J. Vis. Commun. Image R. 74 (2021) 102973

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Sentiment-based sub-event segmentation and key photo selection ☆

Junghyun Bum [a], Joyce Jiyoung Whang [b,*], Hyunseung Choo [a,*]

[a] College of Computing, Sungkyunkwan University (SKKU), Suwon, South Korea
[b] School of Computing, KAIST, Daejeon, South Korea

ARTICLE INFO

Keywords:
Personal photo collection
Event segmentation
Key photo selection
Summarization
Sentiment analysis

ABSTRACT

The number of people collecting photos has surged owing to social media and cloud services in recent years. A typical approach to summarize a photo collection is dividing it into events and selecting key photos from each event. Despite the fact that a certain event comprises several sub-events, few studies have proposed sub-event segmentation. We propose the sentiment analysis-based photo summarization (SAPS) method, which automatically summarizes personal photo collections by utilizing metadata and visual sentiment features. For this purpose, we first cluster events using metadata of photos and then calculate the novelty scores to determine the sub-event boundaries. Next, we summarize the photo collections using a ranking algorithm that measures sentiment, emotion, and aesthetics. We evaluate the proposed method by applying it to the photo collections of six participants consisting of 5,480 photos in total. We observe that our sub-event segmentation based on sentiment features outperforms the existing baseline methods. Furthermore, the proposed method is also more effective in finding sub-event boundaries and key photos, because it focuses on detailed sentiment features instead of general content features.

1. Introduction

Due to the influence of social media platforms such as Instagram, Flickr, and Facebook, people are producing more media than ever before, and the number of people collecting photos, one of the most popular and important pieces of media, has increased [1]. According to [2], approximately 1.4 trillion photos are captured each year, owing to smartphones and similar hand-held devices. With regard to searching, browsing, and organizing personal photo collections, significant research has focused on event detection and representative photo selection. For example, Google Photos automatically identifies visual features and summarizes users' photos into three main categories: people, places, and things. Finding a specific set of events in a large photo collection, however, remains an open issue [3]. Meanwhile, as the number of photos stored on mobile devices rapidly grows, the need for automatic photo collection summarization significantly increases [4].

A personal photo collection is a record of sequential activities over a period. One of the popular approaches to summarizing a photo collection is dividing it into events and selecting key photos from each event. Since an event is closely related to a specific time and place, the time and global positioning system (GPS) information acquired from a photo's metadata can be used to distinguish one event from another. For instance, the time-based clustering method detects a significant time gap between two consecutive photos and segments the photos into an event. Recent approaches include probabilistic models and high-level visual features extracted from convolutional neural networks (CNNs) [5–7]. For example, [8] uses the hidden Markov model and Gaussian mixture model to recognize sub-events in photo sets with specific routines. Because personal photo collections have a wide variety of subjects, there is a limitation to finding an optimal solution through supervised learning. Consequently, few studies have proposed sub-event segmentation.

Photo collection summarization can be defined as the process in which the most important and meaningful moments of a photo collection are highlighted. For this purpose, it should eliminate redundancies and include aesthetically pleasing photos. We construct a summary by first dividing the segments into events or sub-events and then selecting the key photos within the segments. The primary goal of key photo selection algorithms is to find the most representative photos. For instance, although the simplest method is to select the middle photo in an event as the key photo, other algorithms select the photo with maximum a priori probability in an event as the representative photo. In recent studies, several ranking algorithms considering quality, popularity, and user-expectation, have been proposed [9–11]. However, key photos chosen by such algorithms still do not align with the ground truth because they do not consider the subjectivity of an individual such as sentiment and emotion [12]. In this paper, we propose the

☆ This paper has been recommended for acceptance by Xilin Chen.

* Corresponding authors.
E-mail addresses: bumjh@skku.edu (J. Bum), jjwhang@kaist.ac.kr (J.J. Whang), choo@skku.edu (H. Choo).

https://doi.org/10.1016/j.jvcir.2020.102973
Received 12 November 2019; Received in revised form 17 April 2020; Accepted 7 November 2020
Available online 12 November 2020

sentiment analysis-based photo summarization (SAPS) method, which automatically summarizes personal photo collections utilizing metadata and visual sentiment features. The proposed framework is composed as follows. First, we use a Gaussian mixture model to segment the events in the photo collection according to time and place. Second, we break down each event into sub-events based on content and sentiment features. We find the sub-event boundaries by computing novelty scores and quantifying the intra-cluster and inter-cluster similarities between two adjacent photo groups. Finally, we select key photos by using our ranking algorithm that takes into account sentiment, emotion, and aesthetics factors of photos. We evaluate the proposed method by applying it to photo collections of six participants consisting of 5,480 photos in total.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related research, and Section 3 describes the structure and key ideas of our sub-event summarization process. Section 4 discusses the experimental results, and the conclusion and future work are presented in Section 5.

2. Related work

An event is defined as a significant occurrence limited by time and an associated location [13]. For personal photo collections, an event, which includes one or more sub-events, can be defined as a photo-taking session where photo-worthy moments are recorded. Moreover, an event is mostly explicit, whereas a sub-event is either an intended or an unintended subset. Since photos in an event are often visually similar because they share aspects such as people, locations, and scenes [14], it becomes difficult to distinguish them from one another when the same event includes several different sub-events.

Event segmentation in a photo collection refers to setting the boundaries between different events on a timeline. In this regard, exchangeable image file format (Exif) metadata provides the timestamps for photos, which are helpful for indexing personal photo collections. Moreover, [15] has reported that organizing photos by time effectively improves users' performance in photo browsing and retrieval tasks. When an event occurs, a significant number of photos are taken in a relatively short time. Thus, the time-based clustering method detects large time gaps between two consecutive photos. PhotoToc [16] segments the boundary between two photo groups, especially when the time gaps are greater than the adaptive threshold, that is, the local average of the gaps between temporally adjacent photos.

Owing to the limitations of improper or missing timestamps, event segmentation methods using content-based features have also been proposed. For example, a recent study proposed a multi-modal generative model that utilizes time, location, and visual content features [17]. They extract high-level semantic features by taking advantage of CNNs trained in large-scale image collections. The photos in this case are represented by continuous feature vectors consisting of time, location, and deep learning features. It is assumed that all of these components are independent and that each component is generated by a single Gaussian distribution. Event segmentation is performed by computing the parameters that maximize the probability that each photo belongs to the corresponding event.

For tempo-spatial clustering, GPS information in the photo header is crucial; however, some smartphone users turn off their GPS. In such a case, spatial information can be obtained from the visual content of photos. Some pre-trained CNN models [18] are available at Places Database [19], a repository of 10 million scene photographs, labeled with scene semantic categories. The pre-trained models provide baselines for creating the novel scene recognition model on small-scale datasets [20].

In [21], the semantic regularized clustering method was used in a related study to represent photos as semantic visual concepts instead of CNN feature vectors. They first obtain a set of objects/tags/concepts detected in the photos, with their associated confidence values. Then, they reduce the number of semantic concepts to 100, by utilizing a semantic similarity graph and spectral clustering. Contextual event segmentation (CES) [22] proposes a method to detect event boundaries using an unsupervised learning method. CES is an LSTM-based generative model comprising a visual context predictor and an event boundary detector. The visual context predictor predicts the visual context of the upcoming photo, either in the past or future depending on the sequence ordering; the event boundary detector decides whether a photo is an event boundary by comparing the visual context generated from the photos in the past to that in the future.

State-of-the-art key photo selection algorithms rely on a combination of representativeness and aesthetics estimation. For instance, the PhotoToc selects key photos by measuring Kullback–Leibler (KL) divergence between the color histogram of every photo and the average color histogram of all photos. In this case, the photos with the highest number of uniquely colored regions are selected by the KL divergence metric [16]. In [23], they select the photo with maximum priori probability within an event. Shen et al. propose a ranking algorithm by combining the attributes of quality, representativeness and popularity [17]. For quality, they extract content features, such as brightness, color, the size of the object area, and contrast, and then evaluate the photos as aesthetically good or bad on the basis of the atomic visual action dataset [24]. Regarding representativeness, they compute the probability of a photo belonging to a specific event. Finally, popularity is measured by the number of similar photos in the same event. In travel photo album summarization, key photos are selected by considering aesthetic attributes, such as quality, memorability, and interestingness, based on the appearance of a central person or object [25]. Moreover, many state-of-the-art methods have utilized ranking algorithms in an attempt to imitate human selection. Li et al. proposes a new ranking algorithm that considers aesthetic qualities and memorable factors to find representative photos [11]. The aesthetic qualities contain the area and location of the salient region and sharpness of the photo; the memorable factors include the salient people and text information.

Sentiment analysis has been increasingly performed to understand human decision-making in areas such as brand marketing, customer satisfaction, and political forecasting. However, this approach has mainly focused on textual contents, instead of visual ones. In general, the semantic concept of a photo, which is associated with the actual presence of an object or scene, can vary from person to person, since the stimuli that trigger human responses are subjective. Recently, joint visual–textual sentiment analysis with deep neural networks has been proposed for online user-generated contents such as Twitter and Instagram [26]. In another study, an affective image classification algorithm jointly utilized visual features and semantic image annotations such as the categories of certain objects and scenes [27]. Furthermore, a study of visual sentiment analysis called, 'Visual Sentiment Ontology,' automatically analyzes sentiment by using mid-level representation of visual contents [28]. In this case, they apply Plutchik's Wheel of Emotions, which is a well-established psychological model that defines emotional keywords and extracts adjective noun pairs (ANPs) such as "beautiful flower" and "cute dog" from photos. This framework not only consists of more than 3,000 semantic concepts but also provides Sentibank, a library of trained concept detectors, with 1,200 ANPs. In this study, we take advantage of the visual sentiment features of photos based on the Sentibank classifiers.

3. The sentiment analysis-based photo summarization (SAPS)

An overview of our proposed method is presented in Fig. 1. First, the context information of each photo is extracted and the photos are segmented using a Gaussian mixture model in Phase 1. That is, the events are grouped into large segments through tempo-spatial clustering. Second, in Phase 2, each event is divided into sub-events through sentiment-based sub-event segmentation. Third, using our ranking algorithm, key photos are selected based on sentiment, emotion and aesthetics. The following subsections discuss this summarization method and key photo selection algorithm in detail.

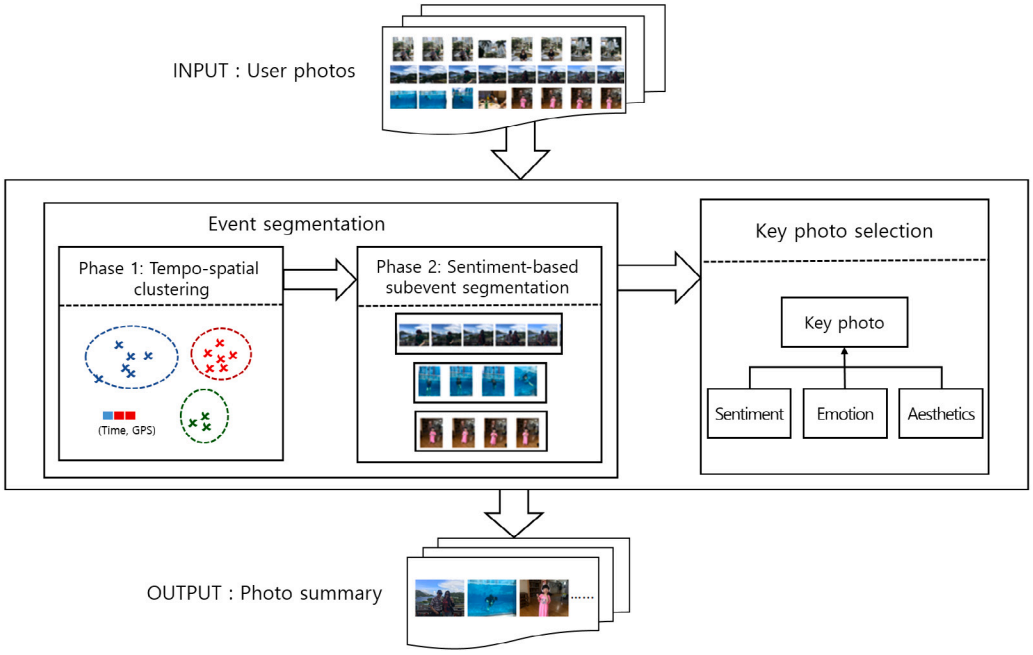


Fig. 1. An illustration of the sentiment analysis-based photo summarization (SAPS) method.

3.1. Features

An event is naturally associated with a specific time and place. Thus, photos are represented by a three-dimensional feature vector of time and location (latitude, longitude) for event segmentation. That is, given $X = \{x_1, x_2, \ldots, x_N\}$ of N photos, each photo can be represented as $x_i = (t_i, lat_i, long_i)$, where t_i is the timestamp and $(lat_i, long_i)$ are the geo-coordinates defined by latitude and longitude.

First, for each photo, Exif headers are processed to extract the timestamps, whose units are denoted in seconds. Then N photos in the collection are sorted according to time. Meanwhile, the latitude and longitude stamps of the Exif headers provide the GPS data. In general, majority of the photos from mobile devices include date/time stamps and GPS data. However, some photos may not contain location information in their headers. In this case, we set them to be the same as the nearest photo in time. We also use $L1$ normalized timestamps and GPS data.

We utilize spatial information from the visual content of photos in the case where GPS data are unavailable. We take advantage of the VGG16 architecture [29], a CNN model pre-trained with Places Database for scene recognition. This classifier provides 365 scene semantic categories. We divide the photo collection into large segments of tempo-spatial dimensions in Phase 1; however, time feature can be diluted by the 365-dimensional spatial information. Therefore, our model is fine-tuned to a smaller number of semantic categories, i.e., five classes. The five categories consist of two indoor (house and restaurant) and three outdoor scenes (street, building, and nature). Transfer learning is a technique that transfers knowledge to a new model by utilizing the initial weights from the pre-trained model. This method can accelerate training and improve accuracy. We focus on the parameter transfer of the pre-trained model and then redesign the structure of the transfer model for scene recognition. Finally, we assign spatial features via one-hot vector encoding from the predictions of our modified CNN model.

Second, we extract content features, such as color and sentiment, to refine the events into sub-events. For this purpose, we adopt the RGB color histogram to represent the visual contents of the photos. In other words, for a color image x with n pixels, a 3 x 256-dimensional color histogram $H(x) = [H_r, H_g, H_b]$; $H_a(x) = (h_1, h_2, \ldots, h_{256})$ with $a \in \{r, g, b\}$ in RGB color space is computed, where $h_i = n_i/n$ is the proportion of pixels whose value belongs to each color bin i. The color histogram of photos is commonly used to compare similarities in visual contents.

Third, we adopt sentiment features to split events that occur at adjacent times and places into sub-events. In this case, the pre-trained Sentibank detectors [28] provide a mid-level representation of ANP terms for a given image. The value of the detectors is equal to the likelihood value for the 1,200 ANP concepts. In other words, given the photo image x, it returns the confidence value associated with the detected ANP concepts in the photo. The confidence values for each concept form a sentiment feature vector to be used in the photo collection. These values are multiplied by the pre-defined sentiment scores for each concept to predict the sentiments for the photos, such as positive, negative and neutral. Then, the content features are fused with sentiment features to represent each photo.

3.2. Event segmentation

In Phase 1, the events are separated into large segments through tempo-spatial clustering, while in Phase 2, the contents and sentiment features are used to split the events into smaller sub-events. This two-phase process prevents event segmentation from being biased, owing to missing data such as GPS data. For example, if the contents (e.g., person and background scene) of a photo captured at completely different places are the same when there is no GPS data, clustering algorithms may group the photos in the same cluster. Even when GPS data are available, the large scale of content features may overwhelm the two-dimensional feature vectors indicating location. Conversely, the two-phase process is advantageous because one large cluster formed with the missing GPS data in Phase 1 can be segmented in Phase 2.

Fig. 2 presents an example of the number of photos taken at each month during one year. In this case, each event corresponds to one peak in the distribution. In general, it is difficult to determine the number of events in clustering applications. Fortunately, we can estimate the number of events from the distribution of the number of photos. More specifically, since the date on which many photos are captured corresponds to one event, an initial estimate is set to be the number of peaks in the timeline. We set the global average number of photos

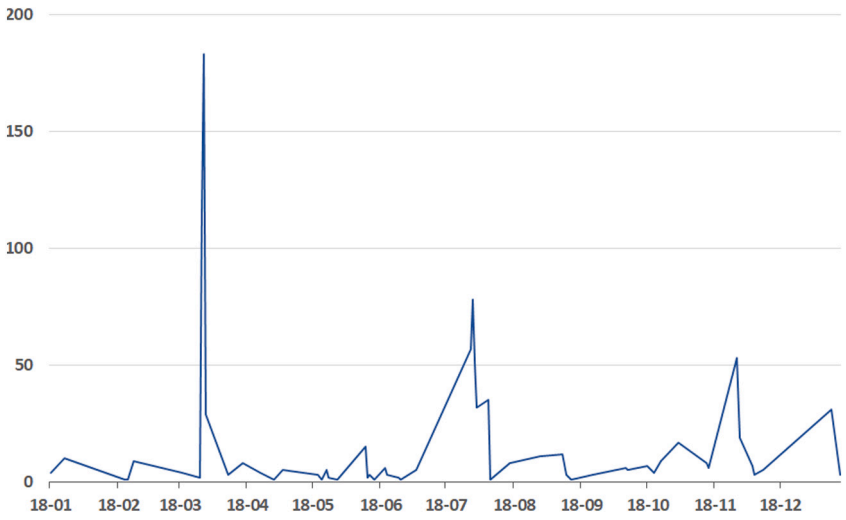


Fig. 2. An example of the number of photos taken at each month during one year.

captured in a day as a threshold, and count the number of peaks that are above the threshold to determine the number of events. Notably, this number can vary depending on the granularity dividing events. Alternatively, a user can specify an initial number for k. For instance, when a user is interested in the top 10 events, then k is set to 10.

Furthermore, we investigate the Davies–Bouldin (DB) index to determine the optimal number of events [30]. DB index calculates the ratio of the within-cluster to between-cluster distances. If clusters are farther apart and less dispersed, the DB index gets the lower score. We set the optimal k to the number of clusters given the lowest DB index.

We assume that the photos in the same event share the same distribution of tempo-spatial features, that is, they are close in both time and location. In this case, each photo $x_i \in X = \{x_1, x_2, \ldots, x_N\}$ corresponds to an unobserved semantic concept class, that is, an event $e_j \in E = \{e_1, e_2, \ldots, e_K\}$, where N and K are the total number of photos and events in the photo collection X, respectively. Thus, the probability of photo x_i generated from event e_j can be formulated as $p(x_i|e_j)$. In our model, each photo is represented by a 3-dimensional feature vector $x_i = (t_i, lat_i, long_i) = (x_{i1}, x_{i2}, x_{i3})$. To simplify the model, three components related to photo x_i are assumed conditionally independent, given the hidden concept event e_j. In (1), a priori probability $p(x_i|e_j)$ can be defined as follows:

$$p(x_i|e_j) = p(t_i|e_j)p(lat_i|e_j)p(long_i|e_j) = \prod_{l=1}^{3} p(x_{il}|e_j), \tag{1}$$

where $x_i = (x_{i1}, x_{i2}, x_{i3})$ and each component x_{il} is generated by a single Gaussian distribution, as shown in (2):

$$p(x_{il}|e_j) = \frac{1}{\sqrt{2\pi\sigma_{il,j}^2}} e^{-\frac{(x_{il,j} - \mu_{il,j})^2}{2\sigma_{il,j}^2}}. \tag{2}$$

When the header of whole photos does not contain GPS data, we utilize 5-dimensional spatial features obtained from our modified CNN model for scene recognition. Here, a priori probability $p(x_i|e_j)$ can be defined as follows:

$$p(x_i|e_j) = \prod_{l=1}^{6} p(x_{il}|e_j), \tag{3}$$

where x_{i1} is time feature and $x_{i2}, \cdots, x_{i6}$ contain spatial information.

Meanwhile, the model parameters can be estimated by maximizing the log-likelihood of joint distribution. The objective function is formulated as

$$\mathcal{L}(X;\theta) \triangleq \log\left(\prod_{i=1}^{N} p(x_i|\theta)\right) = \sum_{i=1}^{N} log\left(\sum_{j=1}^{K} p(e_j)p(x_i|e_j,\theta)\right). \tag{4}$$

where $p(x_i|e_j,\theta)$ is computed according to (1) or (3), with θ given. $p(e_j)$ is the priori probability of event e_j. We adopt the expectation–maximization(EM) algorithm to tune parameters as shown in Algorithm 1. We also describe the entire process of Phase 1 in Algorithm 2.

Algorithm 1 EM algorithm

Step E

Compute the likelihood by (4)

Step M

1. Update posterior of event e_j by Bayes rule:
   $p(e_j|x_i)^{n+1} = \frac{p(e_j)^n p(x_j|e_j)^n}{\sum_{j=1}^{I} p(e_j)^n p(x_i|e_j)^n}$
2. Update model parameters of event e_j :
   $\mu_{il,j}^{n+1} = \frac{\sum_{i=1}^{N} p(e_j|x_i)^{n+1} x_{il}}{\sum_{i=1}^{N} p(e_j|x_i)^{n+1}}$
   $\sigma_{il,j}^{n+1} = \frac{\sum_{i=1}^{N} p(e_j|x_i)^{n+1} (x_{il} - \mu_{il}^{n+1})^2}{\sum_{i=1}^{N} p(e_j|x_i)^{n+1}}$
3. Update model of event e_j :
   $p(e_j)^{n+1} \approx \frac{1}{N} \sum_{i=1}^{N} p(e_j|x_i)^{n+1}$
   $p(x_i|e_j)^{n+1} = \prod_{l=1}^{m} p(x_{il}|e_j)^{n+1}$

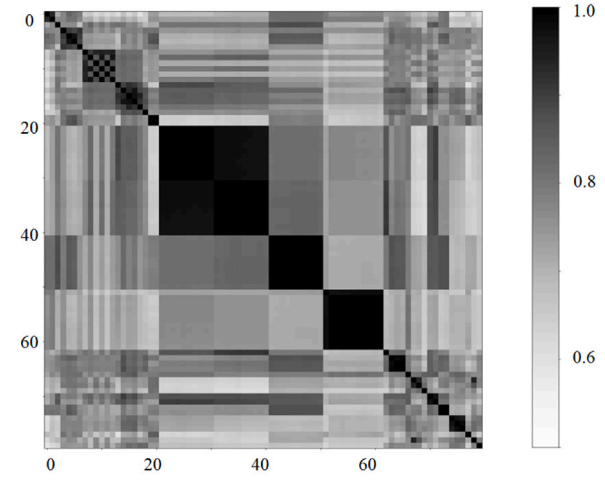
Algorithm 2 Tempo-spatial clustering algorithm

1. Perform metadata extraction and normalize data
2. Set K by performing the model selection using DB index
3. Initialize model parameters θ by K-means
4. For event $j = 1$ to K do
   Update model parameters θ by using EM algorithm
5. Assign photos to the corresponding events based on $p(e_j|x_i)$

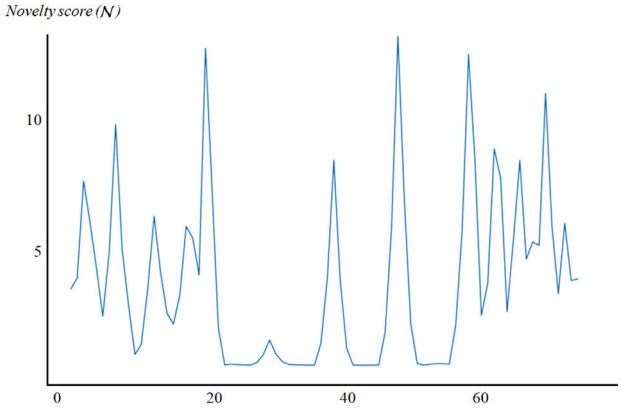
3.3. Sub-event segmentation

A personal photo collection usually includes a combination of photos from intended and unintended events in everyday life. Noisy data, such as screenshots can render sub-event segmentation difficult. Moreover, although tempo-spatial data can distinguish a large segment of events, there is not enough information to split sub-events within an event.

Therefore, we analyze the visual content of the image and perform sub-event segmentation. The visual content of photos can be represented in various ways, such as scale invariant feature transform descriptors, speeded-up robust features descriptors, and CNN features used for object detection. As mentioned in the previous section, we assume that photos within the same sub-event share sentiment similarity. According to [31], the color of an image is highly correlated with its sentiment. Because a visual color induces a psychological



(a) A visualization of similarity matrix S



(b) A graph of novelty scores. The x-axis is the i^{th} photo and the y-axis represents the novelty score of the corresponding photo.

Fig. 3. Similarity matrix displaying event boundaries and corresponding novelty scores.

association with a cold or warm sentiment, we use color histograms of photos as features for sub-event segmentation. Moreover, we obtain sentiment features from the Sentibank classifier of VSO, a framework that performs automatic sentiment analyses. This classifier provides the probability of 1,200 ANP terms such as "little puppy" and "clear sky", and we use these results as sentiment features. In this phase of sentiment-based sub-event segmentation, the photos are represented by a 1968-dimensional feature vector fusing color and sentiment features. In other words, given $X = \{x_1, x_2, \ldots, x_N\}$ of N photos, each photo is represented by $x_i = (h_{i1}, \ldots, h_{i768}, s_{i1}, \ldots, s_{i1200})$, where $(h_{i1}, \ldots, h_{i768})$ is the color histogram and $(s_{i1}, \ldots, s_{i1200})$ is the sentiment feature.

We also distinguish the boundaries between the sub-events using a similarity-based algorithm. We measure the similarity between photo i and j by computing the cosine similarity between feature vectors x_i and x_j. The larger the value, the more similar the visual contents of the given photos. The cosine similarity between the two vectors is defined by (5):

$$s_{ij} = \frac{x_i \cdot x_j}{\| x_i \| \| x_j \|}. \tag{5}$$

A similarity matrix S contains the similarity measure s_{ij} for all photos in the event. Generally, S is a symmetric matrix with a value of 1 in the main diagonal. S can also be visualized as a square image, as shown in Fig. 3. Each pixel i, j is shown in a gray scale proportional to the value of similarity measure s_{ij}.

The areas with high image similarity appear as dark squares in the diagonal. Due to the cosine metric, similar photos have values close to 1, while dissimilar photos have values close to −1. Determining the boundaries of the sub-events is as simple as detecting the borders of a checkerboard. To find the boundaries of sub-events in a group of similar photos, we move along the main diagonal of S and calculate the novelty score, following [32]. In this case, the novelty scores quantify self-similarity and cross-dissimilarity. Next, we correlate a Gaussian-tapered checkerboard kernel, denoted as $\mathcal{G}$, and calculate the novelty scores as shown in (6):

$$\mathcal{N}(i) = \sum_{p,q=-L/2}^{L/2-1} S(i+p, i+q) \cdot \mathcal{G}(p, q), \tag{6}$$

where L is the lag of kernel (window size).

Since the lag of kernel L directly affects the properties of the novelty measure, we use a small kernel (i.e., $L = 8$) to detect boundaries of the sub-events. In addition, we compute the interior region in which the kernel completely overlaps matrix S. We note that when the photos adjacent to photo i are similar, the score is high. If the photos non-contiguous to photo i (the white region of kernel) are similar, the score is low. Contrarily, when they are dissimilar, the score is high. Because we search for photo i that is similar to the adjacent photos but dissimilar to non-contiguous photos, the higher the score, the better the boundary candidate. That is, photo i becomes the sub-event boundary when it acquires the highest score. Finally, determining the sub-event boundaries is a matter of detecting the peaks in the novelty scores as shown in Fig. 3b. A simple approach is determining the peaks where the score exceeds a local threshold. In this case, we set a threshold to a value exceeding the local average from each group of events. Algorithm 3 describes the detailed steps to determine the sub-event boundaries.

Algorithm 3 Sentiment-based sub-event segmentation

1. Extract sentiment-related features
2. Sort the photos according to their timestamps
3. For each segment 1 to K do
   (a) Compute the similarity matrix S using (5)
   (b) Compute the novelty score $\mathcal{N}$ using (6)
   (c) Calculate the local average value (threshold) of $n \in \mathcal{N}$
   (d) Detect the peak in n by the hill-climbing method
   (e) Store photo i if the value of detected peak $\mathcal{N}(i) >$ threshold
4. Form the event boundary list

3.4. Photo filtering

The purpose of summarizing a personal photo collection is two-fold: (1) to cover all aspects of the entire collection, and (2) to summarize them as aesthetically pleasing and meaningful to users. This section describes our algorithm used to achieve the second goal. Ceroni et al.'s work [12] found that when selecting photos from personal photo collections for long-term preservation, the criteria "memory evocation" and "important to me" were rated high, while the objective quality of the photos was the second least important criterion. Since the aforementioned criteria contain a high level of subjectivity, it is difficult to create automatic selection measures. Thus, we use sentiment and emotion as alternative measures for imitating the human selection process.

A pre-processing step of selecting key photos is to filter out low-quality photos (e.g., blurry or dark photos), since the basic factors that determine the quality of a photo are brightness and sharpness. The brightness distribution of a photo plays an important role in determining the quality. In this regard, we measure the distance between the brightness histogram and a uniform histogram using KL divergence, after which the distance is used as the brightness score. Because a well-distributed brightness histogram is similar to the uniform function, KL divergence is relatively low. Fig. 4, shows the brightness distribution in our photo collection. To automatically identify high-quality photos, we only select photos with a brightness score of < 0.55.

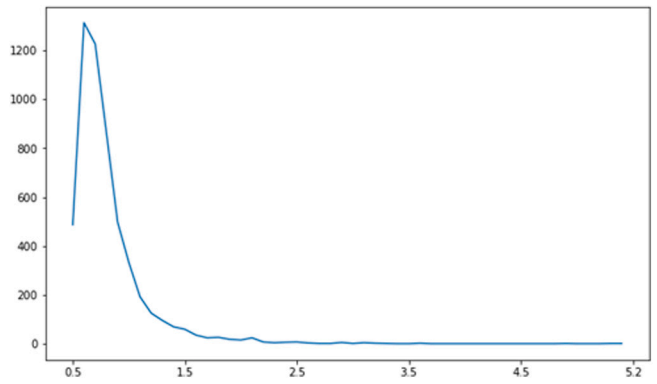


Fig. 4. KL divergence score distribution of our photo collections. A lower score implies a well-distributed brightness histogram.

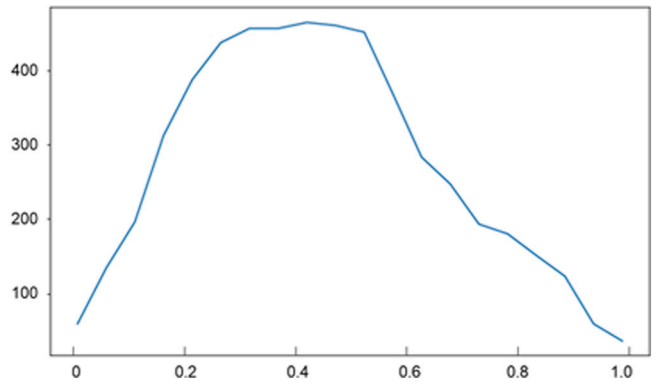


Fig. 5. Blur score distribution of our photo collections. A lower score implies a sharper photo.

Table 1
Attributes of the key photo selection algorithm.

Attributes	Description
Sentiment(Q_1)	Each photo receives a value in the range of −2 to +2, by applying visual sentiment prediction. The larger the value, the higher the positive sentiment.
Emotion(Q_2)	Emotion scores are extracted from each photo ($0 \le f_i \le 1, i = 1, 2, \ldots, 7$). The maximum emotion score for each face detected in the photo is used.
Aesthetics(Q_3)	The placement that respects the "rule of thirds" receives a high score.

Following Tong et al. [33], we also remove blurry photos by employing edge type and sharpness analysis with a Haar wavelet transform process, in which the lower the blur score (%), the sharper the photo. As shown in Fig. 5, we select only the photos with blur score distribution of < 0.63. These parameters are empirically selected to exclude long-tail of distributions.

3.5. Key photo selection

For the key photo selection algorithm in our proposed SAPS method, we assume that users prefer photos with a positive sentiment. Further, we note that photos with high emotion scores are likely to be people-centered photos. Regarding aesthetic quality, we use the "rule of thirds", which places important objects along the imagery lines both vertical and horizontal [34]. The evaluation method of quantifying the quality of attributes is shown in Table 1.

Regarding sentiment attributes, the ANPs (based on the Sentibank classifier) include pre-defined scores ranging from −2 to +2 [28]. A negative score represents negative sentiment (e.g., "crazy fire" or "dead skull"), whereas a positive score represents positive sentiment (e.g., "sunny sky" or "happy baby"). We also use sentiment scores based on the top 100 confidence values out of the 1,200 ANPs. Moreover, the values of sentiment attributes are determined by the average ANP confidence values of the photos, multiplied by the pre-defined sentiment scores. For example, when the confidence values of the top 100 ANPs for photo x_i is c and the sentiment score of the corresponding ANPs is v, the sentiment prediction score Q_1 is calculated as follows:

$$Q_1(x_i) = \frac{\sum_{t=1}^{T} c_t \cdot v_t}{T}, \tag{7}$$

where $T = 100$. Fig. 6 presents examples of the sentiment attribute scores, the top 10 ANPs, and their corresponding sentiment scores and confidence values.

According to [35], people prefer photos with facial expressions and emotions, and this has been used in various applications focusing on face recognition and emotion detection. Thus, we use a real-time emotion classification application that trained on FER-2013 emotion datasets, with a Keras CNN model and open source computer vision library(OpenCV) [36]. This library provides a confusion matrix for the seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. In this case, the confidence value ranges from 0 and 1. Meanwhile, since the emotion score is estimated from the facial expression in the photo, a photo receives a value of 0 when a face is not recognized. In other words, emotion measurement is based on whether a person is included in a photo. However, unlike sentiment attributes, it is difficult to differentiate the seven emotions into positive and negative. Based on the assumption that emotionally well-expressed photos are more interesting to users, we take the maximum value among the seven emotions as an indicator. The emotion metric Q_2 is obtained as follows:

$$Q_2(x_i) = \frac{\sum_{m=1}^{P} \max(f_m)}{P}, \tag{8}$$

where P is the number of persons in the photo and f_m indicates the confidence values of the seven emotions.

The "rule of thirds" is one of the most well-known composition rules used by photographers to create high esthetical photos. In particular, this guideline divides a photo into nine equal parts, with two equally spaced horizontal and vertical lines, respectively, after which important compositional elements are placed along these lines or at their intersections. Object detection requires semantic content understanding. Alternatively, we simply find the contours of objects in a photo and consider the largest contour as the most important object. This algorithm also computes aesthetic metric Q_3, by applying the same formula as [25]. A photo having a size of r by s pixels is defined as follows:

$$Q_3(x_i) = 1 - \left(\frac{d}{1/2\sqrt{(r/3)^2 + (s/3)^2}} \right)^2, \tag{9}$$

where d is the distance between the center of the key object and the nearest intersection.

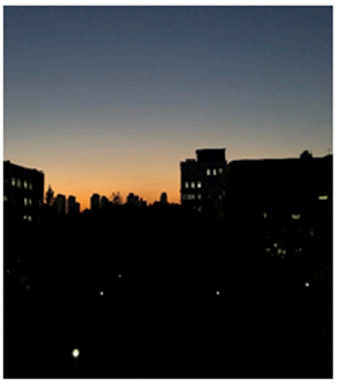
Finally, we prioritize photos with positive sentiment and large facial expressions to select key photos within a sub-event. Moreover, we choose high quality photos by considering the composition rule. The weights can be adjusted according to the user's preference. The final score is computed by combining all three values, as shown in (10):

$$Score(x_i) = \sum_{z=1}^{3} w_z \times Q_z(x_i), \tag{10}$$

where $w_z = 1/3, z \in \{1, 2, 3\}$ (weight).

Top 10 ANPs	Score	Confidence
little_puppy	2	0.98
gorgeous_eyes	1.47	0.98
fluffy_kitty	1.89	0.98
shy_cat	-1.56	0.97
sleepy_puppy	0.7	0.97
silly_face	1.53	0.97
sweet_face	1.53	0.97
smiling_baby	2	0.97
shy_dog	-1.55	0.97
smooth_skin	1.51	0.96

Sentiment_score = 0.82



Top 10 ANPs	Score	Confidence
dusty_sky	-1.83	0.99
clear_night	1.59	0.98
beautiful_night	1.59	0.98
bright_moon	1.71	0.98
dark_night	-1.41	0.97
shiny_moon	1.71	0.97
wet_sand	-1.57	0.97
bright_sea	1.6	0.95
tiny_raindrops	0.62	0.95
bright_city	1.65	0.94

Sentiment_score = 0.14

Fig. 6. The top 10 ANPs along with their scores and confidence values.

4. Experiments

4.1. Experimental setup

To compare the proposed SAPS method with the baseline methods in [16,17,21], and [22] six participants were asked to share photos taken through smartphones during the past year. Although all photos included timestamps, some photos had missing GPS data. Overall, the experiments were conducted in three steps. First, we asked participants to segment all their photos into meaningful events. Second, we asked them to divide each event into smaller sub-events, without providing special guidelines for sub-event segmentation. Finally, we asked them to select one or more key photos within each sub-event. The human-labeled sub-event boundaries and key photos are considered to be the ground truth. Table 2 presents the datasets used in our experiments.

We compute the following precision, recall and F-score metrics (described in [17]) to evaluate the sub-event boundary detection and key photo selection processes:

$$precision_{seg} = \frac{correctly\ detected\ boundaries}{total\ number\ of\ detected\ boundaries} \tag{11}$$

$$recall_{seg} = \frac{correctly\ detected\ boundaries}{total\ number\ of\ ground\ truth\ boundaries} \tag{12}$$

$$F-score = \frac{2 \times precision \times recall}{precision + recall} \tag{13}$$

$$precision_{select} = \frac{correctly\ selected\ key\ photos}{total\ number\ of\ selected\ photos} \tag{14}$$

Table 2
Information about our datasets.

Dataset	No. of photos	No. of events	No. of sub-events (ground truth)	avg. photo of sub-events
User 1	947	32	85	11
User 2	498	21	77	6
User 3	1,884	45	189	9
User 4	802	20	73	10
User 5	601	24	47	13
User 6	753	17	67	11

Based on the aforementioned equations, precision represents the ratio of correctly detected boundaries to the total number of detected boundaries, while recall indicates the ratio of correctly detected boundaries out of the total number of ground truth boundaries. Moreover, the F-score measures comprehensive performance as a harmonic mean value of precision and recall. Similarly, the precision of key photo selection is computed by the proportion of correctly selected key photos among the total number of selected photos.

We perform transfer learning with VGG16 CNN model pre-trained on the places365-standard for scene classification and use the Keras deep learning framework and applications API [29]. We designed the model by reducing the output of the two fully-connected layers from 4,096 to 1,024 and 512 and the output of last softmax layer to 5 to decrease the training complexity. We retrained the fully-connected layer and the fifth block of the convolution layers (Fig. 7). The dataset used for training is part of the places365-standard, as shown in Fig. 9.

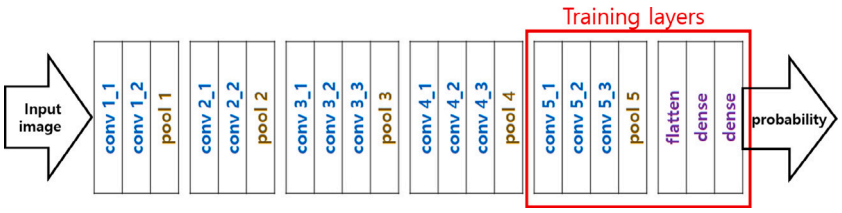
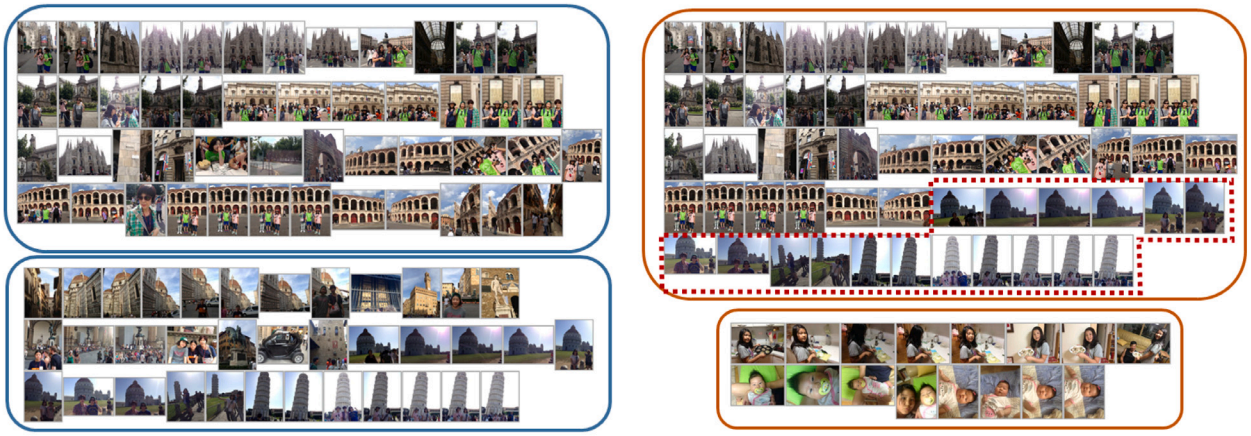


Fig. 7. A schematic diagram of the VGG-16 deep convolutional neural network architecture for our base CNN model.



(a) Examples of event segmentation results of SAPS(w/ GPS)

(b) Examples of event segmentation results of SAPS(w/o GPS)

Fig. 8. Comparison of tempo-spatial event segmentation results between GPS data and spatial information.

(a) Examples of photos from subset of places365-standard datasets for training our model

(b) Examples of correct prediction results on our own datasets from our trained model

Fig. 9. Examples of datasets for training and prediction results.

When GPS data is unavailable because photos do not include GPS data in their headers, spatial information is obtained from a CNN model trained with five scene categories.

In the case where a photo collection does not contain GPS data in the header, as an alternative to GPS data, spatial information is obtained from our CNN model trained on the five scene categories. Spatial information of photos are represented by five dimensional feature vectors encoded in one-hot vectors from the model's predictions.

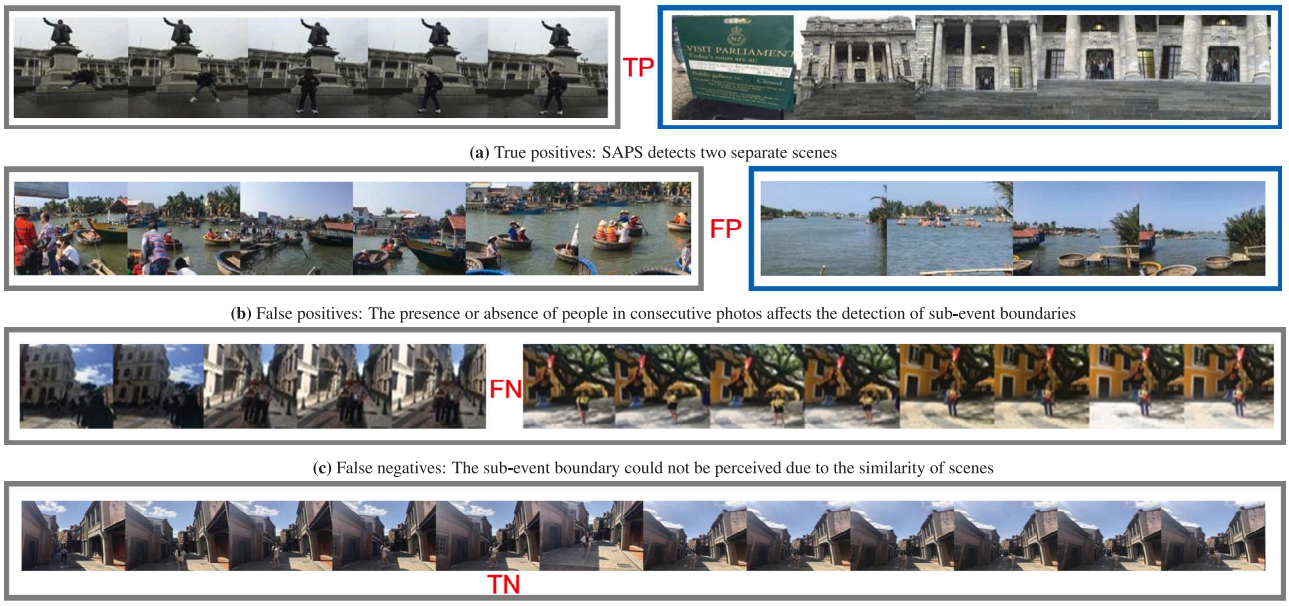
Table 3
Experimental results of SAPS event segmentation.

Dataset	No. of photos	No. of events SAPS(w/ GPS)	No. of events SAPS(w/o GPS)
User 1	947	33	27
User 2	498	20	7
User 3	1,884	49	18
User 4	802	32	29
User 5	601	27	25
User 6	753	19	9

4.2. Experimental results of event segmentation

We cluster photos into events that are adjacent in time and place in Phase 1. Tempo-spatial clustering using time and GPS information forms a large segment of the photos when many photos are captured at a specific place regardless of the content of the photos. When tempo-spatial clustering with time and scene classifications of visual content is performed, events are grouped with photos with high scene similarity within a closely confined time period. For example, in the case of Fig. 8a, photos of the Milan Cathedral and the Leaning Tower of Pisa belong to different events according to their GPS information, whereas in Fig. 8b, they belong to one event because their spatial information is represented by the same place category.

Since we aim to detect event boundaries in chronologically arranged photostreams, time information is the most important factor. To balance between 1-dimensional time and 5-dimensional scene features, we give a low weight to the scene features. As a result, the number of events found after tempo-spatial clustering in Phase 1 is shown in Table 3 . SAPS(w/ GPS) uses GPS data and SAPS(w/o GPS) uses scene recognition information. As we set the scene category to five, SAPS(w/o gps) tends to split into fewer events than SAPS(w/ gps).



(a) True positives: SAPS detects two separate scenes

(b) False positives: The presence or absence of people in consecutive photos affects the detection of sub-event boundaries

(c) False negatives: The sub-event boundary could not be perceived due to the similarity of scenes

(d) True negatives: There is no sub-event boundary since contiguous photos are similar

Fig. 10. Examples of sub-event detection provided by SAPS. The detected sub-events are framed in separate boxes.

Table 4
Experimental results of SAPS sub-event segmentation.

Dataset	No. of photos	No. of detected sub-event boundaries	No. of correctly detected boundaries
User 1	947	87	70
User 2	498	66	58
User 3	1,884	202	143
User 4	802	81	62
User 5	601	46	43
User 6	753	70	46

Table 5
Comparison of the sub-event segmentation methods.

Method	Precision	Recall	F-score
PhotoToc [16]	0.43	0.64	0.51
Multi-Modal (time+GPS) [17]	0.70	0.39	0.48
Multi-Modal (time+GPS+color)	0.38	0.56	0.45
Multi-Modal (time+GPS+CNN)	0.38	0.68	0.48
SR-clustering [21]	0.54	0.53	0.53
CES [22]	0.62	0.64	0.63
SAPS(w/o GPS)	0.60	0.72	0.65
SAPS(w/ GPS)	0.75	0.80	0.77

4.3. Experimental results of sub-event segmentation

In Fig. 10, we present the results of applying the proposed SAPS method to one of the tested photo collections and where the detected sub-events are framed in separate boxes. Our method is capable of detecting sub-event boundaries if there is a significant difference in the visual contents (see Fig. 10a). However, in Fig. 10b, this method detected the wrong sub-event boundary, which indicates that the presence or absence of people in consecutive photos can affect the detection of sub-event boundaries. In Fig. 10c, the photos were taken in different locations, but the sub-event boundary could not be perceived owing to the similarity of scenes. Fig. 10d shows that SAPS detects no boundary since contiguous photos are very similar. There were also difficulties in setting the ground truths for sub-event boundaries owing to groups of irrelevant photos. For example, since the first photo of the second segment in Fig. 10a is not related to the previous or next segment, the detected sub-event boundary can be either the photo or the next photo.

Table 4 shows the detected sub-event boundaries and correctly detected boundaries using the proposed SAPS method. We compared the SAPS method with the following baselines:

- PhotoToc [16]: Time-based clustering that uses an adaptive threshold method to detect noticeable time gaps
- Multi-modal [17]: A generative probabilistic model that uses the combination of visual features such as time, gps, color and CNN to determine the optimal event boundaries
- Semantic regularized clustering (SR-clustering): SR-clustering, as described in [21], which uses visual and semantic features
- CES [22]: A segmentation framework that uses an LSTM-based generative network to decide whether a photo is an event boundary by comparing the visual context generated from the photos in the past, to that predicted in the future.

Moreover, we adjust some parameters of the baseline methods for fair comparison so that they fit in the sub-event segmentation environment. In the multi-modal event segmentation, the value of initial k was set to the number of ground truth sub-events, while time and GPS values were normalized. Meanwhile, color and CNN features were reduced to the dimensions of 64 and 128, respectively, by applying principal component analysis. In addition, we apply SR-clustering (available on github site),[1] by utilizing CNN and Image semantic features. In this method, we used a cut value of 0.2 while the unary and pairwise parameters set to be 0.9 and 0.005, respectively. Additionally, we set the window size to 1 for the boundary prediction function to obtain the optimal value in the case of CES method.[2]

In Table 5, we compare the performance of our algorithm with the baseline methods. The proposed model outperforms PhotoToc [16], SR-clustering algorithm proposed in [21] and generative model with deep learning features in [17]. More specifically, our goal is to segment the whole collection in detail by sub-event level. Therefore, there is a limitation in precise sub-event segmentation using only time or GPS features. Time-based clustering method tends to split into smaller clusters. High-level semantic features enhance event segmentation performance, whereas in terms of sub-event segmentation such features are

[1] https://github.com/MarcBS/SR-Clustering.
[2] https://github.com/GarciaDelMolino/contextual-event-segmentation.

Table 6
Comparison of sub-event segmentation methods with deep learning features.

Method(Model)	Feature(D)	Precision	Recall	F-score
SAPS(Inception-ResNetV2) [37]	1,536	0.50	0.69	0.57
SAPS(DenseNet201) [38]	1,920	0.59	0.72	0.64
SAPS(NASNetLarge) [39]	4,032	0.53	0.65	0.58
SAPS(VGG16-Places365) [29]	4,096	0.52	0.85	0.63
SAPS(proposed method)	1978	0.75	0.80	0.77

Table 7
Comparison of key photo selection methods.

Method	Precision
Similarity-based algorithm [16]	0.52
Representative-based algorithm [23]	0.53
Ranking algorithm [17]	0.61
SAPS(proposed algorithm)	0.69

not helpful in the generative model. In addition, the proposed method outperforms SR-clustering algorithm that uses semantic features extracted from semantic similarity graph and CES that uses visual context predictors.

Inspired by deep learning models showing the best performance in image classification, we performed experiments that take advantage of the visual content features obtained from the pre-trained image classification model. Keras library [40] provides deep learning models with pre-trained weights. We used these models for feature extraction on our datasets. We selected state-of-art models with weights trained on ImageNet. Moreover, we extracted features from VGG16 scene recognition model with weights trained on Places database. For our experiments we use feature vectors provided by the fully connected layer, that is the layer following the convolution base in the case of the VGG16 model. For the other models, we extracted features from the output of the average pooling layer, except in the last classification layer. The experimental results are shown in Table 6. DenseNet201 and VGG16-Places365 achieved high recall scores by finding a large number of sub-event boundaries, but recorded low performance with respect to their precision. Unlike the single image classification problem, a higher performance is achieved when exploiting sentiment features to detect sub-event boundaries in photo collections.

4.4. Experimental results of key photo selection

To evaluate our key photo selection algorithm, we compared the performance of our method with the baseline methods proposed in [16, 21], and [17]. We also asked the participants to select one or more key photos within each sub-event for setting the ground truth of key photo selection.

In general, smartphone users take several photos with similar poses during important events. Thus, we set the ground truth for a group of photos with similar quality, composition, and people within the same sub-event. For fair comparison, we also set the same sub-event segmentation and selected one photo from all sub-events following each method. Table 7 presents the results of key photo selection.

Overall, the proposed SAPS method provided high scores to photos with positive sentiment and emotions, detected from facial expressions. It shows a higher performance than representative or similarity-based algorithms. It also outperformed (by approximately 12%) the ranking algorithm [17], which rates according to quality, representativeness and popularity. We further conduct experiments to verify the most affecting attributes to the performance. It is demonstrated that emotion achieves the highest precision and sentiment achieves the second highest. Thus, both emotion and sentiment features are crucial for summarizing personal photo collections.

5. Conclusion and future work

We proposed the SAPS method that automatically summarizes personal photo collections focusing on visual sentiment features. For this purpose, events are segmented using time and location features; then each event is divided into sub-events by calculating the novelty scores to determine the sub-event boundaries. Next, we summarized the photo collection using a ranking algorithm measuring three attributes: sentiment, emotion, and aesthetics. Finally, we evaluated the proposed method by applying it to real-world photo collections consisting of 5,480 photos in total. The results indicated that our proposed method is effective in finding sub-event boundaries and key photos, as it primarily focused on detailed sentiment features instead of general content features. We plan to extend this method into storytelling to extract certain keywords and automatically annotate textual information.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the G-ITRC (IITP-2019-2015-0-00742), the ICT Creative Consilience program(IITP-2020-2051-001), the Basic Science Research Program (2019R1C1C1008956), and the Engineering Research Center Program (NRF-2018R1A5A1059921) funded by the MSIT, South Korea.

References

[1] D. Kuzovkin, T. Pouli, O.L. Meur, R. Cozot, J. Kervec, K. Bouatouch, Context in photo albums: Understanding and modeling user behavior in clustering and selection, ACM Trans. Appl. Percept. (TAP) 16 (2) (2019) 1–20.
[2] D. Carrington, How many photos will be taken in 2020? - life in focus, 2020, https://focus.mylio.com/tech-today/how-many-photos-will-be-taken-in-2020 [Accessed: 04/10/2020].
[3] A. Pigeau, Life gallery: event detection in a personal media collection, Multimedia Tools Appl. 76 (7) (2017) 9713–9734.
[4] S. Lonn, P. Radeva, M. Dimiccoli, Smartphone picture organization: A hierarchical approach, Comput. Vis. Image Underst. 187 (2019) 102789.
[5] K. Ahmad, N. Conci, How deep features have improved event recognition in multimedia: A survey, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 15 (2) (2019) 1–27.
[6] S.N. Aakur, S. Sarkar, A perceptual prediction framework for self supervised event segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[7] S. Bacha, M.S. Allili, N. Benblidia, Event recognition in photo albums using probabilistic graphical models and feature relevance, J. Vis. Commun. Image Represent. 40 (2016) 546–558.
[8] L. Zhang, B. Denney, J. Lu, Sub-event recognition and summarization for structured scenario photos, Multimedia Tools Appl. 75 (15) (2016) 9295–9314.
[9] A. Ceroni, V. Solachidis, C. Niederée, O. Papadopoulou, N. Kanhabua, V. Mezaris, To keep or not to keep: An expectation-oriented photo selection method for personal photo collections, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 187–194.
[10] T.C. Walber, A. Scherp, S. Staab, Smart photo selection: Interpret gaze as personal interest, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, pp. 2065–2074.
[11] Y. Li, M. Geng, F. Liu, D. Zhang, Visualization of photo album: selecting a representative photo of a specific event, in: International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 128–141.
[12] A. Ceroni, V. Solachidis, M. Fu, N. Kanhabua, O. Papadopoulou, C. Niederée, V. Mezaris, Investigating human behaviors in selecting personal photos to preserve memories, in: International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2015, pp. 1–6.
[13] D. Nolasco, J. Oliveira, Subevents detection through topic modeling in social media posts, Future Gener. Comput. Syst. 93 (2019) 290–303.
[14] J.P. Gozali, M.-Y. Kan, H. Sundaram, Hidden Markov model for event photo stream segmentation, in: 2012 IEEE International Conference on Multimedia and Expo Workshops, IEEE, 2012, pp. 25–30.

[15] A. Graham, H. Garcia-Molina, A. Paepcke, T. Winograd, Time as essence for photo browsing through personal digital libraries, in: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, 2002, pp. 326–335.
[16] J.C. Platt, M. Czerwinski, B.A. Field, Phototoc: Automatic clustering for browsing personal photographs, in: Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, Vol. 1, IEEE, 2003, pp. 6–10.
[17] X. Shen, X. Tian, Multi-modal and multi-scale photo collection summarization, Multimedia Tools Appl. 75 (5) (2016) 2527–2541.
[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.
[20] S. Liu, G. Tian, Y. Xu, A novel scene classification model combining resnet based transfer learning and data augmentation with a filter, Neurocomputing 338 (2019) 191–206.
[21] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S.G. Nikolov, P. Radeva, Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation, Comput. Vis. Image Underst. 155 (2017) 55–69.
[22] A. Garcia del Molino, J.H. Lim, A.H. Tan, Predicting visual context for unsupervised event segmentation in continuous photo-streams, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 10–17.
[23] T. Mei, B. Wang, X.-S. Hua, H.-Q. Zhou, S. Li, Probabilistic multimodality fusion for event based home photo clustering, in: International Conference on Multimedia and Expo, IEEE, 2006, pp. 1757–1760.
[24] N. Murray, L. Marchesotti, F. Perronnin, AVA: A large-scale database for aesthetic visual analysis, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2408–2415.
[25] J. Kim, J. Lee, Travel photo album summarization based on aesthetic quality, interestingness, and memorableness, in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE, 2016, pp. 1–5.
[26] Q. You, J. Luo, H. Jin, J. Yang, Joint visual-textual sentiment analysis with deep neural networks, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 1071–1074.
[27] X. Liu, N. Li, Y. Xia, Affective image classification by jointly using interpretable art features and semantic annotations, J. Vis. Commun. Image Represent. 58 (2019) 576–588.
[28] D. Borth, T. Chen, R. Ji, S.-F. Chang, Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content, in: Proceedings of the 21st ACM International Conference on Multimedia, ACM, 2013, pp. 459–460.
[29] G. Kalliatakis, Keras-VGG16-places365, 2017, https://github.com/GKalliatakis/Keras-VGG16-places365 [Accessed: 04/10/2020].
[30] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. (2) (1979) 224–227.
[31] J. Tang, L. Fu, C. Tan, M. Peng, Research on sentiment classification of active scene images based on DNN, in: 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), IEEE, 2019, pp. 169–172.
[32] J. Foote, Automatic audio segmentation using a measure of audio novelty, in: International Conference on Multimedia and Expo. ICME2000, Vol. 1, IEEE, 2000, pp. 452–455.
[33] H. Tong, M. Li, H. Zhang, C. Zhang, Blur detection for digital images using wavelet transform, in: International Conference on Multimedia and Expo, Vol. 1, IEEE, 2004, pp. 17–20.
[34] L. Mai, H. Le, Y. Niu, F. Liu, Rule of thirds detection from photograph, in: 2011 IEEE International Symposium on Multimedia, IEEE, 2011, pp. 91–96.
[35] V. Vonikakis, R. Subramanian, J. Arnfred, S. Winkler, A probabilistic approach to people-centric photo selection and sequencing, IEEE Trans. Multimed. 19 (11) (2017) 2609–2624.
[36] O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification, 2017, arXiv preprint arXiv:1710.07557.
[37] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[38] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[39] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710.
[40] F. Chollet, et al., Keras, 2015, https://keras.io [Accessed: 04/10/2020].