Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR 2015), Shanghai, China, June 2015.

To Keep or not to Keep: An Expectation–oriented Photo Selection Method for Personal Photo Collections

Andrea Ceroni
L3S Research Center
Leibniz Universität Hannover
ceroni@L3S.de

Vassilios Solachidis
Information Technologies
Institute / CERTH
vsol@iti.gr

Claudia Niederée
L3S Research Center
Leibniz Universität Hannover
niederee@L3S.de

Olga Papadopoulou
Information Technologies
Institute / CERTH
olgapapa@iti.gr

Nattiya Kanhabua
L3S Research Center
Leibniz Universität Hannover
kanhabua@L3S.de

Vasileios Mezaris
Information Technologies
Institute / CERTH
bmezaris@iti.gr

ABSTRACT

When selecting important photos from a personal photo collection – e.g. for creating an enjoyable sub-collection for revisiting or preservation – photos are not considered in isolation. Therefore, collection-level criteria are also taken into account by automated photo selection methods. However, the typical two-step process of first clustering and subsequently picking from the clusters seems to overstress coverage as a criterion when applied to the task of selecting the photos most important to a user. We, therefore, propose a novel expectation-oriented photo selection method, which combines a variety of collection-level and image-level selection criteria in a flexible way. In our evaluation, which is based on large real-world personal photo collections with overall more than 18,000 images, we show that our method outperforms state-of-the-art photo selection methods. In addition, the proposed method does not rely on any manual annotations, making it applicable in realistic settings of personal photo collections.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection Process

Keywords

Photo Selection; User Expectations; Clustering; Coverage

1. INTRODUCTION

With digital photography and the many possible devices, photo taking is effortless and tolerated nearly everywhere. This makes us easily ending up with hundreds of photos, for example, when returning from a holiday trip. Furthermore, photos are also taken of more mundane motives, such as food or everyday scenarios, further increasing the number of photos to be dealt with. So, what to best do with all of these photos? With the decreased storage prices it is not a problem to store the photos somewhere. However, this often ends up as a kind of "dark archive" of photo collections, which are rarely accessed (and enjoyed) again. The mere size of the collection makes going through them as well as manual annotation and sorting of photos tedious tasks.

Furthermore, there is the risk of losing photos by a random form of "digital forgetting" [8]: over decades storage devices break down, and formats and storage media become obsolete, making random parts of photo collections inaccessible. Just consider, how difficult it would be today to access photos stored years ago in .mos format in a floppy disk.

Both the risk of dark archives and of digital forgetting suggest to select, supported by automated methods, the most important photos and to invest some effort into keeping them enjoyable and accessible. However, to *foster adoption*, such automated selection methods have to keep the level of user investment low. We do not rely on any additional user investment such as photo annotation with text [14, 17, 18] or eye tracking information [21], because we believe it is exactly the reluctance of further investment that lets large photo collections unattended on our hard disks.

When developing methods for semi-automatic photo selection, it is important to consider human expectations and practices. An important observation is that photo selection is a complex, partially subjective process, which does not consider images in isolation. Selection decisions also take the context of the other photos in the collection and of the photos already selected into account. Therefore, the aspect of coverage is used in a variety of photo selection methods [3, 10, 14]. In more detail, photo selection is modeled as a two-step process of first clustering the photo collection (for reflecting sub-events in the collection) and subsequently picking the most representative photos from the clusters. While coverage surely plays an important role for many photo selection tasks (see e.g. [21]), we believe that the complex decision making in selecting important and personal photos can be better modeled by avoiding the strict splitting into a two step process, which overstresses the role of coverage. We suggest to model a multifaceted notion of image importance driven by user expectations, which represents what photos users perceive as important and would have selected.

In this paper, we present an expectation-oriented method for photo selection, which relies on such a model of image

ICMR '15, June 23-26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749372.

importance. We aim at modeling user expectations considering information at both image- and collection-level, and we learn their different impact through a single model to predict importance. This information consists in (a) advanced concept detection (to capture the semantic content of images beyond aesthetic and quality indicators), (b) face detection (reflecting the importance of the presence of people in photos), (c) near-duplicate detection (to take the redundancy of many pictures of the same scene as a signal of importance, and to eliminate very similar images), (d) quality assessment (good quality photos might be preferred in case of comparable photos). This is complemented by (e) temporal event clustering and, more generally, collection-level information, to reflect the role of coverage in photo selection.

In summary, in this paper we make the following contributions: (i) we present a novel, effective, low-investment method for selecting important photos from personal photo collections, which is driven by user expectations; (ii) as a first work, we study the role of coverage in a systematic way by combining our expectation-oriented photo selection method with an explicit modeling of coverage in different ways, showing that comparable results to our method can be achieved only when coverage is not considered as a primary selection aspect; (iii) in our evaluation with real-world personal collections, we show that our method outperforms state-of-the-art methods to photo selection that rely on explicit modeling of coverage, when considering human selections as evaluation criterion.

2. RELATED WORK

Automated photo selection has already been studied in various other contexts, such as, photo summarization [10, 17, 18], identification of appealing photos based on quality and aesthetics [9, 23], selection of representative photos [3, 21], and the creation of photo books from social media content [14]. We consider the task of selecting important photos from personal collections (e.g. for revisiting or preservation), which meet user expectations. Among the works mentioned above, image importance has been considered only in [9, 23], nevertheless based on quality and aesthetic criteria.

Different photo selection and summarization works consider coverage by identifying clusters of images based on time and visual content [3, 10, 14]. Differently, our approach does not impose such a strict notion of coverage but rather considers clusters and other global information together with image-level information, learning their different impact in a single model. The works in [17, 18] are closer to ours, as they consider coverage in a relaxed way as part of a multi-goal optimization, but they still consider coverage as a key component and do not use user assessments in their evaluation. Moreover, they make partial use of manually created text to associate semantic descriptors to images, while our concept detection does not require any manual input, once the model has been learned.

Finally, Walber et al. [21] also use human judgments to evaluate selections, but the users have to wear eye trackers when using the system to make automatic selections.

Considering image processing, in our work we perform concept detection, temporal clustering, near–duplicate detection, face-detection, and quality assessment. Recent methods for concept detection (i) select specific locations on the image grid where features should be computed and extract at these locations local descriptors such as SIFT, SURF, and

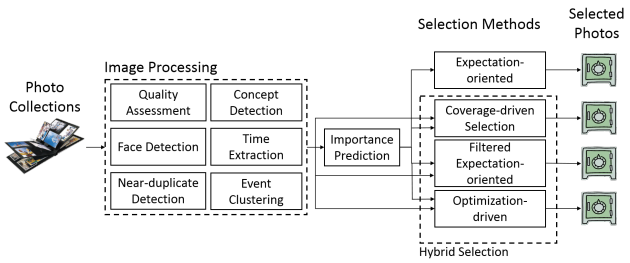

Figure 1: Approach overview of automatic photo selection.

others [19], (ii) build a global image representation from the local features using BoW, VLAD, Fisher vectors [2], (iii) use such representations of ground-truth-annotated training corpora to train concept detectors that rely on machine learning techniques. Many approaches have been proposed for near-duplicate detection, such as employing multi-resolution histograms [22] or aggregating local descriptors into global representations [7]. Since time is the dominant data dimension for sub-event clustering, several time-based image clustering methods have been presented in the literature [4, 6]. They can incorporate other data dimensions, e.g. geolocation information (if any) and visual information. For face detection, one of the most successfully approaches is the Haar-like-feature-based detector introduced in [20], whose modifications and extensions have been presented in [15]. Concerning Image Quality Assessment, a variety of no–reference techniques have been proposed for detecting quality degradations such as image blur, e.g. in [12].

3. PROBLEM DEFINITION

DEFINITION 1. *Let the photo collection P be a set of N photos, where $P = \{p_1, p_2, \ldots, p_N\}$. The photo selection problem is to select a subset S of size θ ($S \subset P$ and $|S| = \theta$), which is as close as possible to the subset S^* that the user would select as the photos most important to her, i.e. S meets user expectations.*

In our model, we represent each photo collection as a set $C = \{P, CL, ND\}$, where P is the set of original photos, and CL and ND are sets of clusters and near-duplicate photos identified in the collection, respectively. A cluster $cl \in CL$ contains a set of photos P_{cl} grouped together with respect to a *defined* notion of similarity, whereas a near-duplicate set $nd \in ND$ is a set of highly similar photos P_{nd}.

Each photo $p \in P$ is modeled as a set of features $p = \{\boldsymbol{q}, \boldsymbol{c}, F, t\}$, where $\boldsymbol{q} \in \mathbb{R}^{n_q}$ is the quality vector of the photo, $\boldsymbol{c} \in \mathbb{R}^{n_c}$ is the concept vector of the photo, F is the set of faces f appearing in the photo, t is its timestamp. Each face $f = \{f_l, f_s\}$ is described by its location f_l and relative size f_s in the photo. For each photo p, we will estimate the importance value I using the extracted features.

3.1 Approach Overview

The overview of our approach to photo selection is presented in Figure 1. Given a photo collection, we apply different image processing techniques (Section 4) in order to extract information used by our computational methods for automatic photo selection. Our main approach is the Expectation-oriented selection (Section 5), which learns to

generate selections by taking into account user selection patterns. In this approach, Importance Prediction is the core idea and it will be presented in Section 5.2. Furthermore, we present three different Hybrid Selection methods (Coverage-driven, Filtered Expectation-oriented, Optimization-driven), with the goal of investigating whether our method can be improved by combining it with state-of-the-art methods that explicitly consider coverage. The Hybrid Selection methods will be discussed in detail in Section 6.

4. IMAGE PROCESSING

The image processing techniques that we employ are concept detection, near–duplicate detection, image quality assessment, image clustering, and face detection.

Concept Detection. Concept Detection involves analyzing the visual content of an image and automatically assigning concept labels to it. This moves the description of an image to a semantic level, where it is possible to identify abstract concepts like *joy*, *cheering*, *entertainment*, as well as more concrete ones like *crowd*, *girl*, *stadium*. We trained 346 concept detectors for the 346 concepts defined as part of the TRECVID 2013 benchmarking activity [13]. As training corpus, the TRECVID 2013 dataset comprising 800 hours of video was used. We used SIFT, SURF, and ORB local descriptors and their color variants [11] for visual feature extraction. Then, PCA was applied on each descriptor for reducing their dimensionality to 80 and VLAD encoding [2] was applied to calculate the final image representation. The methodology that we followed has been introduced in [11].

Near-Duplicate Detection. Photographers can shoot a scene multiple times, which results in creating near–duplicate images that can be evidence of the importance of the scene. We used the method described in [1] to detect near-duplicates. This method detects at most 500 keypoints using the SIFT detector (Harris-Laplace) and extracts their corresponding descriptors using the SIFT extractor. It then forms a vocabulary by applying k-means on SIFT features using 192 cluster centers, and then encodes the image using VLAD encoding. The images with very similar VLAD vectors are efficiently retrieved using a KD forest index (using 5 KD–trees). Finally, Geometric Coding [24] is used to check geometric consistency of image's keypoints and to accept or reject the hypothesis that they are near-duplicates.

Image Quality Assessment. We computed four quality measures, namely blur, contrast, darkness, and noise, along with their aggregated value (weighted pooling using Minkowski metric), following the procedure presented in [12].

Temporal Event Clustering. In a photo collection captured by a single camera, time is a dominant dimension to reveal the sub-events that are represented in the collection. Thus, we apply time-based clustering in order to cluster collections into sub-events. Our method follows the approach presented in [4]. Images are sorted according to the time information extracted from them, and a similarity matrix of the ordered images is constructed using the time information and a sensitivity parameter. Then, for different values of the sensitivity parameter, the novelty scores and then their first derivatives are calculated. First derivatives which are greater than a threshold based on the maximum peak (at least 0.5 times greater than the maximum peak) are selected. The procedure results is a set of boundary lists (one list per sensitivity parameter). Finally, the confidence measure for each boundary list is calculated and the list that corresponds to the largest confidence measure is selected.

	TP	Tot	Precision
front_face_alt	467	576	81.08
front_face_alt_tree	449	544	82.54
front_face_alt2	482	627	76.87
front_face_def	486	733	66.30
merged	495	763	65.74
SCT	442	486	90.95
FE	471	653	72.13
all	439	490	89.59
SCT **and** *FE*	424	457	92.78
SCT **or** *all*	487	576	84.55
(*SCT* **and** *FE*) **or** *all*	484	564	85.82

Table 1: Performances of different face detectors.

Face Detection. We introduce a face detection approach which combines several face detectors to maximize the number of detected faces. We apply the approach in [20] using four pre-trained Haar Cascades detectors (frontalface_alt, frontalface_alt_tree, frontalface_alt2, frontalface_default) publicly provided by OpenCV[1]. In order to reduce the amount of false positive detections, we consider the detected faces as potential facial regions and, in every region, (i) we try to detect other facial characteristics (eyes, nose, and mouth) and (ii) we calculate the percentage of skin-like pixels. If facial characteristics are detected and are located in a valid location (e.g. eyes are centered and on the upper half for facial region) and the skin-like pixels percentage is above a threshold (set to 0.3 after experiments), then the detected region is classified as face. Furthermore, a facial region is accepted as a face if it has been detected by all face detectors (regardless of the existence of facial features or the ratio of skin-color pixels). This last case is effective in dark images where facial characteristics can not be detected and face color is too dark to be considered as skin-like.

This method was tested on a set of 484 images before applying it to the photos considered in this paper, and the results are listed in Table 1. As expected, if we merge the output of the four previously mentioned detectors (*merged*), the number of true positives increases but precision decreases. We experiment various constraints to decrease false detections: if the number of skin-color pixels are above a threshold (*SCT*), if at least one facial element (mouth, eyes, or nose) has been detected (*FE*), if the face has been detected by all detectors (*all*), and combinations of them (last three rows). Given these results, we chose the detector in the last row since it represents a good compromise between absolute number of true positives and precision.

5. EXPECTATION-ORIENTED SELECTION

The photo selection model presented in this section aims at meeting human expectations when selecting photos that are most important to the user from a collection, for instance for revisiting or preservation purposes. This is different from current approaches to photo selection for summarization, which aim at creating summaries that resemble the original collection as much as possible, either based on clustering [3, 10, 14] or explicitly considering coverage [17, 18].

We claim that selecting photos that are important to a user from personal collections is a different task than generating comprehensive summaries: the set of images impor-

[1] https://github.com/Itseez/opencv/tree/master/data/haarcascades

tant to the user might not be a proportioned subsample of the original collection. For instance, considering photos taken during a vacation, a user might ignore photos depicting joyless or boring moments. For this reason, we do not impose a strict notion of coverage but rather consider clusters and other global information as a set of features, along with photo-level features, learning their different impact in a single selection model. We then explicitly learn selection behaviors and preferences from real user data through a wide set of features. A characteristic of our features is that they do not require any manual annotation or external knowledge, differently from other works [14, 17, 18] that make partial use of manually created text associated to photos.

The features are combined via machine learning, providing a model that predicts the probability of a photo to be selected, i.e. its importance. The selected sub-collection is created by ranking photos in the collection based on their predicted importance and by taking the top-k of them, where k can assume any value lower than the collection size.

5.1 Features

Four groups of features have been designed to be used in the photo selection task, based on the information extracted from images as presented in Section 4.

Quality-based features. They consist of the 5 quality measures described before: blur, contrast, darkness, noise, and their fused value. The assumption is that users might tend to select good quality photos.

Face-based features. The presence and position of faces might be an indicator of importance and might influence the selection. We capture this by considering, for each photo, the number of faces within it as well as their positions and relative sizes. In more detail, we divided each photo in nine quadrants, and computed the number of faces and their size in each quadrant. This results in 19 features: two for number and size of faces in each quadrant, plus an aggregated one representing the total number of faces in the photo.

Concept-based features. The semantic content of photos, which we model in terms of concepts appearing in them, is expected to be a better indicator than low-level image features, because it is closer to what a picture encapsulates. We associate to each photo a vector of 346 elements, one for each concept, where the i-th value represents the probability for the i-th concept to appear in the photo.

Collection-based features. When users have to identify a subset of important photos, instead of just making decisions for each photo separately, the characteristics of the collection a photo belongs to might influence the overall selection of the subset. For the same reasons, but moving to a finer granularity, it might be worth considering information about the cluster a photo belongs to. For each photo, we consider the following collection-base features to describe the collection and cluster the photo belongs to: size of the collection, number of the clusters in the collection, number of near–duplicate sets in the collection, size of the near–duplicate sets (avg, std, max, min), quality of the collection (avg, std), faces in the collection (avg, std, max, min), size of the cluster (avg, std, max, min), quality of the cluster (avg, std, max, min), faces in the cluster (avg). Since the redundancy introduced by taking many pictures of the same scene can be evidence of its importance for the user, we also consider whether photos have near-duplicates or not, as well as how big is the near-duplicate set the photo belongs to.

5.2 Importance Prediction and Ranking

Once photos have been described in terms of the features presented above, a prediction model represented by a Support Vector Machine (SVM) [5] is learned to predict the selection probabilities of new unseen photos. Given a training set made of photos p_i, their corresponding feature vectors $\boldsymbol{f}_{p_i}$, and their selection labels l_{p_i} (i.e. *selected* or *not selected*), an SVM is trained and the learned model M is used to predict the importance of a new unseen photo p_{new}

$$I = M\left(\boldsymbol{f}_{p_{new}}\right) \tag{1}$$

i.e. its probability to be selected by the user. To avoid overfitting, the model was trained and evaluated via 10-fold cross validation over the collections and the generated output probabilities were considered in our evaluation.

Ranking. Once the importance of each photo is predicted, photos in the same collection are ranked based on this value and the top-k is finally selected. The parameter k represents the requested size of the selection and has to be specified in advance. It will be discussed during our evaluation (Section 7.1).

6. HYBRID SELECTION

As the evaluation will show (Section 7.3), our expectation-oriented selection clearly outperforms state-of-the-art methods for photo selection based on explicit modeling of coverage. However, we want to better understand the role of coverage in photo selection, in order to see if and in which way our method can be improved by combining it with explicit consideration of coverage. Therefore, we propose and investigate three ways of combining our importance prediction model with coverage-oriented photo selection methods, denoted *hybrid selection* methods and described hereafter.

6.1 Coverage–driven Selection

The coverage-driven selection is based on a two-step process of first clustering and subsequently picking photos from the clusters, which has been already used in other works. First, for a given collection C, a set of clusters CL_C is computed as described in Section 4 and the importance $I(p)$ of each photo $p \in P_C$ is given by our importance prediction model (Equation 1). Given the clusters CL_C, we use the importance $I(p)$ for each photo $p \in P_C$ to pick an equal number of top-ranked photos from each cluster in order to produce the selection S of required size k.

Cluster Visiting. One first issue to resolve in such approach is how to iterate over clusters when picking photos until the requested size of the selection is reached. After experimenting a number of alternatives, we identified a round-robin strategy with a greedy selection at each round as the best performing one. The pseudocode is listed in Algorithm 1. Given an initial set of candidate clusters CL_{cand}, the greedy strategy in each step selects the cluster cl^* containing the photo p^* with the highest importance, according to the prediction model M. The photo p^* is added to the selection S and removed from its cluster cl^*. The cluster cl^* is then removed from the set of candidate clusters for this iteration, and the greedy strategy is repeated until the candidate set is empty. Once it is, all the not empty clusters are considered available again and a new iteration of the cluster visiting starts. This procedure continues until the requested selection size k is reached.

Algorithm 1: Coverage–driven Selection (Greedy)

```
Input  : clusters CL, size k, prediction model M
Output: selection S
Set S = ∅
while |S| < k do
    Set CL_cand = CL
    while |CL_cand| > 0 do
        {cl*, p*} = get_most_important_cluster
        (CL_cand, M)
        S = S ∪ {p*}
        P_cl* = P_cl* − {p*}
        CL_cand = CL_cand − {cl*}
        if |cl*| = 0 then
            CL = CL − {cl*}
        end
    end
end
return S
```

Cluster Filtering. Intuitively, not all the clusters are equally important for the user. We tackle this issue by proposing a cluster filtering method to automatically predict the clusters that are not important for the user, in order to ignore them when picking photos from each cluster. We train a classifier (SVM) to detect and filter out clusters which are not important to the user. First, each cluster is described with the following features: size, quality vector (avg, std), average concept vector, number of faces (avg, std, min, max), number of near-duplicate sets and near-duplicate photos in it, near-duplicate sets size (avg, std, min, max), photo time (avg, std, min, max), photo importance (avg, std, min, max). The label associated to a cluster is *good* if it contains at least one selected photo, *bad* otherwise. Given a training set made of clusters c_i, their corresponding feature vectors $\boldsymbol{f}_{c_i}$, and their classes l_{c_i}, an SVM is trained and the learned model N is used to predict the class $L = N\left(\boldsymbol{f}_{c_{new}}\right)$ of new unseen clusters c_{new}. The model was trained via 10-fold cross validation over the collections.

6.2 Filtered Expectation-oriented Selection

The coverage-driven selection, coherently with all the selection methods based on clustering, is characterized by two steps: first clusters are identified and handled by possibly filtering and sorting them, and then photos in each cluster are ranked based on their predicted importance. Within the *filtered expectation-oriented selection*, we give priority to importance prediction by first ranking photos in a collection based on the predicted importance and then performing cluster filtering. The result is a ranked list of photos, where those belonging to clusters classified as *bad* have been removed. Note that the second phase of this paradigm, which contains cluster filtering in our case, can incorporate any other computation that exploits cluster information.

After the filtering, the selection S of size k is created by choosing the top–k photos in the list, as done in Section 5.2.

6.3 Optimization–driven Selection

Another more flexible way of explicitly incorporating coverage into a photo selection process is to consider it as part of a multi-goal optimization problem. This is done in [18] to generate representative summaries from personal photo collections. In more detail, in this work *quality*, *coverage*, and *diversity* of the summary are jointly optimized and the optimal summary S^* of a requested size k is defined as:

$$S^* = \underset{S \subset P_C}{\arg\max}\, F\left(Qual\left(S\right), Div\left(S\right), Cov\left(S, P_C\right)\right) \qquad (2)$$

where $Qual\left(S\right)$ determines the interestingness of the summary S and it aggregates the *interest* values of the individual photos in the summary, $Div\left(S\right)$ is an aggregated measure of the diversity of the summary measured as $Div\left(S\right) = \min_{p_i,p_j \in S, i \neq j} Dist\left(p_i, p_j\right)$, and $Cov\left(S, P_C\right)$ denotes the number of photos in the original collection C that are represented by the photos in the summary S in a concept space.

We incorporate our expectation-oriented selection within this framework, creating the *optimization–driven selection*, by computing the $Qual\left(\cdot\right)$ function in Equation 2 based on the importance prediction model (Equation 1), that is:

$$Qual\left(S\right) = \sum_{p \in S} M\left(p\right) \qquad (3)$$

Since part of the concepts in [18] are discrete categorical attributes, associated to photos using textual information and external knowledge bases not available in our task, we binarized the elements of our automatically detected concept vector (which includes the probability that a given concept appears in the photo) by using a threshold τ. The threshold has been empirically identified as $\tau = 0.4$ as the value that led to the most meaningful binary results. The rest of the computation of the $Div\left(\cdot\right)$ and $Cov\left(\cdot\right)$ functions in Equation 2 is performed as in the original work.

Regarding the resolution of Equation 2, we experimented the different approaches presented in [18] and we will report the best performing ones in the experimental analysis.

7. EXPERIMENTS AND RESULTS

In this section we evaluate the approaches presented in Sections 5 and 6, and we discuss their performances.

7.1 Experimental Setup

Dataset. For our experiments we use personal photo collections with importance judgments given by the owners of the collections as dataset. We decided to focus on personal collections because we wanted to observe the personal photo selection decisions in a setting that is as realistic as possible. This gives us a ground truth for assessing user expectations.

Given the unavailability of such a dataset, we performed a user study where participants were asked to provide their personal photo collections and to select the 20% that they perceive as the most important for revisiting or preservation purposes. The selection percentage (20%) has been empirically identified as a reasonable amount of representative photos, after a discussion with a subset of users before the study. We obtained 91 collections from 42 users, resulting in 18,147 photos. The collection sizes range between 100 and 625 photos, with an average of 199.4 (SD = 101.03).

Near-duplicates have been detected and filtered by considering the centroid of each set as representative photo, as done in [3]. Similarly to [16], each representative is marked as selected if at least one photo in its set has been marked as selected, and marked as not selected otherwise.

Evaluation Metrics. The selection methods presented in this paper can generate a selection S of size k from the original collection, where k can assume different values. We

evaluate the different methods considering the precision P@k of the selection S of size k that they produce, computed as the ratio between number of photos in S that were originally selected by the user and the size of S. Since the collections in our dataset have high size variability, absolute values of k, although traditionally used in IR tasks, would result in selecting very different relative portions of the collections depending on their sizes. This makes the impact of the selection different between collections. We, therefore, decided to express k as a percentage of the collection size, instead of an absolute value. In particular, we compute the precision for $k = 5\%, 10\%, 15\%, 20\%$, which are indicated as P@5%, P@10%, P@15%, P@20%, respectively. We concentrate the discussion on P@20%, because our ground truth was gathered by asking users to select the 20% of their collections.

Statistical significance was performed using a two-tailed paired t-test and is marked as $^{\blacktriangle}$ and $^{\triangle}$ for a significant improvement (with $p < 0.01$ and $p < 0.05$, respectively), and significant decrease with $^{\blacktriangledown}$ and $^{\triangledown}$ (for $p < 0.01$ and $p < 0.05$, respectively) with respect to the baselines. If not stated otherwise, the significance outcome reported in the tables always refers to the comparisons with both the baselines.

Parameter Settings. The classifiers employed in this paper for importance prediction and cluster filtering, built using the Support Vector Machine implementation of LibSVM[2], have Gaussian Kernels and have been trained via 10–fold cross validation. The open parameters were tuned via grid search to $C = 1.0$, $\gamma = 1.0$. The evaluation has been done using the predictions generated during the 10-fold cross validation, in order to separate training and test data.

7.2 Baselines

Two baselines are considered in our evaluation, one based on clustering and one representing the optimization framework presented in [18].

Clustering. Similarly to what was described at the beginning of Section 6.1, for a given collection C, a set of clusters CL_C is computed. The selection is built by iterating the clusters, temporally sorted, in a round–robin fashion and picking at each round the most important photo from the current cluster (until the requested selection size is reached). The importance of each photo $p \in P_C$ is modeled as $I(p) = \alpha \cdot \|\boldsymbol{q}_p\| + (1-\alpha) \cdot \dim(F_p)$, which is a weighted sum of the quality vector of the photo and the number of faces in it. This notion of photo importance covers different state-of-the-art works, such as [10, 14]. We experimented with different values of the parameter α, identifying the best value as $\alpha = 0.3$, which gives more importance to the number of faces in the photos. We report the performances obtained with this parameter value in our evaluation.

Summary Optimization. We implemented the approach presented in [18] as another baseline, where summaries are generated by optimizing *quality*, *coverage*, and *diversity* as in Equation 2. The *quality* of summaries is computed by summing the *interest* of photos in it, defined as a measure depending on photo quality and presence of portraits, groups, and panoramas. We computed the interest of photos as in the original work, using the concepts *face*, *3 or more people*, and *landscape* available in our concept set to represent portraits, groups, and panoramas respectively. Also *diversity* and *coverage* of summaries are computed coherently with their original computation, as already described in 6.3. We experimented the different approaches presented in [18] for the optimization of the cost functional (Equation 2), and the greedy optimization with equal weights for *quality*, *diversity*, and *coverage* was the one that achieved the best performances. Thus we will report the performances for this setup in the following evaluation, denoting it *SummOpt*.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

7.3 Results

The discussion of the results is organized as follows. First, we show the performances of our expectation-oriented selection with respect to the baselines (Section 7.3.1). Second, we present the results of the hybrid selection methods and we compare them both with the baselines (Section 7.3.2) and with the expectation-oriented selection (Section 7.3.3).

7.3.1 Expectation-oriented Selection

This section presents the evaluation of our expectation-oriented selection with respect to the two baselines. Different importance prediction models have been trained by using the subsets of the features described in Section 5.1, so that the impact of different groups of features on the precision can be analyzed. The results for different selection sizes (k) are listed in Table 2. The two baselines exhibit comparable performances, with *SummOpt* performing slightly better for all considered values of k (5%, 10%, 15%, 20%).

Regarding our model, *quality* features are the ones that perform weakest individually, which has already been observed for other photo selection tasks [21]. *Faces* features alone already show better performances than the baselines: the presence, number, and position of people in photos, largely used as one selection criterion in other works, is indeed a meaningful indicator of importance. The performance achieved when only using *concepts* features is better than the ones of *quality* and *faces*: they are able to capture the semantic content of the photos, going beyond their superficial aesthetic and quality. Examples of concepts with a high importance in the model are *person*, *joy*, *cheering*, *entertainment*, and *crowd*. The model trained with the combination of all aforementioned features, denoted *photo-level* because the features are extracted from photo level, slightly improves the performance of using concept features alone.

If we include global features for each photo representing information about the collection, the cluster, and the near–duplicate set the photo belongs to, we get a comprehensive set of features, which we call *all*. The precision of the selection for this global model further increases for every selection size: this reveals that decisions for single photos are also driven by considering general characteristics of the collection the photo belongs to: e.g. number of photos, clusters, average quality of photos in the collection and in the same cluster, how many duplicates for the photo there are. This is a point of distinction with respect to state-of-the-art methods (represented by the two baselines), because our selection approach does not strictly handle collection-level information by imposing clustering (*Clustering*) or optimizing measures like coverage and diversity along with photo importance only based on quality and presence of people (*SummOpt*). It rather takes this global information in consideration in a flexible way through a set of features, whose impact to the selection is learned from user selections and expectations. The expectation-oriented model using all the available features (named *Expo* in the rest of the evaluation) leads to an improvement of 38.5% and 33.75% over *Clus-*

	P@5%	P@10%	P@15%	P@20%
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Expectation-oriented Selection				
quality	0.3431	0.3261	0.3204	0.3168
faces	0.4506▲	0.3968▲	0.3836△	0.3747△
concepts	0.5464▲	0.4599▲	0.4257▲	0.4117▲
photo-level	0.5482▲	0.4760▲	0.4434▲	0.4266▲
all (Expo)	0.7124▲	0.5500▲	0.4895▲	0.4652▲

Table 2: Precision of the expectation-oriented selection, distinguishing different sets of features.

tering and *SummOpt* respectively, considering P@20%, and even higher improvements when considering smaller values of k (90.4% and 84.6% for P@5%).

Considering the trend of precision performances over different values of k, all the models reach higher precision values for smaller selection sizes. This can be due to the presence of a limited number of selected photos that are relatively easy to identify for the methods, which give them highest selection probability. Another reason is that there might be *noisy* selections originally done by the user, due to the fixed amount of photos required to select: if the user had selected less photos than requested, the selection done to reach the requested selection size might not really reflect her preferences and expectations.

Summarizing, modeling different promising aspects in terms of features and flexibly combining them through machine learning leads (except when using quality information alone) to consistent and statistically significant improvements over state-of-the-art summarization and selection methods.

7.3.2 Hybrid Selection

This section discusses the precision of the hybrid selection methods presented in Section 6 with respect to the baselines, along with a comparative analysis to assess the benefit of using cluster filtering and a greedy visiting strategy.

The results are listed in Table 3, where they have been split based on the three different classes of hybrid selection described in Section 6. For coverage-driven selection, we report results of different combinations: *basic* refers to the coverage–driven selection which only uses our importance prediction model defined in Section 5.2 as photo importance measure, picking photos in a round-robin fashion from clusters temporally ordered; the term *filtered* means the use of cluster filtering, while the presence of the term *greedy* indicates the use of the greedy visiting strategy. The filtered expectation-oriented selection is denoted *F-Expo*.

For the optimization-driven method, we experimented the different optimization methods described in [18] after introducing our importance prediction model in place of the original importance measure used in that work ($Qual\,(\cdot)$). We found out that the best performing method was still the greedy approach but with a parameter combination that gives more importance to the *quality* of the photos (0.6 *Qual*, 0.3 *Cov*, 0.1 *Div*), and we consider the results of this setup in the evaluation. This difference in weights with respect to the *SummOpt* baseline already anticipates that our expectation-based measure of importance has a bigger impact in the performances than the native quality measure defined in [18]. The method will be referred to as *SummOpt++*.

	P@5%	P@10%	P@15%	P@20%
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Coverage-driven Selection				
basic	0.4732▲	0.4113▲	0.3902△	0.3809△
filtered	0.5351▲	0.4617▲	0.4325▲	0.4170▲
filtered+greedy	0.6271▲	0.4835▲	0.4391▲	0.4262▲
F-Expo	0.7065▲	0.5502▲	0.4863▲	0.4600▲
SummOpt++	0.7115▲	0.5533▲	0.4937▲	0.4708▲
Expo	0.7124▲	0.5500▲	0.4895▲	0.4652▲
Filtering with Oracle				
greedy+oracle	0.6499▲	0.5107▲	0.4665▲	0.4484▲
F-Expo+oracle	0.7150▲	0.5606▲	0.4982▲	0.4753▲

Table 3: Precision of the hybrid selection methods.

The results in Table 3 show that all hybrid methods outperform the baselines, with statistical significance, revealing that the inclusion of the importance prediction model to assess photo importance has a strong impact compared to the baselines methods, which model photo importance with simple functions of quality and people occurrence. Similarly to the performances of the expectation-oriented models, both the absolute precision values and the improvements with respect to the baselines increase for decreasing k.

The results in Table 3 also show that cluster filtering increments the precision of the *basic* approach of an amount between 9.48% (P@20%) and 13.1% (P@10%). The greedy visiting strategy leads to improvements as well. Statistical significance tests revealed that the improvements introduced by *filtered* and *filtered+greedy* are statistically significant.

A comparative analysis between the hybrid selection methods shows that *F-Expo* and *SummOpt++* achieve better precision performances than the coverage-driven methods, and a t-test confirms that these improvements are statistically significant. This shows that the measure of photo importance modeled by our importance prediction has a bigger impact in the precision of the selection than coverage, and those methods that strictly model it through clustering (*coverage-driven selection*) get a smaller benefit when incorporating the expectation-oriented model. On the other side, methods that either give priority to expectations (*F-Expo*) or consider expectations, coverage, and global information in a flexible way via optimization (*SummOpt++*) can better exploit the expectation-oriented model.

7.3.3 Expectation vs. Hybrid Analysis

In this section we make a comparative analysis between the expectation-oriented selection model exploiting all the available features (*Expo*), and the hybrid selection models. Considering Table 3, we can observe that the performances of *Expo* are better or comparable with the ones of the hybrid-selection models. In particular, the improvements of *Expo* with respect to the *coverage-driven* methods are statistically significant. The only improvements over *Expo* (which anyway are not statistically significant) are obtained when considering methods that prioritize expectations (*F-Expo*) or possess a relaxed consideration of coverage and global information in general (*SummOpt++*). These results further support our assumption that for the photo selection task,

which we consider, a strong consideration of coverage overstresses this aspect as a selection criterion. Only for the methods with a more flexible consideration of coverage the performances are similar to the pure expectation-oriented method.

Cluster filtering is an attempt to eliminate clusters uninteresting to the user, and in order to further alleviate this aspect we conducted experiments considering only important clusters, i.e. those ones containing at least one selected photo. This is done by assuming to have a perfect classifier, i.e. an *oracle*, to filter out not important clusters and to focus the hybrid selection strategies only on the important ones. Although getting improvements compared to *filtered+greedy* and *F-Expo*, the performances when using such oracle, reported in the bottom part of Table 3, did not lead to consistent and statistically significant improvements with respect to *Expo*. *Greedy+oracle* does not beat *Expo*, while *F-Expo+oracle* only introduces a limited and not statistically significant improvement. These results show that the aspect that mostly drives user selections and expectations is the personal perception of importance, although this can produce unbalanced selections which are not representative of the original collection. Another problem related to clustering, even considering the important ones, might be the decision of how many photos to pick from each of them.

We also conducted experiments based on recall, which were in line with what was observed regarding the precision: both the expectation-oriented model and the hybrid selection methods outperformed the baselines, and the former was overall better than or comparable to the latter class.

8. CONCLUSION

In this paper, we presented an expectation-oriented method exploiting image as well as collection level features for helping the user in selecting the most important photos for creating an enjoyable sub-collection of a personal photo collection for preservation and revisiting. Our experiments with real world photo collections showed that our method outperforms state-of-the-art works in photo selection, which stress coverage as a selection criterion. Therefore, we investigated the role of coverage for our photo selection task in more detail, showing that coverage plays only a secondary role here.

Future works consist in incorporating more discriminative features to better predict user expectations, and designing models that optimize recall when generating selections.

Acknowledgments This work was partially funded by the European Commission in the context of the FP7 ICT project ForgetIT (under grant no: 600826).

9. REFERENCES

[1] K. Apostolidis, C. Papagiannopoulou, and V. Mezaris. CERTH at MediaEval 2014 synchronization of multi-user event media task. In *Proc. of MediaEval 2014 Workshop*, 2014.

[2] R. Arandjelovic and A. Zisserman. All about vlad. In *Proc. of CVPR '13*, 2013.

[3] W.-T. Chu and C.-H. Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *Proc. of MM '08*, 2008.

[4] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *In ACM TOMCCAP*, 2005.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[6] M.-S. Dao, A.-D. Duong, and F. G. De Natale. Unsupervised social media events clustering using user-centric parallel split-n-merge algorithms. In *Proc. of ICASSP '14*, 2014.

[7] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012.

[8] N. Kanhabua, C. Niederée, and W. Siberski. Towards concise preservation by managed forgetting: Research issues and case study. In *Proc. of iPres '13*, 2013.

[9] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proc. of MM '10*, 2010.

[10] J. Li, J. H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Proc. of ICICS-PCM '03*, 2003.

[11] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras. A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection. In *MultiMedia Modeling*, 2015.

[12] E. Mavridaki and V. Mezaris. No-reference blur assessment in natural images using fourier transform and spatial pyramids. In *Proc. of ICIP '14*, 2014.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2013*, 2013.

[14] M. Rabbath, P. Sandhaus, and S. Boll. Automatic creation of photo books from stories in social media. *ACM TOMM*, 2011.

[15] S. Roy and S. K. Bandyophadyay. Face detection using a hybrid approach that combines hsv and rgb. *IJCSMC*, 2013.

[16] A. E. Savakis, S. P. Etz, and A. C. P. Loui. Evaluation of image appeal in consumer photography. 2000.

[17] B.-S. Seah, S. S. Bhowmick, and A. Sun. Prism: Concept-preserving social image search results summarization. In *Proc. of SIGIR '14*, 2014.

[18] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proc. of ICMR '11*, 2011.

[19] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 2010.

[20] P. Viola and M. J. Jones. Robust real-time face detection. *Proc. of IJCV*, 2004.

[21] T. Walber, A. Scherp, and S. Staab. Smart photo selection: Interpret gaze as personal interest. In *Proc. of CHI '14*, 2014.

[22] M.-N. Wu, C.-C. Lin, and C.-C. Chang. Novel image copy detection with rotating tolerance. *Journal of Systems and Software*, 80, 2007.

[23] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *Proc. of MM '10*, 2010.

[24] W. Zhou, H. Li, Y. Lu, and Q. Tian. Large scale image search with geometric coding. In *MM '11*, 2011.