Care to Share? Learning to Rank Personal Photos for Public Sharing

Ido Guy
Ben-Gurion University of the Negev, Beer-Sheva, Israel
eBay Research, Netanya, Israel*
idoguy@acm.org

Alexander Nus
eBay Research, Netanya, Israel*
alnus@ebay.com

Dan Pelleg
Yahoo Research, Haifa, Israel
pellegd@acm.org

Idan Szpektor
Google Research, Tel Aviv, Israel*
szpektor@google.com

ABSTRACT

With mobile devices, users are taking ever-growing numbers of photos every day. These photos are uploaded to social sites such as Facebook and Flickr, often automatically. Yet, the portion of these uploaded photos being publicly shared is low, and on a constant decline. Deciding which photo to share takes considerable time and attention, and many users would rather forfeit the social interaction and engagement than sift through their piles of uploaded photos. In this paper, we introduce a novel task of recommending socially-engaging photos to their creators for public sharing. This will turn a tedious manual chore into a quick, software-assisted process. We provide extensive analysis over a large-scale dataset from the Flickr photo sharing website, which reveals some of the traits of photo sharing in such sites. Additionally, we present a ranking algorithm for the task that comprises three steps: (a) grouping of near-duplicate photos; (b) ranking the photos in each group by their "shareability"; and (c) ranking the groups by their likelihood to contain a shareable photo. A large-scale experiment allows us to evaluate our algorithm and show its benefits compared to competitive baselines and algorithmic alternatives.

ACM Reference Format:
Ido Guy, Alexander Nus, Dan Pelleg, and Idan Szpektor. 2018. Care to Share? Learning to Rank Personal Photos for Public Sharing. In *Proceedings of 11th ACM International Conf. on Web Search and Data Mining (WSDM 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159713

1 INTRODUCTION

Digital photography, as enabled by smartphones, has changed the way people take photos and interact through them. Prior to smartphones and high-quality phone cameras, people preferred looking at photos together or in person over viewing them on a computer, while e-mail was the main vehicle for digital sharing [9]. When phone cameras and image uploading and sharing became ubiquitous, users started taking many ordinary and spontaneous photos of family, friends, and travel. The phone turned into a tool for creating and maintaining social relationships, sharing experiences (either personal or collective), and self-expression and presentation [2, 24]. Specifically, automatic image uploading from mobile phones to social media, such as Facebook and Flickr, or cloud storage, such as Google Drive and Dropbox, soared in popularity. And while some use the cloud to share all their photos (i.e., set their default sharing to "public"), many others upload their photos privately, as a means for backup or limited sharing with just friends or family [2, 16, 17].

This technological change led to new challenges with respect to photo sharing. First, due to volume, many of the uploaded photos remain part of a private repository, because users do not have the time or state-of-mind to go over the photos they took and decide which should be shared. Second, when users do go over their photos, e.g., once a day, a week, or a month, they face a daunting task. And while some users overcome it with a share-all policy, even they would welcome software to help save their followers from a deluge of boring photos. To back this intuition with data, we analyzed millions of photos uploaded from smartphones to Flickr. Our analysis shows that the total number of photos has been on the expected rise, fueled by increasing popularity of mobile photos. And yet, the *portion* of public photos has been on a constant decline.

To address these challenges, we propose the novel task of automatic recommendation of photos from a private collection for the purpose of public sharing. In other words, photos would be assessed by their likelihood to be shared. The most likely ones would be recommended to the user for the actual sharing act.

This novel task is by no means a simple one. Studies revealed that human considerations when selecting photos for sharing involve concerns about social disclosure, as well as factors that traditionally influence privacy management, like family concerns or social support [1, 2, 16]. While some of these considerations are personal and difficult to model, we expect that shareable and non-shareable photos can still be distinguished using automatic analysis.

Prior work found that digital camera users often take *groups* (or "bursts") of photos on the same object or scene, which are sometimes referred to as near-duplicates [6, 15]. Our analysis shows that when users manually select photos for sharing, they typically choose at most one of the photos in each such group. In this work, we therefore propose a three-step algorithm for the photo sharing recommendation task. In the first step, the target photo collection

*Part of the research was conducted while working at Yahoo Research.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159713

Figure 1: Photo stream, ordered from left to right in its original form (1st line), after running the three-step algorithm (2nd line), and after de-duplication (3rd line).

is segmented into near-duplicate groups. The algorithm then ranks the photos in each group by their likelihood to be shared. Finally, the groups themselves are ranked by their prospects to contain shareable photos. Following, we evaluate the final recommendation of photos for sharing under two alternative usage scenarios, in which, either: (a) de-duplication is important; or (b) duplicates are acceptable. If duplicates are unwanted, only the top-ranked photo from each group is presented to the user. Otherwise, several (or all) photos from each group are shown, ordered by the groups' ranking on top of their in-group ranking.

Figure 1 demonstrates the output of our algorithm on a daily photo stream contributed by one of the authors. First, the algorithm segments the stream into three scenes (whiteboard, child, fountain), each composed of between one and three near-duplicate photos. Then it ranks the photos in each group by their likelihood to be shared (e.g., ranking the photo of the fountain without the disrupting car first, or ranking the photo that best captures the child's face first). Then, the groups themselves are ranked by their likelihood to include a publicly shareable photo. Finally, de-duplication is performed (assuming it is desired in this case), leaving one photo per group on the final recommendation list.

We address all three steps in our approach as supervised subtasks. We consider near-duplicate grouping as a sequential segmentation of the photo collection over time and learn a threshold over similarity functions, given a training dataset. In addition, both group ranking and photo ranking within each group are treated as learning to rank (LTR) tasks. We trained our LTR models on a large dataset consisting of daily uploads of smartphone photos to Flickr, where (only some) photos are manually marked as public by the account owner. The goal in the LTR tasks is to rank the public (shared) photos and groups higher, as compared to the private ones.

We note that many works addressed the task of summarizing a photo collection [10, 15, 19, 21–23], but only a few recent studies suggested to select or recommend "interesting" photos from an album without the goal of summarizing the whole album [5, 25] (for more details, see Section 5). While we share some perspectives to these studies in terms of the algorithmic approaches and derived features, our task is inherently different: identifying private photos that are interesting for public sharing, as opposed to photos that are interesting as personal representatives of an album, or that are selected from already public collections.

To evaluate our algorithm, we conducted a large-scale experiment over hundreds of thousands of photos. We evaluated each sub-task of our algorithm, as well as the overall goal of recommending photos for sharing. We compared our three-step algorithm to a direct single-step algorithm that ranks all photos individually without the notion of groups. We also compared our algorithm to baselines such as the last photo taken or the most aesthetic photo. Our results indicate that our algorithm outperforms the respective baselines in each step, as well as in the overall recommendation task, regardless of whether de-duplication is desired or not.

To summarize, the main contributions of this paper are:

- Introducing a novel task of automatic recommendation of photos for sharing.
- Analyzing sharing behavior by Flickr users using smartphones.
- Suggesting a three-phase approach for the task, with a supervised algorithm for each stage, using a learning-to-rank framework.
- Implementing and evaluating the suggested approach using a large-scale log of selective sharing activity by Flickr users.

2 DATASET CHARACTERISTICS

To learn more about users' photo sharing behavior in social networks we analyzed Flickr, a popular social network focused on photos. Our dataset consists of basic metadata for all Flickr photos in the years 2004–2015. This metadata includes the date the photo was taken and the camera model it was taken with (available for over 80% of the photos). It also includes the sharing permissions of each photo, as set by the photo owner: at one extreme, private and accessible to the photo owner only; at the other extreme, public photos accessible to all users; and in-between, photos accessible to user-defined lists of friends, family, or both.

Our dataset indicates that the number of photos on Flickr rises every year. For example, in 2015 it was ×2.5 compared to 2012 and ×16 compared to 2004. This is driven by the increase in mobile photos — the portions rise from close-to-zero in 2004 to over 50% in 2015. Overall, in our dataset, 58% of the photos are private, 30% are public, and only 12% are in the middle ground for friends and family. In this work, we focus on the behavior of public sharing and leave its differences from limited sharing (with friends or family)

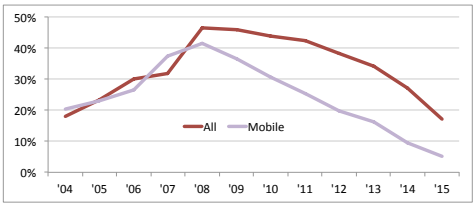

Figure 2: Portion of public photos per year for all Flickr photos and for mobile photos only.

Table 1: Characteristics of all photos and public-only photos in our experimental dataset: total counts and statistics of photos per user-day.

	Total	Avg	Std	Med	Min	Max
All photos	10.3M	58.8	110.1	31	7	7,804
Public photos	402K	2.3	4.3	1	1	191

for future work. We henceforth refer to all types of restricted access as "private" (70%).

Figure 2 shows the portion of public photos by year, for all photos and for mobile photos. From 2004 to 2008, the percentage across all photos increased, from 17.9% to a peak of 46.5%. Since then it is decreasing, down to 17.0% in 2015. One explanation to this sharp decrease is that many photos are submitted via automatic uploaders, which back up all the device photos and whose default permission is private (due to the sensitivity of this feature). Automatic uploading has become especially prevalent in mobile devices in recent years. Indeed, the percentage of public photos within all photos uploaded from mobile devices is much lower than within all photos (starting 2008). Combining this statistic with the surge in overall percentage of photos that originated from mobile devices points at the declining public photo percentage since.

Given our objective to predict which photos would be publicly shared by their owner, we constructed a more goal-oriented dataset. To this end, we focused on Flickr data from mobile devices in the years 2012–2016. We considered all unique pairs of users and calendar days, henceforth referred to as *user-days*. Each user-day consists of all photos taken by a single user in a single day, allowing us to examine sharing behavior at a one-day granularity. The average number of photos per user-day in our data is 7.9 (std: 20.3, median: 3, max: 13,231), with 30.2% of the days having one photo only. As we have seen before, most mobile photos are kept personal. Indeed, 91.8% of our user-days contain no public photos. On the other hand, in 6.9% of our user-days, all photos are public. For our experiments, however, we were interested in days for which both private and public photos exist, and in particular days for which the majority (but not all) of the photos are private, as these imply an explicit selection for sharing. After some analysis, we opted for user-days in which no more than 15% of the photos (but more than zero) are public[1].

The above selection left us with 177K user-days and 10.3M photos, to which we refer as the *experimental dataset*. Table 1 shows the characteristics of this dataset. Overall, the vast majority of the photos are private, with an average of 58.8 photos per user-day, of which only 2.3 are public. The percentage p of public photos per use-day spreads rather evenly across the range (0%, 15%], with an average of 7.8% (std: 4.1, median: 7.7%). For 30.8% of the user-days, $0\% < p \leqslant 5\%$; for 37.7% of the user-days, $5\% < p \leqslant 10\%$; and for 31.5%, $10\% < p \leqslant 15\%$. Finally, we note that the experimental dataset includes a large number of distinct users: 52.9% of them account for only one user-day in the dataset; additional 15.5% account for two user-days; and the maximum number of user-days per user is 828. In our experiments, unless otherwise stated, we randomly split the dataset into 90% training user-days and the rest for testing.

3 RECOMMENDING PHOTOS TO SHARE

3.1 Motivation and Overview

3.1.1 Algorithmic Approach. Our proposed algorithm consists of three steps. First, the target photo upload[2] is segmented into *groups* of near-duplicates. The algorithm then ranks the photos within each group based on their likelihood to be shared. Finally, the groups themselves are ranked by each group's likelihood to contain at least one shareable photo.

We support and experiment with two alternative recommendation scenarios: (a) the user prefers not to see near-duplicates, only a single exemplar from each scene; and (b) the user does not mind duplicates, as long as shareable photos are ranked high. Our algorithm generates the top N photos comprising the *recommendation list*, and can then be used in one of two ways. If near-duplicates are unwanted, the top-N groups that were ranked in the third step are selected. Then, the top-ranked photo from each group is considered, generating a ranked list of N photos. Alternatively, if de-duping is not essential, the recommendation list includes the top $k > 1$ photos from each group, concatenating them to a single list according to the group rankings.

3.1.2 Latent Vector Representation of Photos. Throughout the steps of our algorithm, we use a representation consisting of a 4K-dimensional vector for each photo. This *latent vector* was generated using a deep neural network model. Specifically, we utilized the YFNet-B network, which is used to power Flickr's automatic textual tags feature [**?**]. This network's training goal is different than ours, and its final layer performs classification over ~5,000 tags. Yet, the output of the penultimate layer can be viewed as a compressed "semantic" representation of the input photo, which is then used for inferring which tags are relevant to it. We used this output as our latent vector representation.

The full neural network includes 4 stages. In the first stage, 64 filters of size 7x7 each are applied, with stride 2. In the second stage, two layers of 128 filters are used, each factored by a 3x1 filter following by 1x3 filter, all with stride 1. The third stage consists of 256 filters, each similar to those in stage two (a 3x1 filter followed by a 1x3 filter, all with stride 1). Finally, the forth stage consists of two layers of 512 filters followed by a layer of 256 filters. All filters are similar to those in stages 2 and 3. Each convolution layer applies batch normalization and ReLU for non-linearity. The network was trained over the YFCC dataset [**?**], which consists of 100M photos with textual tags assigned by Flickr users. For further details see [**?**].

[1]Notice this implies each user-day includes at least 7 photos in total.

[2]In our experiments, these are daily mobile uploads.

Table 2: The different features and the steps in which they are used. Except for group-size features, group features are various aggregations over the feature score of all photos in the group: avg, std, variance, median, min, and max.

Feature Name	Step 1	Step 2	Step 3 representative	Step 3 group
color histogram diff.	×			
4K latent vector	×	×	×	
color distribution		×	×	
1,600 semantic tags		×	×	
inappropriateness		×	×	×
aesthetics		×	×	×
first in group		×	×	
last in group		×	×	
position from start		×	×	
position from end		×	×	
time delta from prev	×	×	×	×
time delta from next		×	×	
time delta from first		×	×	
time delta from last		×	×	
max photo score			×	×
group size (numeric)				×
10 group size indicators				×

We next detail the three steps in our algorithm, as well as the direct ranking algorithm. As a companion reference, Table 2 provides a summarization of all features used in each of the three steps.

3.2 Segmentation into Near-Duplicate Groups

As mentioned, we follow prior work that proposed to group photos which are considered near-duplicates [6, 15, 20]. To this end, we extracted a dataset consisting of 70 user-days[3]. Two of the authors annotated them for near-duplicate groups following three guidelines: (i) the photos in the group should capture the same scene; (ii) the photo subject stays the same within the group; and (iii) a user would not want to share more than one photo from the group.

Within each time-ordered stream of photos, we considered only consecutive photos as possible near-duplicates. Thus, annotators scanned the photo stream once, marking each photo that did not meet the guidelines for the current group as the beginning of a new group. Preliminary analysis showed that if the time gap between two photos is more than 15 seconds, they are rarely considered duplicates. We therefore automatically marked such cases (55% of all photos in our sample) as starting a new group, prior to any manual annotation, and excluded them from our gold dataset. This way, we directed the learning task to focus on the more difficult cases. In total, the two annotators labeled 1,486 photos with Cohen's kappa agreement of 0.77, measured on a shared subset of 130 photos. Overall, 70% of these photos were annotated as near-duplicates.

We constructed three features to be used as input for a classifier that determines whether a photo starts a new group (i.e., not a near-duplicate) or continues the previous one (i.e., a near-duplicate). The first two features have already been proposed in prior work on near-duplicates [15, 20]: (a) the time difference (in seconds) between the photo and its predecessor in the photo stream; and (b) the difference between the color histograms of the two photos. We computed the overall histogram difference by averaging the histogram difference for each RGB color, calculated using Hellinger distance.

Color histogram captures photo similarity at a rather low level. To compare photos at a more "semantic" level, we computed the cosine similarity between their latent vector representation, as described in Section 3.1.2. This high-level similarity was taken as a third feature, which, to the best of our knowledge, is novel for the near-duplicate task[4].

Considering the three features and our labeled gold dataset, we examined four types of classifiers for the near-duplicate task (using Weka [**?**]): logistic regression, support vector machine (SVM) with a Gaussian RBF kernel, C4.5 decision tree, and random forest. We used leave-one-out cross-validation to evaluate the performance of each algorithm. In each fold, we performed ten-fold cross-validation over the training set, for hyper-parameter tuning.

3.3 Ranking Photos Inside a Group

Applying group segmentation as described above on our entire experimental dataset showed that very small portions of the near-duplicate groups include more than one shared photo (full details are provided in Section 4). We therefore expect that when users share photos from such a group, they would often select the most appealing in terms of public view. Following, the second step of our algorithm selects the best "shareable" photo(s) within each group. We address this selection as a supervised ranking task, in which the goal is to rank photos that are likely to be shared above non-shareable ones.

To this end, we culled the experimental dataset to only include groups that contain at least one shared photo and one non-shared photo, since only these are useful for training and evaluating the in-group ranking task. We then trained a learning-to-rank (LTR) algorithm, where the target was to rank, in each training group, the shared photos higher than the non-shared ones. As our LTR framework we used an online variant of SVMRank [3] with AROW update [8], whose performance was found competitive for large-scale ranking [4]. The final score provided by the ranker for the top-ranked photo is denoted as the *max photo score*.

As input to the LTR algorithm, we derived from each photo a set of features (see summary in Table 2). Two types of features were considered. The first type refers to features that are extracted from each photo independently. As low-level features, we computed a distribution over a set of 16 main colors in each photo (red, blue, yellow, green, etc.). As high-level features, we included the 4K-dimensional latent vector described in Section 3.1.2. Additionally, we computed semantic features per photo, which were derived from the latent vector. These include automatically-derived semantic tags, such as '*dog*', '*baby*', and '*beach*', each with its likelihood probability [**?**]. We filtered out tags with likelihood smaller than 0.5, resulting in a sparse tag set of only a few tags per photo. Yet, these tags may help identify photo objects that are attractive for sharing, or vice versa. For example, we noticed that the '*child*' tag correlates negatively with shareability. We also assessed the likelihood of a photo to be inappropriate (e.g., for nudity) based on an in-house training set manually labeled for appropriateness. We expect this feature to help in demoting embarrassing photos. Finally, following prior work that found aesthetics to be useful in photo selection [19, 21], we used the approach of Zhang [27] to

[3]Since this task involved looking at personal data, we only considered photos from users who explicitly gave consent for their data to be used for research purposes.

[4]We also experimented with calculating similarity between the photos' derived tags, however this measure performed poorly, due to the low number of distinguishing tags per photo.

compute a single feature whose (continuous) value estimates the degree of aesthetics for each photo.

The second type of features for each photo refers to its relationships with other photos in the group. These include position features: (a) is the photo first in the group?; (b) is it last?; (c) its ordinal position from the group's start; and (d) its position from the end. These features assess whether group position affects photo selection for sharing. We also included features capturing the time delta (in seconds) of the photo from: (a) the previous photo in the group; (b) the next photo; (c) the first photo; and (d) the last photo. These features may help in pointing at various photography efforts as captured by different pause and activity periods.

3.4 Ranking Photo Groups

The third and final step in our algorithm, denoted by *L2RGroups*, ranks the segmented photo groups as a whole. Its goal is to place groups that are expected to contain at least one shareable photo on top of groups whose photos are expected to not be shared. We treat this ranking sub-task as a supervised ranking task, employing the same LTR framework of online SVMRank with AROW update used in step 2 (Section 3.3).

In step 2, the ranked objects were photos within a single near-duplicate group, labeled as either shared or not. Therefore, the ranker would aim to find subtle differences between photos that indicate shareability. Here, on the other hand, the objects are groups within each daily user upload, and each group is labeled as shared only if it contains at least one photo that is labeled as shared. We therefore expect the ranker to rely on more coarse differences between the groups, to distinguish between different scenes, such as a sunset versus parking cars.

As input to the ranker, we derive two types of features from each group. First, we consider all the features extracted for a single photo in step 2, as well as the max photo score provided by the ranker in step 2. To this end, a single photo is selected for each group, from which these features are extracted. Notice that this selection may not necessarily be the same as the selection of shareable photos as derived from the in-group ranking (step 2): here we aim to select a photo representative of the group that would distinguish it from other groups, while at step 2 we aimed to identify the photos in the group that are most likely to be shared. In fact, since we hypothesize that the ranker would look for more coarse-grained differences between photos for the group ranking task, we expect any photo selection rule to be useful here. We examine this hypothesis below, when analyzing group ranking results (Section 4.3).

Features of the second type are derived from the group as a whole. These include the group's size, both as a single numeric feature and as a set of 10 binary bucket features (with the last feature referring to size 10 or higher). In addition, we generated aggregate group statistics for some photo features: aesthetics, inappropriateness, max photo score, and time delta from previous photo. These statistics include the average, median, variance, standard deviation, minimum, and maximum within the group for each target feature.

3.5 Ranking Individual Photos

As already mentioned, an alternative approach that addresses the shareable group recommendation task is an algorithm that ranks all photos by inspecting each one individually, without considering the notion of a group. We implemented this algorithm, denoted as *L2RIndPhotos*, using the same LTR framework of SVMRank with AROW update, and trained the model on user-days, with the goal of ranking shared photos higher than non-shared ones. As features for each photo, we used the features extracted in step 2 (see Table 2), but ignoring any features that are derived from relationships with group members.

We note that *L2RIndPhotos* can also be applied to a scenario where de-duping is desired. This can be achieved in a post-processing step, where each photo that belongs to the same group as a higher-ranking photo is removed from the ranked list. This way, we can also adapt *L2RIndPhotos* for group ranking in a drill-down scenario (see Section 3.1). To this end, after de-duping is performed, each photo in the recommendation list, which represents a single group, can be replaced by its corresponding group.

Table 3: Classifier performance comparison for the near-duplicate grouping task.

Classifier	Accuracy (%)
Random forest	84.7
SVM (RDF kernel)	84.2
C4.5 decision tree	84.1
Logistic regression	82.6

4 EXPERIMENTS

In this section, we describe our main experimental results. These include evaluation for step 1 – near-duplicate group segmentation; step 2 – ranking photos within each group; and step 3 – group ranking. Finally, we describe our evaluation of the full-fledged pipeline for recommending individual photos for sharing, either with or without duplication.

4.1 Near-Duplicate Grouping Evaluation

Our algorithm's first step aims at grouping the stream of photos into near-duplicate groups. In Section 3.2 we defined this as a binary classification task, and set out to evaluate four types of classifiers over our manually-annotated gold dataset. Table 3 presents the accuracy of each of the four algorithms. The random forest model achieved the best performance and was therefore our final choice. Its accuracy using all three features was 84.7%. In practice, the overall accuracy of group segmentation is even higher, since photos taken over 15 seconds after their predecessors are automatically considered as non-near-duplicates, as explained in Section 3.2.

Ablation tests using the random forest model (Table 4) show the advantage of high-level similarity. Using this feature alone, the classifier achieved 83.0% accuracy. On the other hand, leaving high-level similarity out, while only considering time difference and color-histogram similarity, as in previous studies [15, 20], led to a substantively lower accuracy, at 75.1%. This shows that this task requires semantic understanding of photos, while low-level information is insufficient.

4.1.1 Grouping Statistics. As pre-processing for the next algorithmic steps, we used the random forest model, learned on the entire gold dataset with all features, to group near-duplicate pairs

Table 4: Ablation tests for near-duplicate detection using a random forest classifier. *time* stands for the time-difference feature, *colsim* for color-histogram similarity, and *highsim* for the latent-vector (high-level) similarity. Accuracy is presented both when using only the respective feature ('Only') and when using all features but the respective feature ('Exclude'). 'Exclude' for all-features (i.e., no features) reflects always selecting the majority class.

Feature Set	Acc. – Only (%)	Acc. – Exclude (%)
all features	84.7	70.0
highsim	83.0	75.1
time	73.8	84.5
colsim	69.4	84.5

Table 5: Distribution of photos in our experimental dataset by the size of their associated group ('%'), the portion of groups with at least one shared photo ('% shared (groups)'), and the likelihood of an individual photo from the group to be shared ('% shared (photos)').

Group Size	1	2	3	4 – 5	6 – 10	11+
%	43.8	20.5	9.1	9.8	8.8	8.0
% shared (groups)	4.6	7.5	10.9	12.4	14.9	21.0
% shared (photos)	4.6	4.0	4.0	3.3	2.6	1.9

with time delta of 15 seconds or less in our experimental dataset. This pre-processing resulted in an average of 35.9 groups per user-day, 9% of which containing at least one publicly shared photo.

Table 5 shows the distribution of photos by the size of the groups they belong to. The majority of the photos belong to a *non-trivial* group (i.e., size greater than 1), which indicates that near duplicates are common in mobile photography and are therefore important to consider when designing a sharing recommendation algorithm.

Another important aspect of our segmented dataset is that only 1.0% of the non-trivial groups include more than one shared photo (8.6% of these groups have exactly one shared photo), indicating that indeed users typically do not bother to share more than one photo out of a near-duplicate group. Even for groups of size greater than 5, the percentage of groups with more than one shared photo is only 3.3% (16.4% have one shared photo). Table 5 shows that as group size grows, the chances of the group to contain a shared photo increase, but the likelihood of each individual photo in the group to be shared drops.

An additional interesting statistic is shown in Table 6 – the distribution of time delta between consecutive photo pairs in groups. As seen, a substantial portion of the pairs has a time difference of zero seconds, often the result of an automatic camera feature that takes more than one photo per shooting, e.g., photos with different exposure or photos with and without high dynamic range (HDR). In addition, most other pairs are within 1 to 5 seconds from each other, indicating rapid photography in order to capture a "good" photo of a single scene.

Table 6: Distribution (percent) of consecutive photo pairs within groups by time difference (in seconds).

Time Delta	0	1 – 5	6 – 10	11 – 15
group size > 1	36.6	41.4	14.6	7.4
group size > 5	44.8	43.9	8.0	3.3
group size > 10	50.8	42.0	5.3	1.9

Table 7: Performance of in-group ranking algorithms.

Algorithm	P@1	MRR
Chronological	0.301	0.591
Random	0.372	0.626
Distance from centroid	0.422	0.649
Reverse-chronological	0.430	0.660
Aesthetics	0.467	0.683
In-Group-LTR	0.520	0.721

4.2 In-Group Ranking Evaluation

The goal of our algorithm's second step is to learn to select the photos from a near-duplicate group, which are more likely to be shared compared to others in the group. For training and evaluation, we only considered groups with at least one public photo and at least one private photo, as explained in Section 3.3. Overall, our filtered dataset included 162K such groups across 82K different user-days, spanning a total of 684K photos. We compare our algorithmic approach, denoted further on by *In-Group-LTR*, to several baselines: (a) random order; (b) chronological – ranking by the time the photo was taken, from earliest to latest; (c) reverse-chronological; (d) distance from the centroid of the group, where the centroid is calculated over the entire feature space; and (e) aesthetics – ranking by the photo's aesthetic score, from highest to lowest. We used Precision@1 (P@1) as our main metric for this task, to reflect a selection of a single photo per group (i.e., duplication is not desired). For completeness, we also report the Mean Reciprocal Rank (MRR), which reflects the position of the highest ranking shared photo.

Table 7 reports the results. The chronological baseline yields the lowest performance, even lower than random, implying that the first attempt in a group is often a bad choice for sharing. Reverse-chronological ranking, on the other hand, is a better alternative, outperforming both a random choice, and, by a smaller margin, the distance from the centroid. We conjecture that the last photo in a group sometimes indicates user satisfaction with the outcome: after a good shot, no more are needed. Ranking by aesthetic score achieves even better performance, indicating that taking into account photo characteristics is preferable to simply considering its position within the group. Finally, our *In-Group-LTR* algorithm yields a further considerable performance enhancement, up to 52% in P@1, showing the benefit of learning to combine several high-level "semantic" and group-related features.

To further understand the contribution of the different feature categories to the task, we performed ablation tests by training the *In-Group-LTR* algorithm with each feature category by itself and, in addition, with all features excluding it. Results, presented in Table 8, indicate that the latent vector is the most valuable category, achieving high performance by itself (over 50% in P@1). It seems to encode aesthetics well enough, since the exclusion of the aesthetics score hardly has any effect. These results are consistent with

Table 8: Performance when using ('Only') or removing ('Exclude') subsets of features when training *In-Group-LTR*.

Feature Set	Only		Exclude	
	P@1	MRR	P@1	MRR
Position within group	0.394	0.638	0.520	0.721
Time delta within group	0.419	0.660	0.520	0.720
Aesthetics	0.467	0.683	0.519	0.721
Semantic tags	0.486	0.700	0.516	0.718
Latent vector	0.505	0.711	0.512	0.715
All Features	0.520	0.721		

Table 9: Performance of group ranking algorithms.

Ranking Algorithm	P@1	MRR	MAP
Random	0.090	0.242	0.218
Chronological order	0.146	0.277	0.257
Reverse-chronological order	0.164	0.300	0.279
Maximum aesthetic score	0.201	0.369	0.331
Group size	0.214	0.384	0.359
L2RIndPhotos	0.266	0.431	0.387
L2RGroups	0.310	0.470	0.427
L2RGroups – group features only	0.145	0.276	0.256
L2RGroups – no group features	0.304	0.465	0.421

the high performance achieved by high-level similarity features for the segmentation task (Section 4.1) and suggest this type of representation works well. In addition, semantic tags are useful, and combining any two of the three high-level categories (tags, aesthetics, and latent vector) results in performance similar to the full feature-set.

4.3 Group Ranking Evaluation

We next evaluate the sub-task of ranking groups by their likelihood to contain shared photos. Our training and test sets for this task are based on the experimental dataset described in Section 2. As evaluation metrics, we use Mean Average Precision (MAP) in addition to P@1 and MRR, since we are also interested in performance beyond the first hit (i.e., the first group with a shared photo). As discussed in Section 3.1, this sub-task also reflects a "drill-down" recommendation approach, in which groups are shown first and the user can then choose to view all photos of a specific group.

As mentioned in Section 3.4, the *L2RGroups* algorithm derives some of its features from a single representative photo within the group. The selection of the specific photo from which these features are derived is different in nature than the photo ranking task in step 2, which also involves the selection of photos from a near-duplicate group. Yet, here we are using the individual photo's features to derive more coarse-grained differences between groups, rather than subtle ones between individual photos within the same group. We therefore hypothesized that the selection of the specific photo from which the single-photo features are derived is not critical for the group ranking task. This hypothesis was proven correct: using different selection methods, such as first/last in group, highest aesthetic score, *In-Group-LTR*, and even random, yielded very similar performance for *L2RGroups*. For convenience, we opted to use *In-Group-LTR* as our representative selection method, yet as explained this is a rather arbitrary choice.

The direct algorithm *L2RIndPhotos*, which ranks photos individually, can be adapted to the sub-task of group ranking (Section 3.5). Following, Table 9 presents the performance of *L2RGroups* and *L2RIndPhotos* for group ranking, as well as various baselines. *L2RGroups* achieves the best performance across all metrics, by a large margin from *L2RIndPhotos* and the baselines. *L2RIndPhotos* performs better than the baselines, even though it does not include any group-related features. Among the baselines, ranking by group size, which does not consider any photo characteristics, achieves best performance. Indeed, Table 5 shows that larger groups are more likely to include a shared photo.

The bottom section of Table 9 shows the performance of *L2RGroups* with group-related features only and without them. It is clear that group features by themselves do not achieve high performance. Yet, their exclusion leads to some performance decline, indicating that they do contribute to the final model.

4.4 Photo Recommendation Evaluation

Finally, we tested the performance of different algorithms for the overall task of recommending photos to be publicly shared. We considered the following algorithms for the overall task: (a) Chronological - recommend the photos by the time they were taken (earliest first); (b) Reverse-chronological; (c) Aesthetic score – recommend the photos by their aesthetics (from highest to lowest); (d) *L2RIndPhotos*– learning to rank individual photos in a single step; and (e) *ThreeStepRanking*– our three-step algorithm, which segments into groups, ranks photos in each group (*In-Group-LTR*), and then ranks the groups (*L2RGroups*).

We evaluated two recommendation scenarios (see Section 3.1): (a) without filtering of near-duplicates; and (b) retaining only the top k photos in each near-duplicate group. When $k=1$, de-duplication yields a recommendation with only a single photo per group. For all algorithms that are unaware of groups, such as *L2RIndPhotos*, de-dupping can be performed as a post-processing step (Section 3.5). Evaluation is performed over the entire experimental dataset as described in Section 2. As evaluation metrics, we use MRR, P@K, and Recall@K (R@K), with $K \in \{1, 3, 5\}$, which enables us to inspect performance at different fixed recommendation list sizes.

Table 10 depicts the results for this experiment. Inspecting the algorithms' performance with de-duplication, both LTR algorithms significantly outperform all baselines. For example, *L2RIndPhotos* improves P@1 by 44% compared to the strongest baseline – aesthetic score. Yet, *ThreeStepRanking* performs substantially better, with an additional gain of 14.4% over *L2RIndPhotos* in MRR, and 17.5% in P@1. This gap in performance is consistent across all metrics and different recommendation list sizes, showing that splitting between group ranking and photo ranking within groups provides a leverage in ranking shareable photos high.

Interestingly, while increasing the size of the recommendation list enables the display of more shared photos (an increase of 95% in recall from R@1 to R@3, and of 29% from R@3 to R@5), precision substantially drops (a decrease of 69% from P@1 to P@3, 79% from P@3 to P@5). This shows that even for the top recommendations, it is quite difficult to identify the photos that will eventually be shared by the user. One reason may be that the choice of a single

Table 10: Performance of shareable photo ranking algorithms with de-duplication ($k=1,2$) and without it. The best result(s) in a column are boldfaced.

	Ranking Algorithm	MRR	P@1	P@3	P@5	R@1	R@3	R@5
De-dup ($k=1$)	Chronological	0.215	0.118	0.081	0.069	0.087	0.169	0.235
	Reverse-chronological	0.229	0.127	0.090	0.077	0.094	0.187	0.258
	Aesthetic score	0.272	0.143	0.110	0.095	0.105	0.234	0.323
	L2RIndPhotos	0.332	0.206	0.143	0.115	0.150	0.293	0.380
	ThreeStepRanking	0.380	**0.242**	**0.167**	0.132	**0.177**	0.345	0.445
($k=2$)	*L2RIndPhotos*	0.357	0.215	0.153	0.124	0.156	0.315	0.417
	ThreeStepRanking	**0.388**	**0.242**	**0.167**	**0.135**	**0.177**	**0.349**	**0.459**
No De-dup	Chronological	0.228	0.120	0.084	0.071	0.089	0.173	0.242
	Reverse-chronological	0.244	0.127	0.090	0.077	0.094	0.187	0.258
	Aesthetic Score	0.300	0.152	0.117	0.102	0.112	0.250	0.357
	L2RIndPhotos	0.369	0.218	0.157	0.127	0.158	0.324	0.430
	ThreeStepRanking	0.384	**0.242**	0.159	0.128	**0.177**	0.334	0.439

photo within a near-duplicate group may sometimes be arbitrary (all photos look the same) and therefore hard to reproduce.

If duplicates are allowed, the performance of *L2RIndPhotos*, which was trained for this scenario, improves by 5% to 10% in all metrics compared to the de-duplication scenario. On the other hand, the performance of *ThreeStepRanking* somewhat decreases for longer recommendation lists, yet it remains the best performing algorithm. Under the scenario where only two photos are shown per group ($k=2$), *ThreeStepRanking* performs best over all metrics, with a 3% increase in R@5 compared to full de-dupping. This scenario balances between showing a few alternatives from each group but still not over-populating the recommendation list with a single large group. Overall, these results indicate that ranking all photos is not the best option, especially when large groups are ranked high, from which only a single photo is usually shared (recall Section 4.1). We note that we also tested other values of k and found that *ThreeStepRanking* achieves its best performance when $k=2$, while *L2RIndPhotos*'s best performance is with no de-duplication.

5 RELATED WORK

The task we present in this paper – recommending private photos for public sharing – is novel. Yet, a large body of work considered related tasks in which photos are selected from a collection to represent or summarize the collection. We now describe some of these works, which influenced our algorithmic approach, the features used, such as aesthetic score, color histogram, semantic tags, and latent vectors, and the selected baselines.

Li et al. [15] created a summary of photos of a target event. They noticed that the closeness in photo acquisition time implies related image content, constituting a single *scene* within the event. Following, they segmented the photo stream into groups using timestamp and color-histogram differences. Then, the best photo in each scene was added to the summary. Their best-photo selection was based on identifying clear faces, assuming those are desired in a summary. Platt et al. [20] applied a similar approach using timestamp and color histogram similarities for segmentation. Graham et al. [10] inspected timestamp burstiness for creating clusters in a photo stream. They then selected cluster representatives, based on time differences, to form a summary, while ranking clusters by size. Chu and Lin [6] used Platt's timestamp-based clusters and then constructed a near-duplicate graph for each cluster, where edges indicated near-duplicate pairs. A representative was then selected to be the photo with the highest in-degree count. Other works examined unsupervised and supervised clustering of photo collections by events or by points of interest [7, 14, 18].

When summarizing a photo collection, different approaches considered complementing aspects, such as coverage and diversity. Singha et al. [22] summarized personal photo logs spanning several months. They introduced a framework that optimized over attractiveness, diversity, and coverage, as well as a concrete greedy algorithm. The photo factors were based on camera (EXIF, timestamp) and pixel features, as well as user-generated textual tags, face recognition, and photo location. Obardor et al. [19] composed a summary of a photo stream by splitting it into acts, i.e., long sequences of similar photos. They then selected photo representatives by combining aesthetics and narrative components. Guldogan et al. [11] automatically constructed a set of "interesting" photos by considering the view duration and the number of clicks for each photo in the user's photo collection. Tschiatschek et al. [23] summarized photo streams by evaluating each subset of the collection for coverage and diversity using a supervised mixture of sub-modular components. The components inspected how well a subset was similar to the whole collection, with similarity as a proxy to coverage and inner-subset dissimilarity as a proxy to diversity.

Several studies addressed the summarization of a collection of photos taken by several users. Jaffe et al. [12] defined collection summarization as a ranking task, in which the top K photos were selected after ranking. Their algorithm produced a hierarchical clustering of the photos based on their geo-location. Then, each cluster was scored by considering factors such as user and tag distinguishability and photo quality. Finally, the photos were recursively ranked by interleaving photos from prominent sub-clusters. Yang et al. [26] considered collections in which time settings are not calibrated. Their algorithm aligned the collections using visual and geo-location similarities. It then removed duplicates using greedy backward selection. Sadeghi et al. [21] grouped photos in multi-user collections, even when chronological boundaries were broken. They used a graphical model that considered the quality of a photo as well as pairwise similarity between all pairs in the album. The model was trained over manually-curated albums via Mechanical Turk. Photo features included face features, texture and color features, and photo aesthetics.

Recently, several works suggested to diverge from the traditional summarization approach, whose primary goal is to cover a target photo collection. Ceroni et al. [5] argued that a subset of important photos for a user need not cover all her collection. They proposed to select the top K photos deemed most important from a personal collection. They trained an SVM classifier on collections with self-labeled most important photos, and used it to rank the photos in a collection. Their features included quality-based aspects, such as contrast and blur; face recognition; high-level concept detection; and aggregate properties of the time-based clusters. Wang et al. [25] suggested that different types of albums may benefit from different types of importance prediction. They collected public photo "albums" (photos sharing many tags) from Flickr and tagged their types with Mechanical Turk. The annotators also marked the importance of each photo. They then trained a convolutional neural

network in two stages, first to predict importance, then re-training the network's output stage for each album type.

6 CONCLUSIONS AND FUTURE WORK

We introduced the novel task of recommending private photos for public sharing. Such recommendation is especially helpful with the ubiquitous automatic upload, where mobile phone users need to sift dozens of photos that are uploaded daily into social networks and cloud storage services. We presented a three-step supervised ranking algorithm, which first groups together near-duplicates, then selects the best photo candidates for sharing in each group, and finally ranks groups by their likelihood to include a shareable photo. The selected photos from the top ranked groups are presented to the user as recommendations for sharing.

We conducted a large-scale experiment over a dataset of millions of photos uploaded from mobile phones, in which the owners of the photos manually selected just a few to be shared. Our algorithm outperformed other alternative algorithms and baselines, placing photos that were indeed shared higher on the recommendation list than the competitors.

To enable evaluation at large scale, we relied on a specific type of user behavior on Flickr, reflected by selective sharing of photos per-user per-day. As we showed, this data is diverse and derived from a large number of users, yet it may still limit the generalizability of our findings to some extent. Future research is needed to validate and extend our results in other settings. In addition, future work could also examine more fine-grained types of sharing, other than public, such as with family or friends only, or with communities [**?**], and study the differences in sharing behaviors among the various sharing types.

REFERENCES

[1] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of CHI*, 2007.

[2] M. Ames, D. Eckles, M. Naaman, M. Spasojevic, and N. Van House. Requirements for mobile photoware. *Personal and Ubiquitous Computing*, 14(2), 2010.

[3] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting ranking svm to document retrieval. In *SIGIR*, 2006.

[4] D. Carmel, A. Mejer, Y. Pinter, and I. Szpektor. Improving term weighting for community question answering search using syntactic analysis. In *CIKM*, 2014.

[5] A. Ceroni, V. Solachidis, C. Niederée, O. Papadopoulou, N. Kanhabua, and V. Mezaris. To keep or not to keep: An expectation-oriented photo selection method for personal photo collections. In *Proceedings of ICMR*, 2015.

[6] W.-T. Chu and C.-H. Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *Proceedings of MM*, 2008.

[7] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *TOMM*, 1(3), 2005.

[8] K. Crammer, A. Kulesza, and M. Dredze. Adaptive regularization of weight vectors. *MLJ*, 91(2):155–187, 2013.

[9] D. Frohlich, A. Kuchinsky, C. Pering, A. Don, and S. Ariss. Requirements for photoware. In *Proceedings of CSCW*, 2002.

[10] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of JCDL*, 2002.

[11] E. Guldogan, J. Kangas, and M. Gabbouj. Personalized representative image selection for shared photo albums. In *Proceedings of ICCAT*, 2013.

[12] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of MIR*, 2006.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[14] H. Li and X.-S. Hua. Melog: Mobile experience sharing through automatic multimedia blogging. In *Proceedings of MCMC*, 2010.

[15] J. Li, J. H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Proceedings of ICICS-PCM*, volume 3, 2003.

[16] E. Litt and E. Hargittai. Smile, snap, and share? a nuanced approach to privacy and online photo-sharing. *Poetics*, 42:1–21, 2014.

[17] A. D. Miller and W. K. Edwards. Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of CHI*, 2007.

[18] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of JCDL*, 2004.

[19] P. Obrador, R. De Oliveira, and N. Oliver. Supporting personal photo storytelling for social albums. In *Proceedings of MM*, 2010.

[20] J. C. Platt, M. Czerwinski, and B. A. Field. Phototoc: Automatic clustering for browsing personal photographs. In *Proceedings of ICICS-PCM*, 2003.

[21] F. Sadeghi, J. R. Tena, A. Farhadi, and L. Sigal. Learning to select and order vacation photographs. In *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015.

[22] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of ICMR*, 2011.

[23] S. Tschiatschek, R. K. Iyer, H. Wei, and J. A. Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Proceedings of NIPS*, 2014.

[24] N. Van House, M. Davis, M. Ames, M. Finn, and V. Viswanathan. The uses of personal networked digital imaging: an empirical study of cameraphone photos and sharing. In *CHI extended abstracts*, 2005.

[25] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, and G. W. Cottrell. Event-specific image importance. In *Proceedings of CVPR*, 2016.

[26] J. Yang, J. Luo, J. Yu, and T. S. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Transactions on Multimedia*, 14(6), 2012.

[27] L. Zhang. Describing human aesthetic perception by deeply-learned attributes from flickr. *CoRR*, abs/1605.07699, 2016.