Context in Photo Albums: Understanding and Modeling User Behavior in Clustering and Selection

DMITRY KUZOVKIN, Technicolor, IRISA, Univ Rennes
TANIA POULI, Technicolor
OLIVIER LE MEUR, Univ Rennes, CNRS, IRISA
RÉMI COZOT, Univ Rennes, CNRS, IRISA
JONATHAN KERVEC, Technicolor
KADI BOUATOUCH, Univ Rennes, CNRS, IRISA

Recent progress in digital photography and storage availability has drastically changed our approach to photo creation. While in the era of film cameras, careful forethought would usually precede the capture of a photo, nowadays a large number of pictures can be taken with little effort. One of the consequences is the creation of numerous photos depicting the same moment in slightly different ways, which makes the process of organizing photos laborious for the photographer. Nevertheless, photo collection organization is important both for exploring photo albums and for simplifying the ultimate task of selecting the best photos. In this work, we conduct a user study to explore how users tend to organize or cluster similar photos in albums, to what extent different users agree in their clustering decisions, and to investigate how the clustering-defined photo context affects the subsequent photo selection process. We also propose an automatic hierarchical clustering solution for modeling user clustering decisions. To demonstrate the usefulness of our approach, we apply it to the task of automatic photo evaluation within photo albums and propose a clustering-based context adaptation.

Additional Key Words and Phrases: Photo albums clustering, Photo collection organization, Image assessment, Photo selection

1 INTRODUCTION

The democratization of digital photography has introduced significant changes to our approach to the photography process. No longer limited by the constraints imposed by the film medium, users tend to accumulate large collections of photos, where often multiple repetitions of the same scene are captured, deferring the choice of the best photos to later. Along with the change towards the 'afterthought' model of dealing with photos, new challenges have appeared. By accumulating multiple large photo albums, users face difficulties not only in the process of selecting the best photo, but also in more basic tasks, such as browsing, sharing or showing their photos.

Without visual clues for navigating through an album, seemingly simple collection processing tasks can become challenging [18, 34, 41]. To overcome this, a photo album can be organized or clustered into particular events, time series, or groups of photos of high similarity. This problem has attracted significant research, where different information is employed to provide an automatic clustering of a photo collection [4, 6, 7, 10, 11, 21, 25, 38]. Yet, the nature of the users' decisions and motivations during a manual clustering of photo collections, which could provide valuable insights for automatic approach, has remained little studied.

With that in mind, we design and conduct a user study on the clustering of photo albums. We investigate how users group similar photos together and evaluate the level of agreement in users' grouping decisions for different types of content. We also search for common traits that could be helpful in modeling the users' behavior. Based on our findings, we propose an automatic clustering method, which is based on the visual similarity distance between photos and utilizes an adaptive cut approach to automatically define photo moment boundaries at

Authors' addresses: Dmitry Kuzovkin, Technicolor, IRISA, Univ Rennes, Rennes, France, kuzovkin.dmitry@gmail.com; Tania Pouli, Technicolor, Rennes, France, tania.pouli@technicolor.com; Olivier Le Meur, Univ Rennes, CNRS, IRISA, Rennes, France, olivier.le_meur@irisa.fr; Rémi Cozot, Univ Rennes, CNRS, IRISA, Rennes, France, remi.cozot@irisa.fr; Jonathan Kervec, Technicolor, Rennes, France, jonathan.kervec@technicolor.com; Kadi Bouatouch, Univ Rennes, CNRS, IRISA, Rennes, France, kadi.bouatouch@irisa.fr.

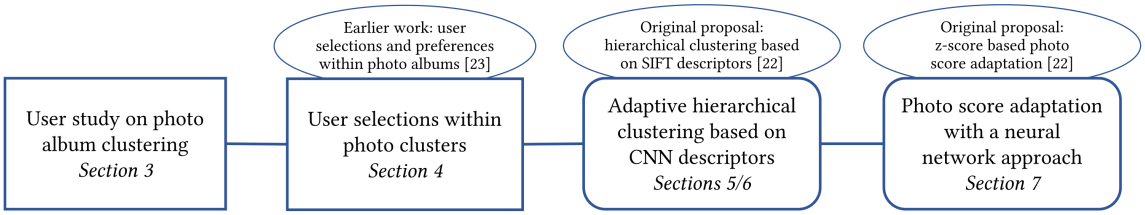


Fig. 1. Outline of the principle contributions of the current article and connections with our earlier works.

different granularity levels. To derive statistics about the photo selection decisions of users within groups of similar photos inside albums, we rely on the dataset provided by our earlier work on user selections within photo albums [23]. We augment this data with the clustering user decisions obtained in the current work.

To demonstrate the usefulness of our clustering approach, we apply it to the task of image assessment in the context of a photo album, where an independent image quality score is adapted using the context provided by our clustering. Although the task of individual image assessment was addressed in numerous recent methods [5, 8, 27, 30, 33, 40, 44], the majority of existing techniques represent average user preferences modeled from a large variety of content, without considering album-specific information. In our earlier work, we approached this problem with an adaptive solution that leverages the independent photo scores with inter-album connections between similar photos, to better model user selection decisions [22]. Inspired by the same idea, we propose a more robust machine learning based solution that utilizes independent scores and corresponding clusters' statistics together.

The general outline of this work and connections with our earlier works are given in Figure 1. The paper is organized as follows. In Section 2 we provide an overview of the existing methods and studies that cover the field of photo albums processing. In Section 3 we describe the experimental procedure of our user study on the photo albums clustering and our findings on the users' decisions. In Section 4 we investigate the nature of photo selection decisions within the photo clusters obtained from users. In Section 5 we introduce our approach for automatic photo album clustering. In Section 6 we demonstrate and analyze the performance results of the proposed clustering method. Finally, in Section 7 we demonstrate a possible application of the automatic clustering, where an independent image score is adapted to the corresponding album and cluster context.

2 RELATED WORK

When dealing with captured images, both amateur and professional photographers usually perform similar tasks: they browse and organize their photo collections, select the best candidates and proceed with editing, sharing or other usage of the selected photos. To organize a photo collection, users may manually cluster photos belonging to the same event, location and so on, or automatic clustering methods may be employed. Similarly, to rank or select photos, users may proceed manually or might rely on automatic quality or aesthetic assessment methods. In the following, we discuss key approaches in these two categories. Although these tasks are often performed in conjunction by users, they are rarely addressed together by automatic methods, despite the fact that photo album clustering could potentially provide useful context information in the subsequent process of photo assessment.

2.1 Photo Album Organization and Clustering

The context of a photo can be defined through its connections with other photos depicting the same moment or location captured by the photographer. To define the natural boundaries of each moment and organize photos into clusters that share the same context, different information may be employed. Temporal information, when

available, can be used to naturally define such boundaries [7, 38], or geo-location information data can be employed [10]. Although these approaches allow for a coarse event-based clustering of the photo collection, they are not sufficient to group visually similar images. For this purpose, similarity information extracted from the images can be employed [4, 6, 9, 11, 25]. More general approaches of event recognition were also proposed [2, 13, 25], which split photo albums into separate events on a near-semantic level, employing multiple feature types (e.g. time stamps, GPS data and various photo features). The use of different image features also found its application in the related domain of image retrieval: different image descriptors, their aggregated codings, as well as image representations based on convolutional neural networks (CNNs) are largely employed in the retrieval task [37, 39].

An alternative approach to album organization is using hierarchical clustering [48]. When applied to a photo collection, this approach assigns photos to the branches of a hierarchical tree, where the height of branches represents the similarity between photos, according to some criterion. Hierarchical clustering provides a convenient way to organize data in image collections, as it does not require information about the potential number of clusters. The obtained tree representations of the albums can be used for image navigation and browsing [10, 21]. However, these approaches are aimed to create a non-linear browsing structure, which can be inconvenient for some tasks, notably photo selection.

To simplify the photo browsing experience, redesigns of the typical photo browsing user interface have been proposed [12, 42]. These approaches often provide an organization of the photo collection based on visual image similarity or on topic- or event-based correspondence. For example, color palette similarity can be used to pack images into rectangular blocks of unified appearance [12]. Although such an interface representation can be useful for visualizing a photo collection, it also eliminates a linear browsing experience.

In our earlier work [22] we proposed a linear, multi-level organization of photo albums using a hierarchical clustering approach based on the visual similarity between images. In the current work we explore the clustering aspect in more depth. First, we design a user study to evaluate the extent of users' agreement in their album clustering decisions. By using this information, we obtain insights about the users' behavior and also evaluate the performance of automatic clustering solutions. Motivated by our observations, we improve an earlier proposed clustering approach with an adaptive cut technique, which is further strengthened by the use of deep features in the similarity distance computation.

2.2 Photo Aesthetics and Quality Assessment

Selection of the best photos within an album is a common post-capture task. However, the selections made by users depend on multiple factors, ranging from objective characteristics of image quality, such as sharpness, exposure, noise or other artifacts, to more subjective aspects of photo aesthetics, such as scene composition, color style and interestingness of a depicted object; these factors can be further affected by photo semantics, such as for example, the presence of known people in a photo and their facial expressions [3, 31, 46]. The task of evaluating these characteristics for photo selection can be aided by automatic image assessment methods.

The area of automatic photo assessment has received considerable attention in recent years, aiming to model human preferences by learning from user data. Hand-crafted features, either describing aesthetic properties inspired by photography practices or objective quality criteria, form the basis of many automatic photo assessment techniques [8, 30, 33, 40, 44, 49]. Generic image descriptors employed for image classification were also found to be applicable in the task of aesthetics classification [32]. Several studies have also identified image features and characteristics that have the largest influence in the users' decisions [3, 31, 46]. Most recently, deep learning techniques have been put to the task of photo assessment [5, 16, 17, 19, 27, 28, 43]. The level of abstraction achieved by CNNs allows to model image properties that cannot be readily obtained with hand-crafted features.

In addition to the features used for assessing the aesthetic quality of photos, it is also important to note the influence of the dataset used, in particular for learning-based methods. The majority of existing aesthetics prediction models are trained and tested on one of a few known datasets, notably the *Photo.Net* dataset [8] and the *AVA* dataset [35], both based on scores acquired from peer reviews in popular photographers' social networks. Although these datasets provide a large variety of photos, two inherent limitations should be considered. First, the scores in these datasets are biased towards high-quality, often post-processed photos. Second, the notion of context is not preserved, as each photo is viewed and assessed independently, outside the original collection.

2.3 Context in Photo Assessment

A photo typically forms part of a larger collection or photo album, representing a particular event. When we view our own photos and decide which ones to keep, we tend to consider them in the context of that event. The methods discussed in the previous section assess images independent of their context. Although some aspects of context would be difficult, if not impossible, to model, such as the personal feelings associated with the depicted place, person or event, the characteristics of the album and relations between similar photos could significantly influence the perceived quality of a photo.

Some recent works have approached the use of individual preferences in image assessment by weighting the results of a general assessment model by user's adjustments [50] or by learning a ranking model from a subset of user evaluated photos [36]. In absence of personalized data, useful photo context characteristics can be determined by considering each photo within the cluster of related similar photos from the same scene. While some methods collect features on intra- and inter-cluster levels to learn a prediction model for unseen photos [4], other methods focus on pairwise comparison predictions within a group of similar photos [5].

In our previous work [22] we proposed a context-aware approach for photo assessment, where an independent image quality score was adapted to the multi-level album context extracted with hierarchical clustering. While this approach was found effective when using objective quality metrics (such as sharpness), a later study [23] showed certain limitations of the approach when modeling more complex evaluations with no predefined assessment criteria. In the current work, we complement the photo selection decisions collected in our earlier work with the clustering decisions for the same photo albums, and use this information to derive statistics about users' selection behavior. In addition, we propose a new solution to the context-adaptive assessment, which employs the clustering information in a machine learning approach.

3 USER STUDY ON PHOTO ALBUM CLUSTERING

To get a better understanding of how users organize and cluster photos within a collection, we conducted a user study on a set of different photo albums. The users grouped photos within the albums into clusters of different similarity level, by means of an interface specially designed for the study (Figure 2).

3.1 Photo Albums

This user study is based on the photo albums data collected in our earlier study on photo selection within albums [23]. It includes photo collections from different sources: PEC dataset [1], YFCC100M dataset [45], CUFED dataset [47] and personal albums [23]. The selected six albums represent typical scenarios of photos taken by amateur photographers, where multiple similar and near-duplicate photos are present, and the number of repetitive photos varies for each collection. The albums represent a large range of image quality, and different degrees of similarity. From each source album, 50 consecutive photos were selected, maintaining the structure of the original collection but limiting the duration of the study. A more detailed description of each album can be found in [23].

Each user was presented with a pair of albums related to different photo scenarios: one album represented a typical family event, such as wedding or birthday, and the second album represented a travel photo collection.

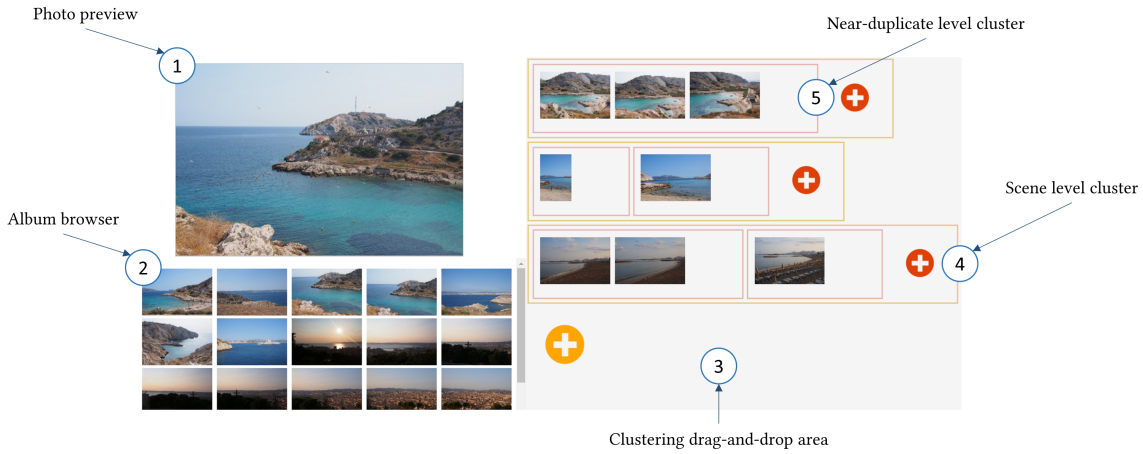


Fig. 2. Interface of the clustering user study. Users can browse the entire album on the left side and by using drag-and-drop gestures, they can move images to the right side and assign to the clusters of their choice. Using the dedicated buttons, new clusters of different depth can be created, to achieve the desired organization.

In total, 30 participants took part in our study (5F/25M), with ages ranging between 23 and 57 years. Each pair of albums was processed by 10 different users. All participants could be characterized as amateur or casual photographers, with varying levels of photographic experience and interest.

3.2 User Study Design

Each user was asked to put himself in the role of the photographer of each photo collection, whose ultimate task would be to create a curated photo album in which the best and the most representative photos are selected. With this in mind, the user was presented with the task of grouping similar photos from a photo album together, according to two levels of similarity: scene clusters and near-duplicate clusters. The following descriptions were given for each level:

- *Scene clusters.* "Photos depicting the same scene (for example, identified by the same background), but with significant changes present. Examples of possible changes: viewpoint changes, new persons or objects appearing throughout the photos. A scene cluster is formed by multiple smaller groups of high similarity (near-duplicate clusters)."
- *Near-duplicate clusters.* "Photos depicting the same objects or persons in the same scene. Although some variations between photos are possible, usually they do not largely change the scene. Examples of possible changes are: pose changes, small viewpoint changes, quality changes (e.g. blur or exposure changes)."

Although it is possible that users might exhibit a different behavior when assessing their own photos, we have opted for this experimental design in the interest of collecting statistical tendencies of users' behavior.

The motivation behind the two level clustering structure is the following. Photos from the same scene (usually depicting the same place or moment) are often seen together when browsing an album. At the same time, groups of near-duplicate photos provide an immediate photo context when selecting which photos to keep, as users are likely to compare photos with one another to identify the best one. Furthermore, in the selection task, the scene context can be used for a more refined decision after selecting among near-duplicates, to keep only few photos from the entire scene.

The essential challenge in the design of such a user study was to provide the ability of multi-level clustering for users, where they could assign an image both to the scene and near-duplicate cluster. Since the near-duplicate clusters are logically enclosed into the higher level scene clusters, the user study interface was created in a similar two-level manner, based on a drag-and-drop principle. The interface is shown in Figure 2: users can freely browse the entire photo collection and cluster images by dragging and dropping to the right part of the screen. A photo can be either assigned to an existing cluster, or a new cluster can be created.

Before the start of the experiment, each user was presented with two practice collections of 12 photos each, in order to get familiar with the task and the interface. Then, they could proceed to cluster photos in the two complete collections assigned to them. No expected number of clusters was defined for users, neither the approximate number of images within a cluster was suggested. Each album had to be clustered before proceeding to the next one, but no time constraint was defined. On average, users took around 30 minutes to complete the task for the assigned pair of albums.

Overall, users perceived the given task positively, while the following and similar comments were given by a few observers: *"I have to do such a grouping myself quite often when going through many photos after a travel"*, *"I might not explicitly put the photos into some folders, but in my mind I select the photos within the clusters of similar photos like this"*.

3.3 Clustering User Agreement

After all images in the album were processed by a user, each image was assigned clustering labels indicating its scene and near-duplicate clusters, allowing us to analyze the inter-user agreement on their clustering decisions. The analysis is performed using the Adjusted Rand Index (*ARI*) [15], which provides a measure of similarity between two data clusterings. The Rand Index (*RI*) considers all possible pairs of data elements and counts the number of pairs that are assigned to the same or different clusters between two given partitions. The *ARI* is the corrected-for-chance version of the original RI, which is adjusted by the expected value of *RI*. The *ARI* is close to 0 for random labeling and equal to 1 when the clusterings are identical.

The per-album user agreement is given in Figure 3, and more detailed results can be found in Table 1. The *ARI* is calculated for the clustering provided by each possible pair of users, and the values in the table represent the obtained mean and standard deviation values across all users.

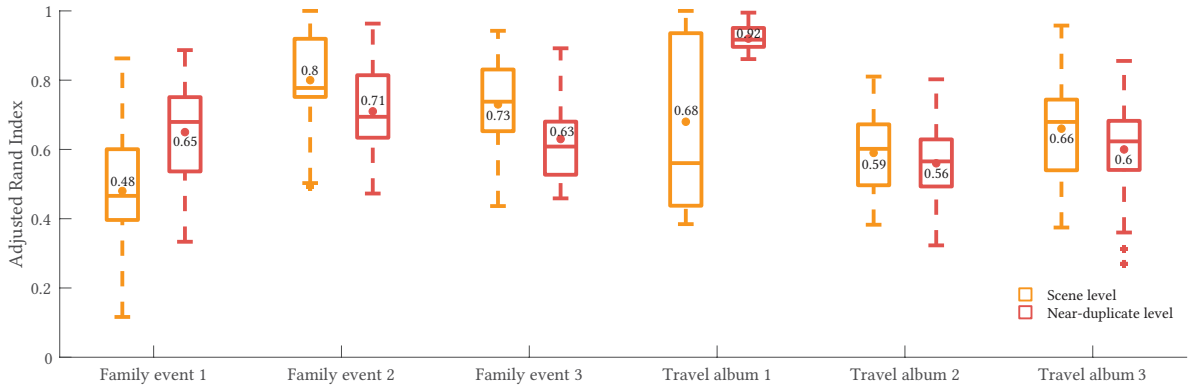


Fig. 3. Per-album user agreement: an *ARI* was computed between each pair of users. The box plots represent the per-album distributions of *ARI*, where a dot and the associated value indicate the average per-album *ARI*.

Table 1. Per-album user agreement and performance of the analyzed clustering methods. *SC* denotes scene level clustering results, *ND* denotes near-duplicate level clustering results. The first value represents the mean ARI across all users. The second value (given in parentheses) represents the standard deviation of the ARI. The user agreement results are discussed in Section 3.3, and the performance results of automatic clustering methods are discussed in Section 6.

	User Agreement		Time-SIFT [22]		Adaptive Time-SIFT		Adaptive Time-CNNR	
	Scene level (SC)	Near-duplicate level (ND)	SC	ND	SC	ND	SC	ND
Family event 1	0.481 (±0.19)	0.645 (±0.15)	0.338 (±0.18)	0.269 (±0.09)	0.546 (±0.20)	0.716 (±0.12)	0.470 (±0.15)	0.567 (±0.08)
Family event 2	0.802 (±0.14)	0.715 (±0.12)	0.880 (±0.12)	0.735 (±0.14)	0.880 (±0.12)	0.605 (±0.10)	0.710 (±0.07)	0.624 (±0.11)
Family event 3	0.732 (±0.13)	0.625 (±0.11)	0.366 (±0.09)	0.143 (±0.06)	0.600 (±0.15)	0.537 (±0.07)	0.495 (±0.12)	0.464 (±0.05)
Travel album 1	0.682 (±0.24)	0.922 (±0.03)	0.795 (±0.25)	0.916 (±0.04)	0.795 (±0.25)	0.889 (±0.01)	0.742 (±0.18)	0.930 (±0.02)
Travel album 2	0.592 (±0.11)	0.561 (±0.11)	0.495 (±0.05)	0.241 (±0.11)	0.481 (±0.14)	0.617 (±0.08)	0.569 (±0.09)	0.545 (±0.11)
Travel album 3	0.657 (±0.16)	0.598 (±0.14)	0.325 (±0.06)	0.162 (±0.07)	0.006 (±0.00)	0.414 (±0.05)	0.492 (±0.07)	0.616 (±0.11)
Average	0.658 (±0.16)	0.678 (±0.11)	0.533 (±0.13)	0.411 (±0.09)	0.551 (±0.14)	0.629 (±0.07)	0.580 (±0.11)	0.624 (±0.08)

Our findings suggest that users do not always achieve a high agreement, however a certain level of agreement is present for all albums. The obtained results demonstrate that the difficulty of album clustering depends on its content and the presence of particular features. For example, the lowest scene clustering agreement is found for the wedding album *Family event 1*, where the average *ARI* is equal 0.481 (also, the standard deviation is rather high in this case and equal 0.19). This can possibly be explained by the sparsity of particular scene landmarks, as the entire album is related to the wedding ceremony in the church, and also due to the presence of multiple close-up shots, which do not provide many indications about the general setting. At the same time, the near-duplicate clustering agreement for the same album is higher, as the *ARI* is equal 0.645, which can be explained by the presence of multiple repetitive shots that are easy to identify.

Almost perfect user agreement can be found for near-duplicate clustering (average $ARI = 0.92$) of the *Travel album 1*, which contains numerous almost identical shots, picturing people in front of landscapes. Also the scene clustering agreement is relatively high for this album (average $ARI = 0.68$). According to the acquired results and users' remarks after the experiment, albums with people present in photos are easier to cluster, as the boundaries between the captured moments are clearer. On the contrary, albums consisting of landscape or object photos make the clustering task more difficult. For instance, in *Travel album 2* we can find a photo sequence that presents a panoramic capture of the surrounding landscape. While the scene level concept is applicable here, the division into near-duplicate clusters does not achieve a considerable agreement by users, with individual user clusterings varying significantly.

4 USER SELECTIONS WITHIN CLUSTERS OF PHOTOS

The acquired user clustering decisions can provide additional insights on another subject—the nature of user selections within photo albums. In our previous study [23], we analyzed how users select the best or most important photos in a collection. Since the present clustering study was performed on the same albums, we combine the acquired information and analyze the possible influence of photo clusters in the selection process.

The data aggregation scheme is given in Figure 4. As the current clustering study and the earlier selection study were performed by different users, the aggregation of data cannot be performed in a direct manner. To overcome this, we create one generalized per-album clustering and combine it with multiple user selections inside each album.

To create one aggregated clustering per album from the multiple user partitions obtained in the user study, we apply the ensemble clustering approach by Lu et al. [29], which finds one combined clustering through an optimization scheme. An example of the acquired ensemble clustering is given in Figure 4.

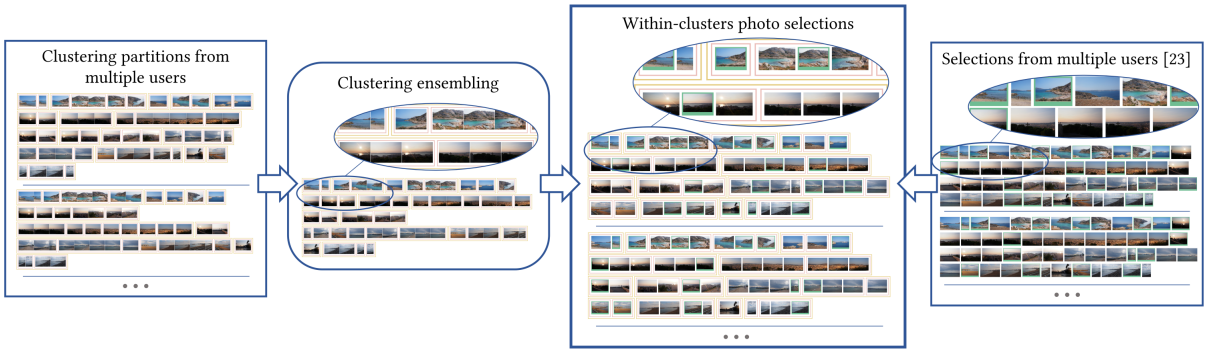


Fig. 4. Data aggregation for the combined study on clustering and selection. Multiple clustering partitions are transformed into one ensemble clustering using [29]. Photo selections acquired in our earlier study [23] are incorporated into the computed ensemble clustering, to achieve within-clusters photo selections.

Once a per-album ensemble clustering is computed, we incorporate the user selection decisions into it. The photo selections acquired in our earlier study [23] represent the decisions provided by multiple users for each album. These decisions are directly combined with the computed ensemble clustering, to obtain within-clusters photo selections for different users. Examples of the output clustering structure along with user selections from one user can be seen in Figure 4. By analyzing the photo selections within such context, we can assess the influence of groups of similar photos in users' decision making.

The results of the first analysis are given in Table 2. In this table, we compute ratios of selection for different entities within the albums:

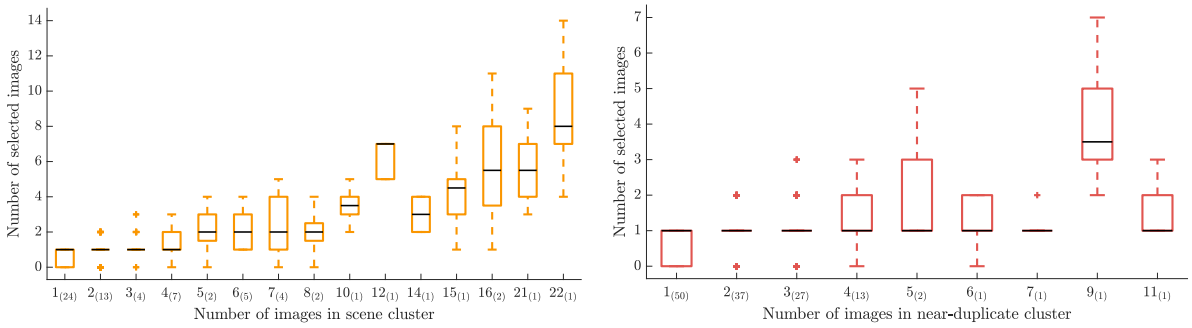
(1) ratio of selected photos indicates a ratio between the number of selected photos within an album and the total number of photos in the album;
(2) ratio of selected scene clusters indicates a ratio between the number of the scene clusters where at least one photo was selected and the total number of scenes;
(3) ratio of selected near-duplicate clusters indicates a ratio between the number of the near-duplicate clusters where at least one photo was selected and the total number of near-duplicate clusters;
(4) ratios of selected singleton clusters indicates a ratio between the number of clusters consisting of only one image that were selected and the total number of such clusters.

Several conclusions can be drawn from these results. First, on average, users have selected around 37% of the photos. A key observation is that in the album *Family event 1* the ratio of selected photos is higher than in others and equal to 0.57, which can be explained by the fact that this album has less repetitive content and contains a number of unique portraits of people not reappearing in other photos. Second, we can observe that the average ratio of selected scene clusters differs from the average ratio of selected near-duplicate clusters (0.818 versus 0.696). This appears natural, as scenes generally contain a wider range of photos, which leads to a higher chance that at least one image will be selected within them. Finally, the ratio of selected singleton clusters (consisting of only one image) largely varies for different albums, suggesting that a unique photo of an object or a person does not have a higher chance to be selected, perhaps contrary to intuition. Departing from this latter conclusion, the next step would be to investigate the relation between the number of selected images in a cluster and its total number of images.

The average number of selected images per cluster of different size is given in Figure 5. It can observed that while for the scene clusters the median value of selected photos gradually increases with the size of the cluster,

Table 2. Selection ratio within albums. Ratio of selected photos represents overall ratio of selected photos. Ratio of selected clusters represents the ratio of clusters where at least one photo is selected.

	Ratio of selected photos	Ratio of selected scene clusters	Ratio of selected near-duplicate clusters	Ratio of selected singleton (one-image) clusters
Family event 1	0.570	0.931	0.868	0.700
Family event 2	0.326	0.883	0.779	0.333
Family event 3	0.328	0.729	0.553	0.550
Travel album 1	0.322	0.822	0.705	0.275
Travel album 2	0.390	0.875	0.704	1.000
Travel album 3	0.308	0.669	0.568	0.471
Average	0.374	0.818	0.696	0.555



(a) (b)

Fig. 5. Dependency between a total number of images in a cluster and a number of selected images within it. Data is aggregated across all albums, and the distributions represent number of selections by different users. Number in parenthesis indicates how many clusters of corresponding size present in the ensemble clustering data. (a) Selections made within scene clusters. (b) Selections made within near-duplicate clusters.

the same is not true for the near-duplicate level clusters. Their median value of selected images per cluster holds around 1, except for some larger clusters which do not provide reliable statistics, since there are only few large-size clusters found across the albums.

Although the subject of within-clusters photo selection merits further study, the above observations suggest that users tend to discard a considerable amount of content within clusters. As users tend to capture multiple repetitive photos to ensure a good result, they also inherently create a lot of redundancy in their collections, increasing the time and effort required to manually review them. Consequently, automatic approaches to both cluster and rate photos in an integrated manner might be beneficial.

5 AUTOMATIC APPROACH FOR PHOTO ALBUM CLUSTERING

Following our analysis of users' behavior in clustering and selection within photo albums, we explore automatic solutions for both tasks. In this section, we focus on automatically clustering photo albums, and we propose a method to partition photo albums into related moments, similar to users' decisions. Starting with the previously

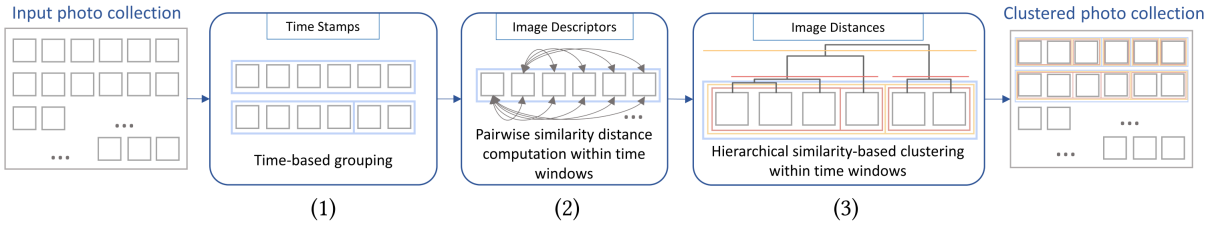


Fig. 6. Clustering procedure framework.

proposed hierarchical clustering approach [22], we introduce several changes that provide more robust clustering on a larger variety of photo albums.

We propose a photo album clustering method that groups together similar photos related to the same captured moments, while preserving linear time organization. This method serves two purposes: (a) provide the basis for a user interface that facilitates album browsing for photo selection, and (b) automatically extract the context of each photo, which can be subsequently used for different purposes, such as an album-based adaptation of independent image scores.

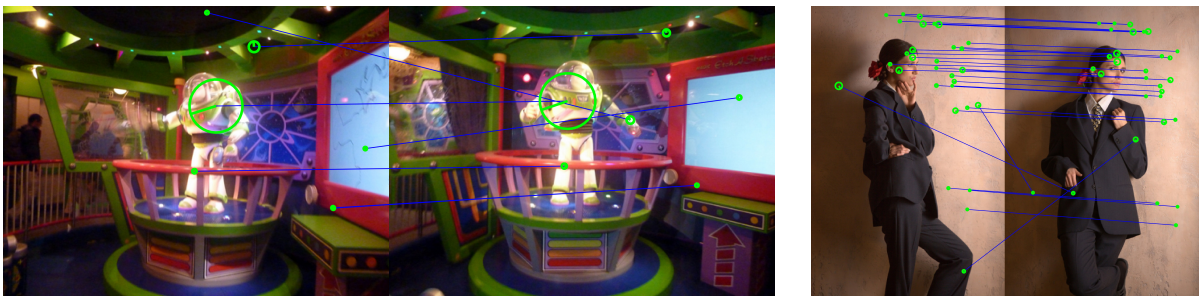
The clustering of an album is performed in three steps, as illustrated in Figure 6. First, time-based groups are detected within the entire album. Second, an image similarity distance is computed for each pair of images inside each time group. Then, the computed distances are used to create a hierarchical tree representing the structured relationship between images. The obtained hierarchical trees within each time window are cut at two levels to obtain the two-level clustering: the resulting scene and near duplicate clustering levels represent different level of visual similarity between photos. The expected method's output is a photo collection, clustered similarly to a user clustering demonstrated in Figure 4.

In the following sections, we provide a detailed description of each step of the proposed method. We also describe two image similarity distances based on different types of image descriptors, which serve to guide the hierarchical clustering: a SIFT-based image distance and a CNN-based image distance. Finally, we describe our proposal of an adaptive cut for hierarchical clustering.

5.1 Time-based Grouping

In our approach we aim to preserve the time linearity of the album by only grouping similar images if they correspond to approximately the same moment, and conversely, avoid grouping similar images if they are taken at different times (e.g. if a user revisits the same location on another day). To better preserve such time linearity and avoid clustering similar scenes from different time occasions together, we begin with a prior time-based grouping of the collection.

The entire album is split into sequential temporal windows using the available photo time stamps, extracted from the EXIF data. If no reliable timestamps are available, the entire album is treated as a single temporal window in the subsequent steps. The temporal windows are computed using the method proposed by Platt et al. [38]. Their approach presents a useful property for us: as the window boundaries depend on time differences across the image neighborhood, the granularity of time clustering is automatically adapted to the time span of the album. If, for example, the input album spans one hour, where photos from the same time window are separated by seconds, the different time windows are split apart by minutes. In contrast, if the input album represents a day trip, where photos in the same time window are separated by minutes, the time window gap can approach one or several hours.



(a) (b)

Fig. 7. Examples of SIFT-descriptor based matching limitations. (a) The presence of significant motion blur can affect the image gradients and keypoints' characteristics, thus leading to a low number of matches. (b) In case of considerable change in human pose, the most prominent keypoints can become hidden. Image credits: Sean MacEntee and Nikolay Kuzovkin.

The obtained temporal windows serve as a basis for further similarity-based clustering. By performing hierarchical clustering within each window, the images are clustered into scene-level and near-duplicate level clusters, representing different degrees of similarity. The hierarchical clustering requires a distance metric computed for each data pair. In our case the distance is based on the visual similarity between two images, which is computed for each image pair inside the temporal window.

5.2 SIFT-Descriptor based Similarity Distance

Similarly to our earlier approach [22], we first explore a metric based on SIFT features [26]. For each pair of images I and J, two sets of SIFT descriptors are compared using the Euclidean distance, and the number of matches [26] is computed for both directions $m_{I\to J}$ and $m_{J\to I}$. Then, the distance between the two images is defined as follows:

$$d_{SIFT}(I,J) = 1 - \frac{M(I,J)\cdot P(I,J)}{N(I,J)}, \tag{1}$$

where $M(I,J)$ is the average number of matches $M(I,J) = (m_{I\to J} + m_{J\to I})/2$, and $N(I,J)$ is the average number of detected SIFT descriptors in both images $N(I,J) = (n_I + n_J)/2$. $P(I,J)$ is defined as a measure of pair matches consistency:

$$P(I,J) = \frac{\min(m_{I\to J}, m_{J\to I})}{\max(m_{I\to J}, m_{J\to I})}. \tag{2}$$

In contrast to the earlier approach described in [22], we now add the average number of descriptors $N(I,J)$ as a normalization factor in the distance computation. This ensures that our distance metric is bounded on the interval $[0,1]$, which in turns allows for the use of an adaptive approach to determine the thresholds for cutting the hierarchical tree, removing the fixed thresholds of the previous method.

5.3 CNN-Descriptor based Similarity Distance

The traditional SIFT descriptors are generally capable of identifying matches between images even in the presence of distortions and rotations, which is advantageous for the task of matching a series of similar photos. However, their capability is limited in the presence of strong blur, large in-scene rotations or significant pose changes of the objects (e.g. Fig. 7). Such changes can be handled more effectively by image descriptors based on the activations

of convolutional neural networks, as they are able to provide a more generalized representation of image features. Due to this, in our approach we have employed another similarity metric, based on CNN-computed global image descriptors [39].

Radenović et al. [39] proposed a CNN fine-tuning scheme to train global image descriptors for the task of image retrieval (CNNR). As a typical global image descriptor, it can be applied to other tasks as well. In our case, we use their descriptor computation based on the fine-tuned ResNet [14], which provides a descriptor vector f of dimensionality 2048 for each image. The vector f is l_2-normalized, therefore similarity between two images can be evaluated with their inner product. As the later step of hierarchical clustering is based on the distance measure, we compute the CNNR-based image distance between images I and J as follows:

$$d_{CNNR}(I, J) = 1 - f_I^T f_J, \tag{3}$$

where f_I and f_J represent the l_2-normalized feature vectors, and the computed distance d_{CNNR} is also bounded on the interval $[0, 1]$.

5.4 Hierarchical Clustering with an Adaptive Cut

The hierarchical clustering approach has several useful properties for photo albums clustering. First, the hierarchical tree structure reflects the similarity connections between photos in an organized manner: the height of tree branches reflects the visual similarity. Second, by cutting the hierarchical tree at different heights, we obtain output clusters corresponding to different levels of similarity. This way, it is also possible to obtain clusters that are enclosed into each other, which reflects the natural organization made by photographers. For instance, in our approach, photos are connected on the near-duplicate level of high similarity and on the scene level of more general similarity. No predetermined number of clusters is required as an input, which is crucial when dealing with different photo collections, as no preliminary information on the potential number of clusters is available.

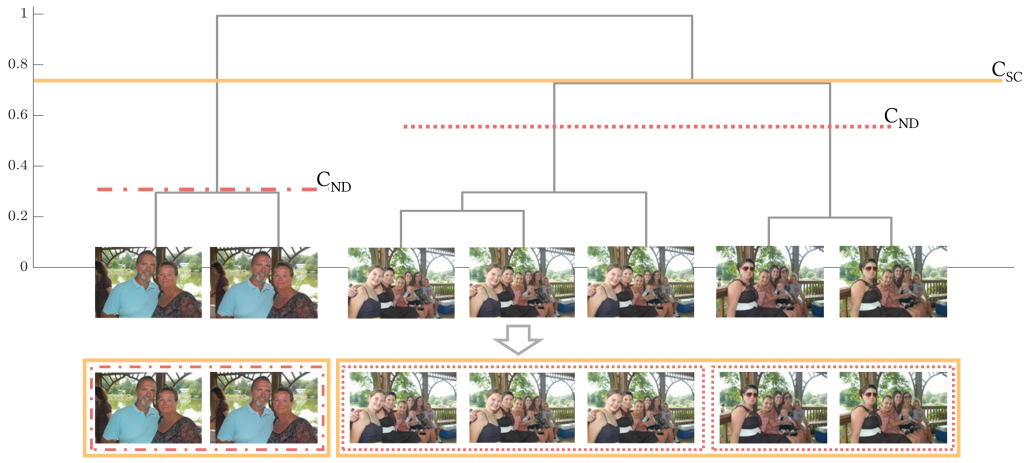


Fig. 8. Example of adaptive granularity in users' decisions and a corresponding clustering cut adaptation. The photos from the second scene (five photos on the right) contain many repetitions, so users tend to find possible differences between them, creating two separate near-duplicate clusters. Our hierarchical clustering approach performs in a similar manner. After creating a distance-based tree, this tree is cut on two levels. First, using the collection-defined threshold C_{SC} to obtain division into scene clusters. Second, the near-duplicate clustering threshold C_{ND} is computed within each scene, to further divide it into the clusters of near-duplicate photos.

ALGORITHM 1: Hierarchical clustering via adaptive cut computation

Input : Set of pairwise image distances $d(I, J)$ in album $D_A = \{d(1,2), d(1,3), ..., d(I_A, J_A)\}$, set of temporal clusters $T = \{t_1, t_2, ..., t_i\}$

Output: Set of scene clusters $S = \{s_{i_1}, s_{i_2}, ..., s_{i_j}\}$, set of near-duplicate clusters $N = \{n_{i_{j_1}}, n_{i_{j_2}}, ..., n_{i_{j_k}}\}$

Compute scene and near-duplicate level clusters within each temporal cluster:

$C_{SC} = \overline{D_A}$;

foreach *temporal cluster* t_i *in* T **do**

  $D_{t_i} = \{d(1,2), d(1,3), ..., d(I_{t_i}, J_{t_i})\}$;

  $tree(t_i) \leftarrow$ `ConstructHierarchicalTree`(D_{t_i});

  $S_i = \{s_{i_1}, s_{i_2}, ..., s_{i_j}\} \leftarrow$ `PerformCutIntoSceneClusters`$(tree(t_i), C_{SC})$;

  foreach *scene cluster* s_{i_j} *in* S_i **do**

    $D_{s_{i_j}} = \{d(1,2), d(1,3), ..., d(I_{s_{i_j}}, J_{s_{i_j}})\}$;

    $tree(s_{i_j}) \leftarrow$ `ConstructHierarchicalTree`$(D_{s_{i_j}})$;

    $C_{ND} = \overline{D_{s_{i_j}}}$;

    $N_{i_j} = \{n_{i_{j_1}}, n_{i_{j_2}}, ..., n_{i_{j_k}}\} \leftarrow$ `PerformCutIntoNearDuplicateClusters`$(tree(s_{i_j}), C_{ND})$;

  end

end

Similarly to our earlier proposed approach [22], we adopt the agglomerative hierarchical clustering approach, also called the "bottom up" approach, where each element (image) belongs only to its own cluster at the starting point, and pairs of clusters are merged together in a hierarchical manner until the tree is formed at the top of hierarchy. Then, typically, the tree can be cut at different levels to obtain image clusters of different similarity, as described in [22]. In the previous approach, the cut thresholds are defined on two fixed levels, which allow to split the images into scene level and near-duplicate level clusters. However, the fixed cut thresholds limit the flexibility of the approach and cannot successfully adapt to different types of collections.

After observing the user clustering behavior, we noted that when presented with a scene of only few repetitive similar photos, even if the photos are not exactly near-duplicates of each other, users tend to keep them in the same near-duplicate cluster. On the contrary, when presented with many repetitive photos from the same scene, users tend to highlight differences between photos and split them into clusters of higher granularity (e.g. Figure 8).

To simulate the user behavior, we define the cut threshold in an adaptive manner. The outline of the main steps is given in Algorithm 1. The scene clustering cut threshold C_{SC} is defined as the mean value $\overline{D_A}$ of the image distances in the entire album. Then, a hierarchical tree is constructed inside each temporal window t_i, and the scene clustering cut C_{SC} is applied on this tree to create separate scene clusters. At the second stage, each scene s_{i_j} is further split into near-duplicate clusters in a similar manner, where the cut C_{ND} is defined as the mean value $\overline{D_{s_{i_j}}}$ of the image distances in the processed scene s_{i_j}. An illustration of this process can be found in Figure 8.

We have also experimented with different grouping criteria for constructing the hierarchical tree. In [22] the entire tree was linked using a *single linkage* criterion, where the distance between two potential clusters to link is defined as the shortest distance over all pairs of images. This approach often leads to the chaining phenomenon, when more distant clusters are grouped together through a chain of intra-similar elements between them. This was found to be beneficial for the SIFT-based clustering on the scene level, as the lower number of matches may be not sufficient to link images with a smaller visual overlap. However, for the clustering based on CNN features, the higher abstraction level of the descriptors led to the over-merging of scene clusters. In this case, a *complete linkage* strategy is preferable, where the distance between farthest elements in two candidate clusters is

considered, leading to more compact clusters. Therefore, for the scene tree construction, we use *single linkage* in the case of SIFT descriptors, and *complete linkage* in the case of CNN descriptors. For the near-duplicates tree construction inside each scene, we use *complete linkage* for both types of descriptors, as its compactness property aids in our adaptive process, where the users' tendency to find clusters of higher granularity on the near-duplicate level.

6 CLUSTERING PERFORMANCE

The performance analysis for our clustering method is conducted in a similar manner to the analysis of users' clustering agreement. The Adjusted Rand Index is computed between the computed clustering partitions and the partitions provided by each user, and then the average value is computed.

Table 3. Average *ARI* performance of the analyzed clustering methods. For our proposed methods, the prefix *Time-* indicates a preliminary temporal clustering applied before the similarity-based hierarchical clustering.

	User Agreement	Time-based clustering [38]	Affinity propagation [11]	SIFT [22]	Time-SIFT [22]	Adaptive SIFT	Adaptive Time-SIFT	Adaptive CNNR	Adaptive Time-CNNR
Scene level	0.658	0.372	0.459	0.471	0.533	0.481	0.551	0.573	0.580
Near-duplicate level	0.678	0.116	0.407	0.411	0.411	0.606	0.629	0.610	0.624

First, we compare the average performance of several clustering methods, as shown in Table 3. We compare the proposed adaptive hierarchical clustering with our earlier proposed approach [22] and with existing state-of-the-art methods. For this purpose, we have selected two clustering methods that, similarly to our method, do not require a prior estimate of the potential number of clusters and thus can operate completely automatically. The first is the time-based clustering by Platt et al. [38], which we have described in Section 5.1. The second is the affinity propagation clustering proposed by Frey et al. [11]. Their approach iteratively searches for the most representative exemplars, while the associated data points are used to define cluster boundaries. This approach was applied to the task of clustering images of human faces and to perform image data summarization [9]. Since their approach works with any similarity measure, we employ it using the earlier described CNNR-descriptors based similarity, as it has shown the best performance in our tests. Additionally, as each of these two methods does not provide a specific multi-level structure and produces one clustering output, we evaluate this output both against our scene and near-duplicate user partitions.

The time-based clustering provides the lowest performance compared to the other methods tested. It is not unexpected, since the temporal information alone is generally not sufficient for such task. The affinity propagation clustering based on visual similarity shows reasonable performance for both clustering levels, close to our previously proposed SIFT-based hierarchical clustering, when temporal information was not used. Nevertheless, as this method was not designed for multi-level clustering, its performance is lower than our proposed approaches.

For the proposed hierarchical clustering solutions, the detailed per-album results are given in Table 1 (detailed results are given only for the time-aided method), and the results for the entire dataset are summarized in Figure 9 for both the time-aided method and for the case when no time information is available.

When using an adaptive cut threshold, our method largely outperforms the original Time-SIFT method of Kuzovkin et al. [22]. It is also interesting to note that the Adaptive Time-SIFT generally outperforms the Adaptive Time-CNNR method for the albums focused on people's photos, while the CNNR-based method shows better performance in landscape-focused albums, where often the landmarks are more abstract and more challenging to model by SIFT descriptors.

We have also observed that *Travel album 3* is particularly challenging for the SIFT-based methods, since it does not contain reliable time information to perform the first pre-clustering and facilitate further similarity-based

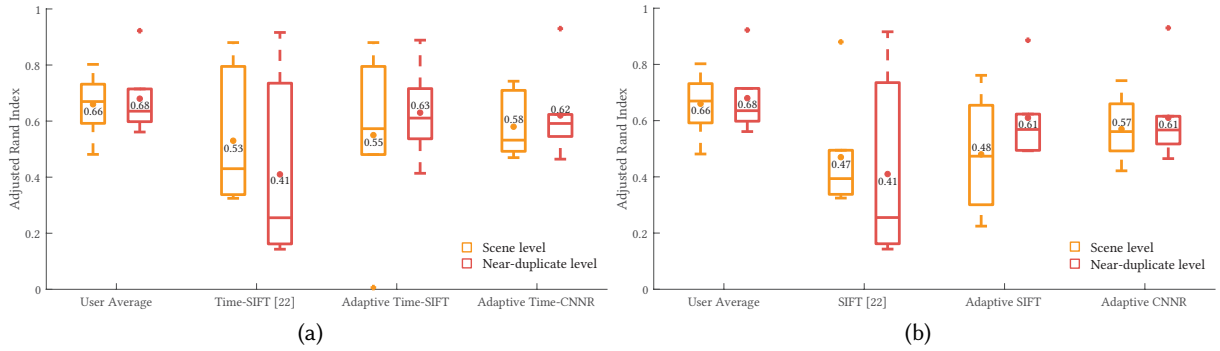


Fig. 9. Performance of clustering methods. The box plots represent *ARI* distributions across the albums, where a dot and the associated value indicate the average *ARI* for each method. (a) Performance with an aid of preliminary temporal clustering. (b) Performance with no time information available.

clustering steps. Also, it contains numerous low quality blurred shots, with large viewpoint changes, limiting the ability of SIFT to detect appropriate matches (as it was demonstrated in Figure 7). In this case, the CNNR descriptors generally demonstrate better matching performance.

Another test case that we have considered is the potential absence of time information, thus no time-based pre-clustering can be performed, and the entire collection is clustered directly. The average performance for this scenario can be found in Figure 9b. It should be noted that we have kept the adaptive cut computation the same for the Adaptive CNNR method, but we have introduced a small change for the Adaptive SIFT method. Considering possible instabilities of SIFT distance in lack of reliable image matches, we have defined the scene cut threshold as $C_{SC} = \overline{D_A} - \sigma_A/2$, when the time stamps are not available. The corresponding results are reported.

The performance of the SIFT-based methods noticeably drops for the scene level in the absence of time information. This also confirms the limited capability of SIFT-based matching to find the matches of higher abstraction, which are necessary in this case. At the same time, no significant deterioration in performance is observed for the CNNR-based method, for both clustering levels. Therefore, the CNNR-based clustering is more robust overall and is able to provide relevant results even with no time information available (although the time linearity of the photo album might be not preserved in this case).

7 CONTEXT-BASED SCORE ADAPTATION

In this section we explore an application of the proposed automatic clustering for context-adaptive photo assessment. As we observed in Section 4, users tend to consider the context of photos when assessing them. We attempt to model the user preferences by using an independent image score together with the obtained clustering information. In our analysis, we compare the computed image scores with user preference scores [23], which reflect how often each photo in an album was selected by users.

For the automatic score computation, we adopt three independent photo assessment metrics, which were also previously examined in [23]. All analyzed approaches are based on CNN models and recently demonstrated state-of-the-art performance in photo evaluation:

- The approach by Kong et al. [19], where importance of different photographic attributes is weighted by the image content, giving proper relevance to what should be considered important in an image.

- The method proposed by Jin et al. [16], which introduced a fine-tuning scheme with sample weights to assess images from a wide range of possible origin and aesthetic quality.
- The NIMA method proposed by Talebi et al. [43], which evaluates both image quality and aesthetic attractiveness of an image.

7.1 Z-score based Adaptation

As a first context-adaptation approach, we apply our previously-proposed method, which is based on the direct z-score weighting [22]. In that approach, an independent score for each photo in the album is transformed into three z-scores [20] for three different context levels: collection level, scene cluster level and near-duplicate cluster level, using the statistics of the entire collection and different corresponding clusters:

$$z_{I_L} = \frac{s_I - \mu_L}{\sigma_L}, \tag{4}$$

where μ_L and σ_L denote mean and standard deviation of the scores, computed on one of three levels $L \in C, SC, ND$. The obtained z-scores for each level are combined into the global score Z_I, as their average. We consider Z_I as the new album-adapted image score.

The performance of z-score based adaptation is given in Table 4 (original scores represent the original methods' results without adaptation applied). The Pearson correlation coefficient was computed between the computed scores and the user preference scores [23] and averaged across all albums. It can be seen that the z-score adaptation results only in limited improvement over the original scores. In addition, the advantage of better clustering is not evident in this case, as often *Adaptive Time-SIFT* clustering and *Adaptive Time-CNNR* clustering provide lower correlation than the adaptation based on *Time-SIFT* clustering [22]. Finally, instead of automatically computed clustering, we apply the ensemble *User clustering* (Section 4) as an adaptation base, to estimate the overall feasibility of the z-score adaptation approach. The adaptation based on the *User clustering* also does not lead to a clear improvement of the results, which suggests that the employed adaptation approach may be too simplistic to model the user behavior.

Table 4. Correlation with the user preferences after z-scores based context adaptation [22]. *Original scores* column represents the original methods' results without adaptation applied. *User clustering* column represents the results of adaptation using the ensemble clustering acquired in the clustering study. Other columns represent results of adaptation with automatically computed clustering.

	Original scores	User clustering	Time-SIFT [22]	Adaptive Time-SIFT	Adaptive Time-CNNR
Kong et al. [19]	0.274	0.232	**0.261**	0.248	0.217
Jin et al. [16]	0.230	0.254	**0.260**	0.252	0.243
NIMA [43]	0.169	0.248	0.217	0.185	**0.219**

7.2 Neural Network based Adaptation

As a more powerful alternative to the z-score based adaptation of image scores, here we explore a different approach for predicting user scores, based on a shallow neural network trained for a regression task. The utilized network is a multilayer perceptron with one hidden layer consisting of ten neurons and the tan-sigmoid function used before the output layer with the linear function. The Levenberg-Marquardt algorithm [24] is used as the optimization approach.

As the input features, we enrich the independent photo score with the following statistics computed from the cluster partitions: (1) original image score s_I, (2) collection level z-score z_{I_C}, (3) scene level z-score $z_{I_{SC}}$, (4)

near-duplicate level z-score $z_{I_{ND}}$, (5) number of images in the scene cluster $|SC|$, (6) number of images in the near-duplicate cluster $|ND|$.

For the train-test process we use the nested cross-validation procedure. In the nested cross-validation, an outer k-fold loop is used to divide the data into training and test folds. Then, each training fold is further split into training and validations folds on an inner loop, to select the best model parameters. In our case, on each outer loop, one album is set apart as a test fold, to test model's performance on it. In the inner loop, two albums are used in the validation fold, and the remaining albums are used in the training fold. By testing on the test fold album from the outer loop, we avoid possible model overfitting.

Table 5. Correlation with user preferences after neural network based context adaptation.

	Original scores	User clustering	Time-SIFT [22]	Adaptive Time-SIFT	Adaptive Time-CNNR
Kong et al. [19]	0.274	0.546	0.485	0.482	**0.489**
Jin et al. [16]	0.230	0.517	0.440	**0.463**	0.439
NIMA [43]	0.169	0.543	0.415	**0.459**	0.457

For each clustering approach, we retrain and test the network on its corresponding clustering results. The obtained results are shown in Table 5. Compared with the original scores and the straightforward z-score based adaptation, this method achieves a significant improvement. The adaptation based on the *User clustering* provides the highest increase in correlation with the user preferences, which confirms the feasibility of the chosen adaptation. Moreover, all three automatic clustering approaches lead to improvement over the original unadapted scores. The score adaptation using the *Adaptive Time-SIFT* or the *Adaptive Time-CNNR* clustering methods provides similar performance. Overall, the best performance is achieved for the method of Kong et al. [19], using the *Adaptive Time-CNNR* clustering as an adaptation base. At the same time, for two other scoring approaches, their best performance is achieved using the *Adaptive Time-SIFT* clustering. Nevertheless, the adaptation results between these two clustering approaches are comparable, conforming with a similarity in their clustering performance.

The obtained results confirm that the clustering information can be successfully used to improve the performance of the independent photo assessment solutions. The *User clustering* provides the largest improvement, since the clustering partitions are derived directly from user decisions in this case. At the same time, the automatic clustering solutions also show a degree of correlation with user preferences, suggesting that the context of a photo as defined by this form of clustering can assist in the task of automatically scoring and selecting photos within albums.

Table 6. Ablation study on data features in a neural network based adaptation. The average correlation is shown for the independent scores by Kong et al. [19] and the adaptation output, when the input is the original score supplied with indicated data feature(s). We can see that the best performance is achieved when all the features are used together.

	Original scores [19]	z_C	z_{SC}	z_{ND}	$\|SC\|$	$\|ND\|$	(z_C, z_{SC}, z_{ND})	$(\|SC\|, \|ND\|)$	All features
User clustering	0.274	0.371	0.393	0.413	0.477	0.455	0.424	0.520	0.546
Adaptive Time-CNNR		0.365	0.393	0.423	0.431	0.450	0.409	0.470	0.489

As an additional point of interest, we have performed an ablation study, where our solution has been trained and tested using only a subset of data features. From the results given in Table 6, we can conclude that the introduction

of scene and near-duplicate cluster sizes provides an important contribution to the overall performance. This finding also corresponds with our user study conclusion on dependency between a total number of images in a cluster and a number of selected images within it. At the same time, while the z-scores based features alone provide lower improvement, they provide a noticeable increase in performance in case of complete feature combination. On the whole, the ablation study demonstrates an importance of each data feature and confirms our earlier intuition on the nature of photo selection.

8 CONCLUSION

To better understand how users manage large photo collections with often repetitive content, we have performed a user study evaluating photo clustering within albums. We have observed that although overall users achieve a certain level of agreement in their decisions, particular types of content appear to be easier to cluster than others. More specifically, higher user agreement is achieved for albums focused on events with people present. This conclusion also conforms with our earlier study [22] on the nature of selections in photo albums, where we have found that users achieve higher selection agreement for photos of people. In addition, we have incorporated the user selections into the output clustering structure obtained in the current study, and we have used this aggregated data to acquire useful statistics on user decisions within the clusters.

An automatic clustering method for photo collections was also proposed, based on the hierarchical clustering approach and augmented by an adaptive cut technique and the use of deep features. The performance of the automatic clustering solutions was evaluated against the acquired user clustering data. The proposed CNNR-based clustering with an adaptive cut demonstrated the most robust performance, even in absence of time information in the album. Finally, we introduce an application of context-based score adaptation, where a score obtained from the independent photo assessment method is adapted to the context extracted via clustering. The obtained results show that a score adaptation using a small neural network based on the features acquired by means of the *Adaptive Time-CNNR* clustering can bring significant improvement in correlation with user provided preference scores.

Our studies on user clustering and selection decisions also demonstrate that for certain types of content, user agreement can be noticeably low. In these cases, the potential of the automatic modeling would be limited, as it would be impossible to create a model that would suit every user. This in contrast suggests that personalized automatic solutions, which could adapt not only to the nature of the photo collection, but also to the individual way of photo organization by a user, would be valuable, presenting an interesting direction for future work.

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] Liangliang Cao, Jiebo Luo, Henry Kautz, and Thomas S. Huang. 2008. Annotating collections of photos using hierarchical event and scene models. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2008).

[3] 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* July (2015).

[4] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, 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*. ACM, 187–194.

[5] 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.

[6] 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*. ACM, 829–832.

[7] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. 2005. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications* 1, 3 (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] Delbert Dueck and Brendan J Frey. 2007. Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

[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] Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science* 315, 5814 (2007), 972–976.

[12] Ai Gomi, Reiko Miyazaki, Takayuki Itoh, and Jia Li. 2008. CAT: A hierarchical image browser using a rectangle packing technique. In *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE, 82–87.

[13] Jesse Prabawa Gozali, Min Yen Kan, and Hari Sundaram. 2012. Hidden Markov model for event photo stream segmentation. *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2012* (2012), 25–30.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2, 1 (1985), 193–218.

[16] 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.

[17] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1733–1740.

[18] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. 2006. Understanding photowork. *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06* (2006), 761–770.

[19] 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.

[20] Erwin Kreyszig. 2007. *Advanced engineering mathematics*. Wiley publishing.

[21] Santhana Krishnamachari and Mohamed Abdel-Mottaleb. 1999. Image browsing using hierarchical clustering. In *Computers and Communications, 1999. Proceedings. IEEE International Symposium on*. IEEE, 301–307.

[22] 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*. ACM.

[23] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch. 2018. Image Selection in Photo Albums. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM.

[24] Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* 2, 2 (1944), 164–168.

[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 (2003), 390–402.

[26] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1150–1157 vol.2.

[27] 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.

[28] 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 *Proceedings of the IEEE International Conference on Computer Vision*. 990–998.

[29] Zhiwu Lu, Yuxin Peng, and Jianguo Xiao. 2008. From Comparing Clusterings to Combining Clusterings. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2 (AAAI'08)*. AAAI Press, 665–670.

[30] Yiwen Luo and Xiaoou Tang. 2008. Photo and Video Quality Evaluation: Focusing on the Subject. In *Proceedings of the 10th European Conference on Computer Vision: Part III (ECCV '08)*. Springer-Verlag, 386–399.

[31] Luca Marchesotti and Florent Perronnin. 2013. Learning beautiful (and ugly) attributes. *British Machine Vision Conference* (2013), 1–11.

[32] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. *Proceedings of the IEEE International Conference on Computer Vision* (2011), 1784–1791.

[33] 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)*. IEEE, 887–891.

[34] Andrew D Miller and W Keith Edwards. 2007. Give and Take: A Study of Consumer Photo-sharing Culture and Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 347–356.

[35] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2408–2415.

[36] 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)* (2017), 1206–1214.

[37] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. 2010. Large-scale image retrieval with compressed Fisher vectors. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3384–3391.

[38] John C. Platt, Mary Czerwinski, and Brent A. Field. 2003. PhotoTOC: Automatic clustering for browsing personal photographs. *ICICS-PCM 2003* 1 (2003), 6–10.

[39] F. Radenović, G. Tolias, and O. Chum. 2018. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1.

[40] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. 2015. The beauty of capturing faces: Rating the quality of digital portraits. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[41] Kerry Rodden and Kenneth R Wood. 2003. How Do People Manage Their Digital Photographs?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 409–416.

[42] Ben Shneiderman and Jack Kustanowitz. 2005. Meaningful presentations of photo libraries: rationale and applications of bi-level radial quantum layouts. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*. IEEE, 188–196.

[43] H. Talebi and P. Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[44] Xiaoou Tang, Wei Luo, and Xiaogang Wang. 2013. Content-Based Photo Quality Assessment. *IEEE Transactions on Multimedia* 15, 8 (Dec. 2013), 1930–1943.

[45] 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.

[46] Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. 2014. Smart Photo Selection: Interpret Gaze As Personal Interest. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2065–2074.

[47] 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.

[48] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.

[49] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The Design of High-Level Features for Photo Quality Assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, Vol. 1. IEEE, 419–426.

[50] 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*. ACM, 211–220.