the responsibility of taking one best shot during the action of photo capture itself. Eventually, they might obtain an extremely large collection of photos, where for each significant moment, the best shot has to be chosen among many similar photos.

Assessment of photographs in a photo album is a non-trivial task. The selection of the best photos is a subjective process that is also highly affected by the photo album context. For instance, a particular photo might appear of low quality when observed in an isolated manner, while inside the belonging album, the same photo may be the best candidate when compared with other similar

photos. Thus, the selection of photos in the album is, to a large extent, a comparison-based process.

To facilitate and eventually automate this time-consuming task, computational modelling of these human decisions would be necessary. Recent progress in computer vision and machine learning techniques has led to a wealth of image assessment techniques [8, 20, 26, 28, 29, 31], where an image is usually assigned a ranking score, or a label of high or low aesthetics. Going a step further, certain works have attempted to analyze and understand the image features and general characteristics that affect people’s decisions regarding aesthetic quality or beauty of a photograph [16, 30, 41]. Among other applications, automatic image assessment can be used to assist different tasks, such as image retrieval [12, 15] or automatic video thumbnailing [37].

Although such approaches could help guide automatic photo selection or rating decisions, a few major drawbacks limit their usefulness when applied to users’ photo collections. By their nature, such methods are typically trained or optimized on large, general collections accumulating photographs from multiple users. As such, they inherently represent average user preferences modeled over a large variety of content. Further, the models or features learned in such methods are often biased towards professional level photographs, as these types of images are likely to be preferred by the average user when faced with a varied selection of photographs. Finally, the evaluation of each photo is performed independently, where possible connections to other similar photos in the collection are not taken into account.

Given the above limitations, classic approaches for photo aesthetics assessment can be directly applied to more general tasks, such as image retrieval, but may be less capable to reproduce user selections within a photo album, which is the focus of our work. A framework was recently proposed for adapting a general image quality or aesthetics score to the context of a collection [23]. By clustering images according to their degree of similarity, individual image scores could be scaled, such that final selection preferred images that were the best in their cluster, even if they were assessed as being low quality independently. Nevertheless, this framework was demonstrated only using a sharpness metric as an independent assessment criterion, which covers only one of a multitude of image characteristics considered both by users and image aesthetics models [30, 43].

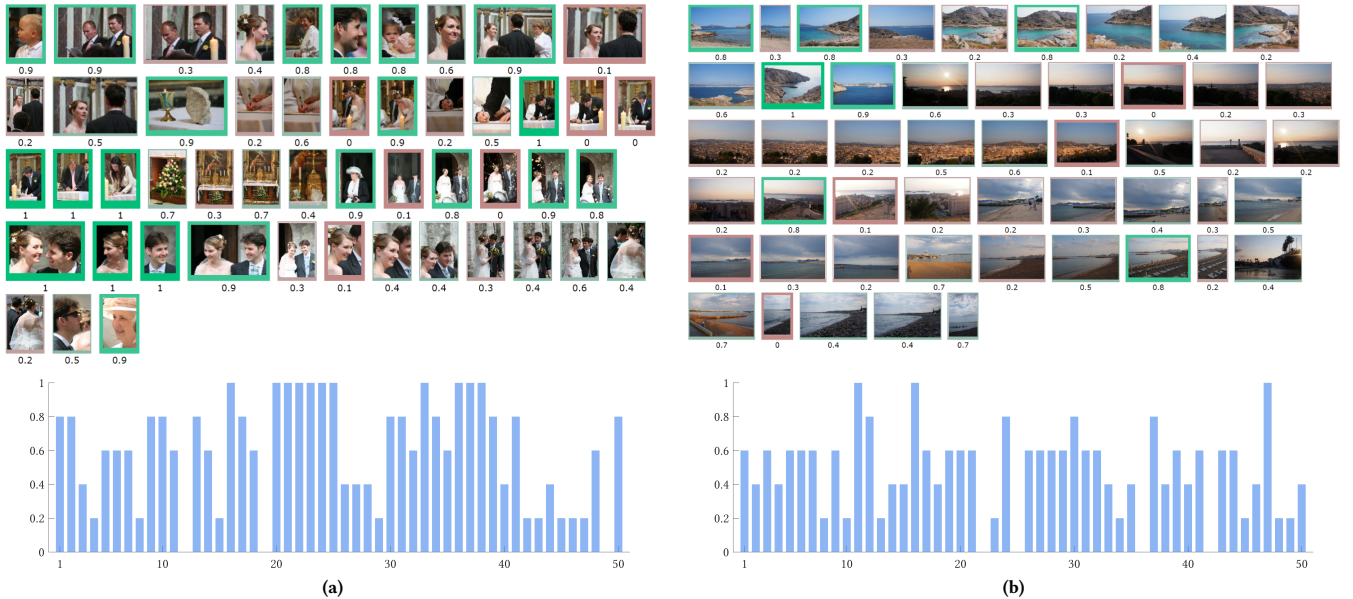


Figure 1: Demonstration of two albums from our user study (*Family event 1* and *Travel album 2*), with a visualization of user preferences shown around each image and confidence bar charts, which indicate confidence of user preferences. (a) Family event with high confidence of user selections. As it can be seen, in the albums with people’s photos, users tend to agree more in their selections. (b) Travel album with lower confidence of user selections. In the albums with landscape photos, users tend to agree less.

To better define how to assess images within the context of their surrounding collection, a deeper understanding of the behavior of users when faced with this task is necessary. To that end, we perform a user study, where, for a variety of photo albums, users have to place themselves in the role of the photographer and identify the images they would like to keep. Our study considers photo albums covering typical scenarios from vacation albums to specific events such as weddings or birthday parties, allowing us to analyze user behavior and cross-user agreement across different situations. Based on this data, we evaluate existing approaches to assess their suitability for selecting photos within collections in a manner consistent with user preferences. Our evaluation considers several state-of-the-art deep-learning methods for image assessment applied to images independently, as well as their adaption with a clustering-based framework [23].

2 RELATED WORK

Photo selection is a complex task, where a selection decision depends on objective characteristics of image quality (e.g. sharpness, dynamic range and presence of artifacts), on subjectively perceived attributes, which make an image attractive (e.g. photo composition and color style), and on semantic aspects, such as presence of important people in a photo and their face expressions [2, 30, 41].

To model human preferences in image evaluation, machine learning techniques are often used as they can learn from pre-evaluated data. Hand-crafted features and different types of image descriptors have been used by several approaches to guide the image analysis and evaluation [8, 28, 31, 32, 35, 36, 39]. Similar to the factors that

affect user decisions, such hand-crafted features may be inspired by photography practices or can take the form of more objective quality metrics. However, these features are not fully capable of modeling and predicting the complex and highly subjective notion of aesthetics, which tends to limit their performance in image evaluation.

To encompass higher level concepts that may influence how people evaluate images, methods for assessing image memorability and interestingness were proposed. The term of memorability is linked with the likelihood that a user will recognize the same photograph after a certain time delay [16, 19]. The term of interestingness is linked with the ability of a certain image or video to draw attention of a user to its content and keep this attention for an extent of time [9, 13]. Such methods are primarily concerned with evaluating the effect a new, unseen image might have to a user, and hence, they might be not directly applicable to the image selection task.

A more recent trend in photo aesthetics assessment is represented by the wealth of deep learning based methods [4, 17, 20, 26, 27, 29, 38, 42], since the task of photo assessment can largely benefit from the abstract feature modeling achieved by convolutional neural networks (CNN). The first CNN-based methods were representing the entire image by a fixed-size cropped patch [26] or multi-patch aggregation models [27], while providing a binary label of low or high aesthetics as an output. Later proposals were able to handle the input images directly, while preserving their aspect ratio, providing a ranking score to an image [17, 29] instead of a binary label. Certain methods combine the estimation of technical image quality along with aesthetics, such as the proposal by Talebi

et al. [38]. The model proposed by Kong et al. [20] can assess multiple meaningful photographic attributes from the content, and then estimate an overall ranking score from them.

Despite the number of studies conducted in the area of independent image assessment, little attention has been paid to context-dependent and personalized assessment of images. While the majority of automatic approaches assess the analyzed photo against non-related photos (or features) from the learned dataset, in a real-life scenario each photo is usually evaluated in comparison with similar photos from the same album. In addition, the actual process of photo selection in photo albums is largely affected by individual preferences, the capturing conditions and other properties associated with a particular album and a particular user.

The usage of individual preferences for image aesthetics assessment was demonstrated in a few recent works. In the work of Yeh et al. [44], the influence of each extracted feature is weighted by the user’s adjustments, either provided manually or learned from a photo example. Adaptation of the general photo assessment model was also proposed by Park et al. [33], where a ranking model for personal preferences is learned from a subset of test images assessed by users, and is then used for adaptation. However, existing user-adaptive methods still require some amount of user interaction to be able to model specific preferences. In absence of such information we can still discover and employ useful patterns in photo albums, such as relations between similar photos taken in the same scene and characteristics of other photos from the entire album.

The characteristics of the associated photo context can be extracted by considering and comparing photos within coherent clusters [3, 25], which are typically constructed by detecting the natural boundaries in the captured image series. Existing approaches for collection clustering are usually based on temporal information [7, 34] or image similarity [3, 5, 23, 25]. A versatile technique for clustering data where the number of clusters is unknown is hierarchical clustering, which was demonstrated as advantageous in photo collection based applications [10, 22, 23].

Once a clustering is obtained, several approaches may be considered for analyzing the cluster contents and assessing images in their newly defined context. The method proposed by Ceroni et al. [3] utilizes the features collected both on intra- and inter-cluster levels, where a Support Vector Machine based prediction model is learned to predict the selection probabilities for unseen photos. Although their method takes the characteristics of each cluster into account, the learned model is not completely album-adaptive, as it is learned over numerous non-related albums and users. The nature of selection decisions within a given group of similar photos taken in the same scene was studied in the recent work by Chang et al. [4]. In their proposed method, pairwise comparisons are learned with a Siamese CNN, and a relative ranking of images in the group is produced. To handle a similar problem on the album-wise level, the method by Wang et al. [42] complements their Siamese CNN architecture with event type information. Another approach to the context assessment was proposed by Kuzovkin et al. [23], where an independent score provided by an external method is adapted to the multi-level photo context extracted with hierarchical clustering.

Although the problem of automatic photo assessment has attracted a lot of attention in recent years, it has been difficult to

evaluate the relative merits of the proposed methods in real-life scenarios, such as image selection in complete photo albums. In part, this is due to the absence of appropriate ground truth data for image selection within albums, and, from the other side, due to the lack of knowledge on how users perform this task themselves. In this work, we attempt to address these matters through a user study, which allows us to compare different methods and provides useful insights on the user decision patterns.

3 USER PHOTO SELECTION ASSESSMENT

To better understand how users select their preferred photographs within a collection, we perform a user study evaluating user selection decisions on a series of different photo collections. For each album, average *user preference scores* are calculated, indicating how often a particular image was chosen by the study participants, as well as *user confidence scores*, giving an indication of user agreement for each decision.

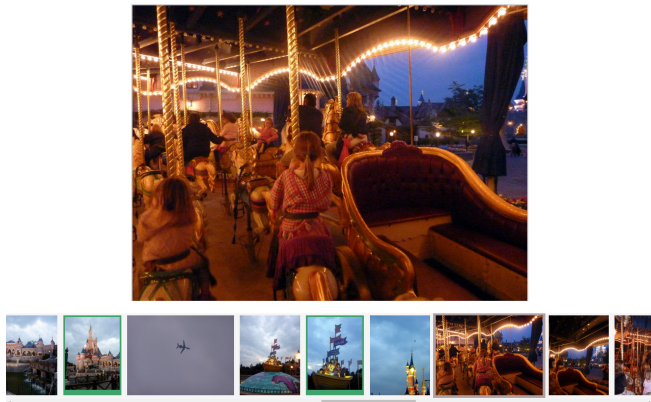


Figure 2: Interface of the conducted selection user study. User can freely browse through the entire collection and perform an image selection on key press.

3.1 User Study Design

Photo Album Data: For the purpose of this user study, we selected six photo albums covering a variety of typical scenarios where amateur photographers may opt to take a large number of photos. Photo albums were selected from several different sources, including PEC dataset [1], YFCC100M dataset [40], CUFED dataset [42] and personal albums of the authors. We have limited our search to collections that were not altered by image processing software, and where no evident pre-selection was applied before, thus possibly containing multiple similar and near-duplicate photos, and reflecting a typical modern photo album taken with a digital camera or a smartphone. As the initial collections vary in the number of photos, and to limit the duration of the experiment, we have extracted 50 photos from each album, in their original consecutive order.

Each user was presented with a pair of albums, where one given album represented a typical family event, such as wedding or birthday, while the second represented a travel photo collection. Two example albums are shown in Figure 1.

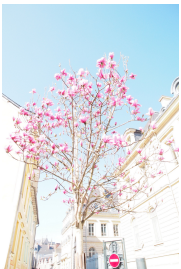
				
Kong et al. [20]	0.55	0.58	0.54	0.53
Jin et al. [17]	3.13	4.59	4.86	5.06
NIMA [38]	4.87	5.01	4.76	4.54

Table 1: Scoring of different photos in the same album by the analyzed image assessment methods. In this case, one pair of similar images is always scored higher than another pair, by each tested method. Thus, a global ranking of photos in an album would not provide an expected selection, if the presence of different scenes is not taken into account.

The analyzed albums demonstrate different characteristics, which allow us to study various real-life scenarios and also identify particular behavior of assessed methods:

- *Family Event 1* is a wedding photo album, with a moderate number of repetitive photos (from 1 to 3 photos for each captured moment on average), where the pictures are taken with a semi-professional camera.
- *Family Event 2* is another wedding photo album captured with a semi-professional camera, but with a higher number of repetitive photos (in some cases, more than 5 photos for each same moment).
- *Family Event 3* album represents a family birthday gathering taken mostly indoors, with a point-and-shoot camera. This collection presents a large number of fuzzy shots, where the points of interest are not well defined, with a moderate-to-high number of repetitive photos (from 3 to 6 photos on average).
- *Travel Album 1* represents a common scenario of vacation photos, where photos of landscapes are mixed together with photos of people posing in front of a landscape. The photos are taken with a semi-professional camera, and multiple highly similar photos of the same moment are taken (more than 5 photos on average).
- *Travel Album 2* consists only of landscape photos, taken with a semi-professional camera, and the number of repetitions is moderate: 2 to 3 photos are taken for each captured moment.
- *Travel Album 3* represents photographs taken during an amusement park visit. In this album, no people are present: it consists of multiple pictures of architecture, landscapes and objects (usually from 2 to 5 photos per same scene). The photos are taken with a point-and-shoot camera and various cases of blurred or under-exposed photos are present.

Participants: In total, 30 participants took part in our study (7F/23M), with ages ranging between 24 and 55 years. Each pair of albums was evaluated by 10 different users. All participants could be characterized as amateur or casual photographers, with varying levels of photographic experience and interest.

Task: For each shown photo album, users were presented with a browser-based interface as shown in Figure 2, and were tasked with putting themselves in the role of the photographer of that collection to select the *best, more representative, or most important photos* in their opinion. No limit was placed on the number of photos selected in each album. Before the start of the experiment, each user was presented with two practice collections of 12 photos each, in order to get familiar with a task and the interface. Then, once the user was ready, they could proceed to selecting photos in the two complete collections assigned to them.

All photos of each album were simultaneously visible as thumbnails, while a larger version of the examined photo was also shown. Users could navigate within the collection freely, with a possibility of viewing all photos before making any selection. Each album had to be completed before moving to the next one. Overall, users took around 20 minutes to complete the entire task.

3.2 Analysis of User Selection Results

User selections were recorded and averaged across the ten users assessing each album, obtaining a normalized *user preference score* for each image. Higher values in this case indicate that an image was selected more often, with a value of one identifying images that were selected by all users. The user preference score for an image i is computed as

$$s_{u_i} = \frac{N_{sel_i}}{N_{users}}, \quad (1)$$

where N_{sel_i} is the number of times image i was selected, and N_{users} is the total number of users.

To better visualize agreement between users, we additionally compute a *user confidence score* which indicates how decisive the preference score for the particular image. The confidence score is calculated using the inverse of the triangular function over the preference score

$$c_{u_i} = \frac{1}{\text{tri}(2s_{u_i} - 1)}, \quad (2)$$

which produces higher confidence when the preference is closer to 1 (every user has selected the image) or to 0 (no user has selected

	Travel album 2	Travel album 3	Family event 2	Family event 3	Family event 1	Travel album 1
Kappa agreement	0.179	0.210	0.334	0.351	0.393	0.472
Kong et al. [20]	0.302	0.120	0.175	0.263	0.384	0.397
Jin et al. [17]	0.258	0.009	0.080	0.075	0.236	0.721
NIMA [38]	0.269	0.196	0.115	0.044	0.128	0.260

Table 2: User selections agreement and performance comparison for analyzed methods. The kappa agreement values are given in the first row (the albums are sorted from the lowest to the highest user agreement). The performance values for each method are computed as the Pearson correlation coefficient between user preferences and scores provided by analyzed methods.

the image). The per-image preference and confidence scores for two example albums used in our study can be seen in Figure 1.

As photo selection is a subjective process, the provided user selections can and do vary between users for each album. To assess the consistency of the selections of different users, the inter-agreement between observers for such data can be computed in different ways. A common choice to estimate the inter-rater agreement is the Kendall’s W coefficient [18] or Cohen’s kappa statistics [6]. The Kendall’s W coefficient is generally used for ordinal ratings, while kappa statistics are applied for nominal ratings, which is our case (*selected/not selected* labels). As in our case ten users rate each album, we employ the modified Fleiss’ kappa measure [11], which is a generalization of the original kappa for more than two raters. Values for this measure can be interpreted as follows according to [24]: $\kappa < 0.2$ indicates slight agreement, $0.2 \leq \kappa < 0.4$ indicates fair agreement, $0.4 \leq \kappa < 0.6$ indicates moderate agreement. The computed Fleiss’ kappa values for each album are given in the first row of Table 2.

Most albums assessed lead to a fair to moderate agreement between users, with the exception of two albums where slight agreement was found (*Travel album 2* and *Travel album 3*). Looking at the content of the albums, several interesting conclusions may be drawn. We observe that albums with higher agreement contain a larger number of people portraits, with repetitive similar photos of the same person or group (including *Travel album 1*, which contains multiple people portraits taken in front of landscapes). At the same time, *Travel album 2* and *Travel album 3* do not contain people’s photos and consist mostly of landscapes and architecture photos. These latter albums demonstrate a larger variance in user selections: the notion of an attractive landscape appears to vary much more than the understanding of a well-captured portrait or group photo.

This suggests that for users it may be easier to perform photo selection of people’s photos within an album, even when the presented people are unknown. In fact, closer observation of users’ selections during the study reveals that facial expressions were a critical factor guiding their decisions when multiple photos of the same people were present. On the other hand, unique photos of people were almost always selected, irrespective of the quality or expression present.

Further, in *Travel album 2* we find a particular example of a photo sequence, where the concept of best photo selection may not be directly applicable: this photo sequence present a panoramic-like capture of the surrounding landscape (seen in the second and third rows of the second album in Figure 1). In this scenario, it

is unlikely that a user would want to keep a single photo, as the series is intended for a particular type of post-processing. Indeed in this case we note that users have shared their selection across the series with no particular photo showing higher selection preference. Another similar scenario could occur when capturing a bracketed series for later construction of a high dynamic range image. Such use cases can frequently occur given the general availability of advanced photo processing tools even on mobile devices.

Another challenging example where user selections become divided is when nearly identical photos are present, with no discernible differences in quality. In such cases, we found that user votes were approximately equally shared between the photos in question, meaning that no single image led to a higher preference, despite users wanting to keep at least one representative image of such scenes.

4 EVALUATION OF IMAGE ASSESSMENT METHODS

As discussed in the previous section, photo selection is a highly subjective task, where, depending on the type of images and collections assessed, even user agreement may be low. As such, automating this task is a daunting challenge. Although a wealth of approaches exist for rating or assessing images, it is unclear how well they perform for selecting images within a photo collection. To evaluate the applicability of different methods for this task, we compare their assigned image scores with our experimental findings.

Given the promising advances in this field with recent deep learning based methods, we opt for evaluating the following, CNN-based methods, which were mentioned in Section 2:

- The approach by Kong et al. [20], which has recently shown the state-of-the-art performance in the independent image aesthetics assessment. In their method, different photographic attributes are estimated and weighted by the image content, giving proper relevance to what should be considered important in an image. It is possible that their proposed content adaptation might be applicable in photo albums as well.
- The method proposed by Jin et al. [17], which introduced another fine-tuning scheme with sample weights that should allow to assess images spanning a wide range of aesthetic quality. In addition, it was demonstrated that their method can be applied for automatic image cropping, therefore we can expect it to cope better with similar repetitive images often present in photo albums.

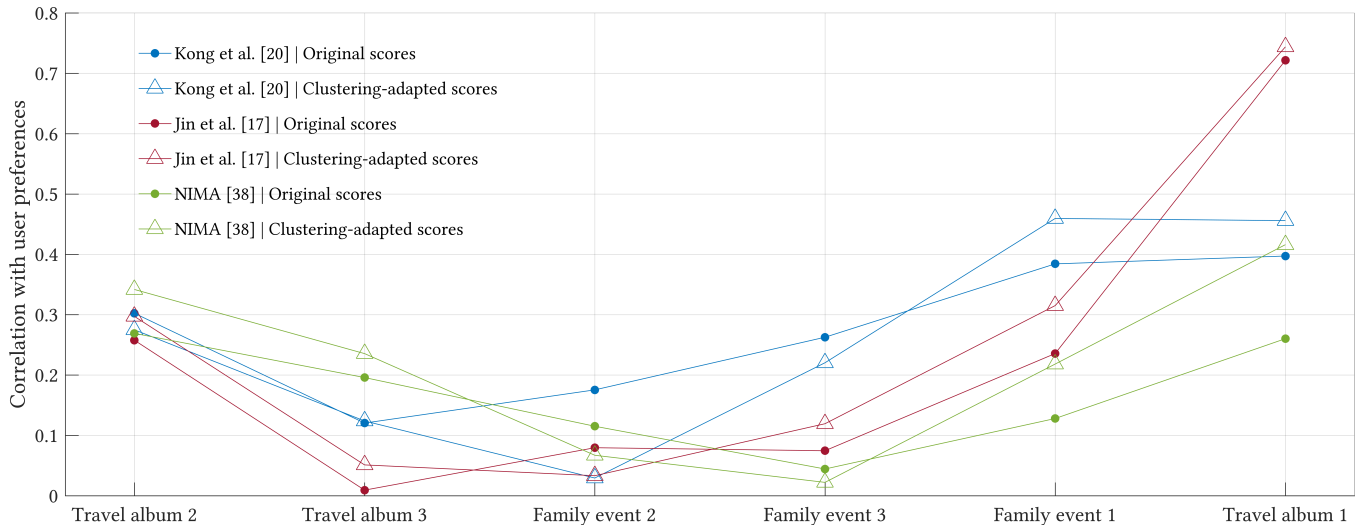


Figure 3: Per-album correlation between computed scores and user preferences, given for the original independent scores and the scores after clustering-based adaptation. The albums are sorted from the lowest to the highest kappa user agreement.

- The NIMA method proposed by Talebi et al. [38], which was designed to estimate both technical image quality and aesthetic attractiveness of an image. In our analysis, we use their NIMA MobileNet [14] version of the CNN architecture.

4.1 Performance of Analyzed Methods

The methods assessed provide a ranking score for each processed image, which we compare against the average user preference scores determined in our experiment. To estimate the performance of each analyzed method, we compute the Pearson correlation coefficient between the method’s photo assessment scores and the user preference scores. With this approach, even if the scores from each method are not computed on the same scale, we can estimate the extent of correlation between the computed independent image scores and the user evaluations. The computed correlation values are detailed in Table 2 and visualized in Figure 3.

Several observations arise from this correlation analysis. The method by Jin et al. [17] often performs poorly in the complete photo albums, despite its potential for dealing with image crops. Nevertheless, for *Travel album 1* it shows the best correlation with the user preferences, among all the methods. This album consists of a number of very similar photos of landscapes and landscapes with people, which is possibly the scenario where the approach by Jin et al. [17] demonstrates its best performance. In addition, their approach appears to respond to the presence of blur in images, and our observations show that in this album the users often selected the most sharp photos among similar ones.

The NIMA method [38] performs worse in the albums where people’s photos are present, but it shows its best performance in landscape-focused *Travel album 2* and *Travel album 3*. However, the degree of inter-observer agreement is relatively low for these collections, therefore no certain conclusions can be made.

On average, the highest performance is demonstrated by the approach of Kong et al. [20]. Despite their primary aim of addressing

general aesthetics scoring of photos, their produced score demonstrates a noticeable correlation with the user preferences. Possibly, this is due to their proposed content-dependent weighting scheme, which provides better estimation of different type of the image content. Thus, their approach could be potentially suitable for aiding the selection process as well.

5 CLUSTERING-BASED SCORE ADAPTATION

As it was shown in the previous section, the applicability of independent image assessment methods is limited in complete photo collections. Although in some cases this could be explained by the method’s occasional failure on certain images with complex content, another important reason is that the photo context is not considered by such methods. The illustration of this point is given in Table 1: it can be observed that photos from different scenes in the same album often receive non-comparable scores, which can lead to inaccurate ranking of photos in the album, if it is performed in a global manner.

The direct approach of using independent image scores in a photo album would be to rescale and normalize the scores linearly, in accordance with other scores computed in the album. However, we have observed earlier that the original scores given by methods are often poorly correlated with user preferences. One reason for this could be the elimination of the context present in photo albums, as each photo is assessed independently, without consideration of its surrounding photos. For example, even if multiple pictures of the same scene were taken, suggesting that the user found that scene important, each picture from the scene could receive a low score and be potentially rejected. For this reason, we attempt to utilize the notion of context in the photo assessment within an album.

As discussed in Section 2, a possible approach to define the photo context is through album clustering, where the entire collection is clustered into groups of similar photos. For this purpose, we apply the framework proposed in the approach of Kuzovkin et al.

[23], where the entire photo collection is clustered into clusters of different similarity levels to define the relevant context for each photo.

The context of a photo is modeled using three enclosed hierarchical levels: collection level (containing the whole collection), scene cluster level (containing photos depicting the same scene), and near-duplicate cluster level (reflecting very similar images). This context hierarchy is obtained by combining time-based clustering together with similarity clustering based on the SIFT descriptors.

To evaluate how the presence of context affects the performance of the state-of-the-art methods tested in the previous section, each album from our user study was clustered and its scores were adapted, according to [23]. To adapt an image-based score to its associated context, z-scores [21] are computed for each image and for each clustering level, considering both the image score and statistics of the cluster, as follows:

$$z_{I_L} = \frac{SI - \mu_L}{\sigma_L}, \quad (3)$$

where μ_L and σ_L denote mean and standard deviation of the scores, computed on one of three levels $L \in C, SC, ND$, which define collection level, scene cluster level and near-duplicate cluster level, respectively. After z-scores for each level are computed, they are combined into the global score Z_I , as an average of the adapted scores from three levels.

In Figure 3, we also demonstrate the result of the z-score context adaptation for the analyzed methods. It can be seen that for the methods by Jin et al. [17] and NIMA method [38] the correlation with user preferences increases for the most of the albums, after the performed adaptation. Especially noticeable increase in performance can be observed for NIMA method [38], which is also reflected in the average correlation increase, shown in Table 3. For the method of Kong et al. [20], the performance gain is not particularly evident. In this case, the performance has increased for the albums where similar repetitive sequences of photos are largely present, such as *Family event 1* and *Travel album 1* (also the kappa user agreement is the highest for these albums). However, for other albums the effect of adaptation is opposite, with a decrease in the correlation. This could be also due to the nature of these albums, as some of them contain a lower number of repetitive photos.

	Original scores	Clustering-adapted scores [23]
Kong et al. [20]	0.274	0.261
Jin et al. [17]	0.230	0.260
NIMA [38]	0.169	0.217

Table 3: Average correlation across all albums for original scores and scores after clustering-based adaptation.

It is also important to note that the utilized clustering is less robust in *Travel album 2* and *Travel album 3*, due to occasional large viewpoint changes for the same captured scenes and presence of severe blur in some photos. Additionally, the album *Family Event 2* represents a special case for all methods, where the correlation before and after adaptation is rather low, even in presence of

multiple similar photos. We found that the number of low-quality photos (such as blurred ones) is smaller in this album, while, as we previously observed, the user preferences in this album were often guided by more complex factors, such as face expressions or people’s poses. Due to this, the original scores may not always be accurate, which can in turn lead to an unreliable adaptation. Moreover, the employed adaptation approach sometimes may be too simplistic to model the subtleties of user employed criteria.

6 CONCLUSION

Our work studies the performance of image assessment methods when applied to the task of photo selection in photo albums. Despite the wealth of work available for assessing the quality or aesthetics of images, most existing methods consider images independently, without knowledge of their surrounding collection or context. To understand how users perform this task, we have collected a selection of photo albums covering different events and quality levels, and conducted a user study on them where users were asked to select which photos they would like to keep in each collection. Our findings suggest that users consider several elements in their decisions, varying from the quality of images, to the depicted scene or people. More interestingly, we find that users often show very different selection decisions between them, highlighting the difficulty of this task for an automated method. The low agreement for some albums also confirms that in certain cases a personalized user preferences modeling would be necessary.

Our comparisons of image evaluation scores from several state-of-the-art methods relative to the results of our user study show that, in most cases, independent image assessment solutions correlate to a limited degree with user selections. At the same time, most methods performed better for albums where user agreement was higher. To assess whether additional knowledge of the context of images could improve the results of automatic image evaluation approaches, we adapted independent image scores using an album clustering approach [23]. Although this adaptation showed some benefit in higher user-agreement albums, no clear conclusions could be drawn in lower agreement cases.

Given the large quantities of photographs captured for any key event in our lives, efficient solutions that can aid users in the cumbersome task of photo selection are likely to become increasingly necessary. In this work, we show that existing methods can work for some scenarios, but are globally far from being able to predict user decisions in the context of their own albums and photo collections. Nevertheless, our study provides some insights in user behavior that we hope may serve as a basis for future work, such as further computational analysis which image features affect user selection decisions within photo albums.

REFERENCES

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van. 2013. Event Recognition in Photo Collections with a Stopwatch HMM. In *2013 IEEE International Conference on Computer Vision*. 1193–1200.
- [2] Andrea Ceroni, Vassilis Solachidis, Mingxin Fu, Nattiya Kanhabua, Olga Papadopoulou, Claudia Niederee, and Vasileios Mezaris. 2015. Investigating human behaviors in selecting personal photos to preserve memories. *2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (2015).
- [3] Andrea Ceroni, Vassilios Solachidis, Claudia Niederee, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. 2015. To Keep or Not to Keep: An Expectation-oriented Photo Selection Method for Personal Photo Collections. In

- Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR '15)*, 187–194.
- [4] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. 2016. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 148.
 - [5] Wei-Ta Chu and Chia-Hung Lin. 2008. Automatic Selection of Representative Photo and Smart Thumbnailing Using Near-duplicate Detection. In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08)*. 829–832.
 - [6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
 - [7] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. 2005. Temporal Event Clustering for Digital Photo Collections. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 3 (Aug. 2005), 269–288.
 - [8] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. *Computer Vision – ECCV 2006* (2006), 288–301.
 - [9] Claire-Hélène Demarty, Mats Sjöberg, Mihai Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do, and Hanli Wang. 2017. Predicting Interestingness of Visual Content. In *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 233–265.
 - [10] Boris Epshtein, Eyal Ofek, Yonatan Wexler, and Pusheng Zhang. 2007. Hierarchical Photo Organization Using Geo-relevance. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*. ACM, 18.
 - [11] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.
 - [12] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, and Shipeng Li. 2011. The role of attractiveness in web image search. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 63–72.
 - [13] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The Interestingness of Images. In *2013 IEEE International Conference on Computer Vision*. 1633–1640.
 - [14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).
 - [15] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. 2016. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications* 75, 2 (2016), 1301–1331.
 - [16] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), 145–152.
 - [17] Bin Jin, Maria V Ortiz Segovia, and Sabine Süsstrunk. 2016. Image aesthetic predictors based on weighted CNNs. In *2016 IEEE International Conference on Image Processing (ICIP)*. 2291–2295.
 - [18] Maurice G Kendall and B Babington Smith. 1939. The problem of m rankings. *The annals of mathematical statistics* 10, 3 (1939), 275–287.
 - [19] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. *2015 IEEE International Conference on Computer Vision (ICCV)* (Dec. 2015), 2390–2398.
 - [20] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Computer Vision – ECCV 2016*. Springer, 662–679.
 - [21] Erwin Kreyszig. 2007. *Advanced engineering mathematics*. Wiley publishing.
 - [22] Santhana Krishnamachari and Mohamed Abdel-Mottaleb. 1999. Image browsing using hierarchical clustering. In *Proceedings IEEE International Symposium on Computers and Communications*. 301–307.
 - [23] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch. 2017. Context-aware Clustering and Assessment of Photo Collections. In *Proceedings of the Symposium on Computational Aesthetics (CAE '17)*. ACM, 6:1–6:10.
 - [24] J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.
 - [25] Alexander C Loui and Andreas Savakis. 2003. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Transactions on Multimedia* 5, 3 (Sept. 2003), 390–402.
 - [26] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. 2014. RAPID: Rating Pictorial Aesthetics Using Deep Learning. *Proceedings of the ACM International Conference on Multimedia - MM '14* (2014), 457–466.
 - [27] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 990–998.
 - [28] Yiwen Luo and Xiaoou Tang. 2008. Photo and Video Quality Evaluation: Focusing on the Subject. In *Computer Vision – ECCV 2008*. Springer, 386–399.
 - [29] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-Preserving Deep Photo Aesthetics Assessment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 497–506.
 - [30] Luca Marchesotti and Florent Perronnin. 2013. Learning beautiful (and ugly) attributes. *British Machine Vision Conference (BMVC)* 7 (2013), 1–11.
 - [31] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Surcka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. *2011 International Conference on Computer Vision (ICCV)* (Nov. 2011), 1784–1791.
 - [32] Eftichia Mavridaki and Vasileios Mezaris. 2015. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *2015 IEEE International Conference on Image Processing (ICIP)*. 887–891.
 - [33] Kayoung Park, Seunghoon Hong, Mooyeol Baek, and Bohyung Han. 2017. Personalized Image Aesthetic Quality Assessment by Joint Regression and Ranking. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (March 2017), 1206–1214.
 - [34] John C. Platt, Mary Czerwinski, and Brent A. Field. 2003. PhotoTOC: automatic clustering for browsing personal photographs. *Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia 1* (Dec. 2003), 6–10 Vol.1.
 - [35] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. 2015. The beauty of capturing faces: Rating the quality of digital portraits. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–8.
 - [36] Philipp Sandhaus, Mohammad Rabbath, and Susanne Boll. 2011. Employing aesthetic principles for automatic photo book layout. In *International Conference on Multimedia Modeling*. Springer, 84–95.
 - [37] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. ACM, 659–668.
 - [38] Hossein Talebi and Peyman Milanfar. 2017. NIMA: Neural Image Assessment. *arXiv preprint arXiv:1709.05424* (2017).
 - [39] Xiaoou Tang, Wei Luo, and Xiaogang Wang. 2013. Content-Based Photo Quality Assessment. *IEEE Transactions on Multimedia* 15, 8 (Dec. 2013), 1930–1943.
 - [40] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Commun. ACM* 59, 2 (Jan. 2016), 64–73.
 - [41] Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. 2014. Smart Photo Selection: Interpret Gaze As Personal Interest. (2014), 2065–2074.
 - [42] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. 2016. Event-specific image importance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4810–4819.
 - [43] Maria K. Wolters, Elaine Niven, and Robert H. Logie. 2014. The Art of Deleting Snapshots. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, 2521–2526.
 - [44] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. 2010. Personalized Photograph Ranking and Selection System. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. 211–220.