Multimed Tools Appl (2016) 75:2527–2541
DOI 10.1007/s11042-015-2658-6

Multi-modal and multi-scale photo collection summarization

Xu Shen[1] · **Xinmei Tian**[1]

Received: 1 October 2014 / Revised: 22 April 2015 / Accepted: 23 April 2015 /
Published online: 24 May 2015

Abstract With the proliferation of digital cameras and mobile devices, people are taking much more photos than ever before. However, these photos can be redundant in content and varied in quality. Therefore there is a growing need for tools to manage the photo collections. One efficient photo management way is photo collection summarization which segments the photo collection into different events and then selects a set of representative and high quality photos (key photos) from those events. However, existing photo collection summarization methods mainly consider the low-level features for photo representation only, such as color, texture, etc, while ignore many other useful features, for example high-level semantic feature and location. Moreover, they often return fixed summarization results which provide little flexibility. In this paper, we propose a multi-modal and multi-scale photo collection summarization method by leveraging multi-modal features, including time, location and high-level semantic features. We first use Gaussian mixture model to segment photo collection into events. With images represented by those multi-modal features, our event segmentation algorithm can generate better performance since the multi-modal features can better capture the inhomogeneous structure of events. Next we propose a novel key photo ranking and selection algorithm to select representative and high quality photos from the events for summarization. Our key photo ranking algorithm takes the importance of both events and photos into consideration. Furthermore, our photo summarization method allows users to control the scale of event segmentation and number of key photos selected. We evaluate our method by extensive experiments on four photo collections. Experimental results demonstrate that our method achieves better performance than previous photo collection summarization methods.

✉ Xinmei Tian
xinmei@ustc.edu.cn

Xu Shen
shenxu@mail.ustc.edu.cn

[1] University of Science and Technology of China, Hefei, China

Keywords Photo collection summarization · Event segmentation · Key photo selection

1 Introduction

Photo is a major form of content creation, providing unlimited usages in the daily life. The proliferation of digital cameras and smart phones greatly facilitates the photography process and encourages people to take more and more photos. However, large amount of redundant photos corresponding to a specific event may be taken and stored in users' mobile devices. It leads to a growing demand for photo collection summarization tools, which can help people to organize, browse and search their photo collections.

Photo collection summarization and event-driven summarization of videos have drawn increasing attention and considerable research has been proposed [4, 5, 8, 9, 14, 15, 17, 18, 20, 21, 24–26]. Since a photo collection often records the sequential activities of people, an effective way to summarize it is to segment the photo collection into events and then select key photos from each event for summarization. The key photos are the ones which are both representative and with high-quality in the events. Previous photo summarization works mainly rely on photos' time and low-level visual features only (such as color, texture, etc.) in the event segmentation step and select fixed number of key photos for each event in the key photo selection step [4, 14, 15, 20]. However, event segmentation is a very challenging problem since event is a highly abstract and high-level concept. Previous event segmentation methods consider the low-level visual features of a photo only, but ignore many other useful and important information, for example the high-level semantic features and location information. As a consequence, those event segmentation methods cannot achieve satisfactory performance. Besides, previous photo collection summarization methods select the key photos mainly based on the representativeness only and do not take photo's quality into consideration. In fact, both quality and representativeness should be simultaneously considered in the key photo selection stage. Moreover, prior summarization methods provide no flexibility in the event segmentation granularity and only choose a pre-defined number of key photos from each event individually, leading to very limited flexibility for people to manage the summarization results.

In order to tackle those problems, we propose a multi-modal and multi-scale photo collection summarization method. In our method, we adopt multi-modal features including time, texture features [23], high-level semantic features [3] and location features [13] for photo collection event segmentation. High-level semantic features are extracted by a deep convolutional neural network trained on a large scale of image collection [12]. Meanwhile, GPS features are extracted from the EXIf of photos. With a combination of time, location and semantic features, we can model the event concept of photo collections more precisely. For better selection of key photos, we propose to evaluate photo's importance by considering photo's quality and representativeness simultaneously. In addition, we also present an effective way to rank all photos in the photo collection based on the importance of both photos and events. In this way, user can control the scale of the summarization by choosing arbitrary number of key photos from this rank.

Our photo collection summarization method can be summarized as follows. Firstly, we represent photos in the collection by their multi-modal features (time, high-level semantic feature, GPS feature). Secondly, we use a Gaussian mixture model to cluster these photos into different events based on their multi-modal feature, in which users are allowed to control the scale/granularity of the events. Thirdly, our proposed photo ranking algorithm is introduced to rank photos in the photo collection. Finally, multi-scale photo collection

summarization is achieved by choosing arbitrary number of key photos from the ranking list.

We evaluate our photo collection summarization method on a set of photo collections consisting of 2849 photos collected from 4 users. The experimental results indicate that our multi-modal feature based event segmentation algorithm performs better than low-level feature based segmentation models [5, 21], and our key photo selection results fit the users' need better than previous methods [5, 17].

The rest of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 details our proposed algorithm, including photo representation, event segmentation and key photo selection. Extensive experiments and analyses are presented in Section 4, followed by the conclusion and future work in Section 5.

2 Related work

Generally, photo collection summarization process has three steps. First, photos are ranked in chronological order and represented by feature vectors. Second, photo collection is segmented into different clusters (events). Third, key photos are selected from each event for summarization. Therefore, we review the related work from those three aspects: photo representation, event segmentation and key photo selection.

Photo representation As time stamp is a key factor for event clustering, lots of photo summarization works are built based on analysing the date/time information of the photos [8, 14, 20, 21, 26]. To overcome the limited information provided by time/date, subsequent researches represent photos by incorporating content information, including color histogram [8, 14, 16], texture histogram [7, 8], low-frequency DCT feature [5], Exchangeable Image File Format (EXIf) [7, 8, 20]. For information in EXIf, Gong and Jain utilized scene brightness value derived from focal length and aperture diameter [6, 7], while aperture, exposure time and focal length were extracted in [17]. Although Gong and Jain point out that Global Positioning System(GPS) is an important feature of photos [6], GPS has not been incorporated for photo collection segmentation yet. In addition, Deep Learning feature has achieved progress in image representation [12], to incorporate this high-dimensional feature into our algorithm, we need to reduce the feature dimension. Lots of algorithms are proposed to do feature dimension reduction, including feature mapping and selection [10], subspace learning [22], PCA [2] and sparse representation [2].

Event segmentation Generally, event segmentation of photo collection can be achieved by setting boundaries in the timeline. Platt proposed to set a boundary between two photos when their time gap was greater than 1 hour [20]. Later, he improved this work by using adaptive threshold [21]. Graham et al. extended this local threshold-based algorithm by introducing the intra-cluster rates and inter-cluster time gap to refine the original clusters [8]. Gargi came up with a bottom-up and adaptive approach that marks long interval with no capture as the end of an event and sharp local increase in the frequency of captures as the start of an event [26]. In [5], boundaries are selected by applying confidence score, dynamic programming or Bayes information criterion (BIC) criteria on the similarity matrix of photos. More generally, the event segmentation process can also be treated as a clustering problem. Loui and Svakis presented a $2-$class K-means algorithm for event clustering, followed by checking color similarity of photos within these clusters [14]. Gong and Jain [6] used hierarchical agglomerative clustering algorithm to group photo collections. Hidden

Markov model with learned parameters was adopted for photo collection segmentation in [21]. Mei et al. addressed this problem by utilizing a unified probabilistic framework to model all the photos, then event was discovered by fitting the generative model [17].

Key photo selection Current key photo selection algorithms mainly rely on photo's representativeness only [4, 5, 17]. Cooper et al. simply recommend the earliest photo in each event as key photo [5]. Mei et al. choose photo with the maximum a priori probability among photos in current event as the representative photo [17]. In [4], key photo is automatically selected by examining the mutual relation between near-duplicate photo pairs in photo clusters. There are two drawbacks in current key photo selection algorithms. First, photo importance is only evaluated by representativeness, while the image quality and the importance of events are ignored. Second, only a fixed number of photos are selected from events, thus users have no access to multi-scale summarization of the photo collection.

3 Multi-modal and multi-scale photo collection summarization

The framework of our proposed photo collection summarization model is presented in Fig. 1. First, photos are represented by their time stamp, GPS information, and high-level content feature. Second, photos are clustered into events via a Gaussian mixture model in the multi-modal feature space. Finally, we build a photo rank for multi-scale summarization of the photo collection. User can choose arbitrary number of key photos from this rank. The multi-modal features utilized for photo representation are described in Section 3.1. Section 3.2 and Section 3.3 present our event segmentation algorithm and key photo selection algorithm, respectively.

3.1 Multi-modal photo representation

The multi-modal features used for photo representation in our event segmentation algorithm include time, color, texture, location, and deep learning feature.

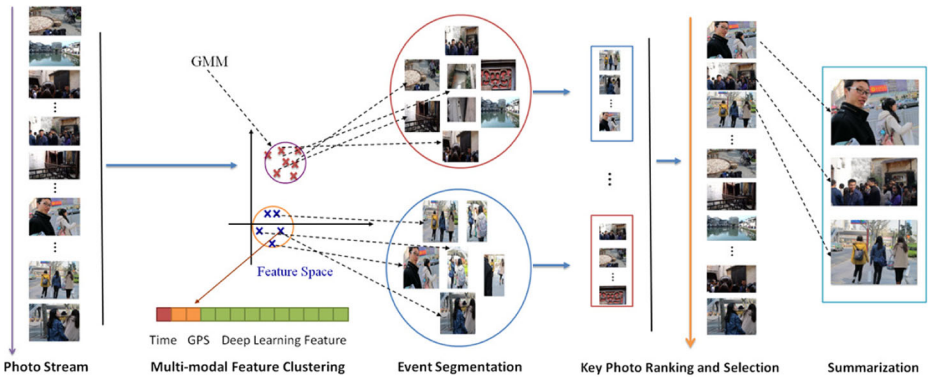

Fig. 1 Illustration of the proposed photo collection summarization method. Photos are firstly ordered according to their time stamp. With photos represented by multi-modal features, Gaussian mixture model is applied for event segmentation. Finally, key photos are ranking and selected for summarization

Time (TM) – For each photo, the EXIf headers are processed to extract the date/time stamp. If the time stamp is unavailable, we rely on the modification time of photo instead.

Location (G) – Generally, GPS information can be obtained from the longitude and latitude stamp in EXIf headers. If location information is not included in the headers, we set it to be the same as the closest photo in time stamp.

Color (C) – For a color image x with N pixels,we derive a color histogram $(64D)$ $H(x) = [h_1, h_2, ..., h_{64}]$ in RGB color space, where $h_i = N_i/N$ is the proportion of pixels whose value falls into bin i.

Texture (TX) – Coarseness, contrast, directionality(20D) of the Tamura descriptor [23] are adopted to describe the texture feature of photos.

Deep Learning feature (DL) – With the help of the open source deep learning framework called Caffe [28], we implemented the deep convolutional neural network in [12] to extract a 4096D deep feature of images. In [12], Krizhevsky et al. designed a deep convolutional neural network with millions of parameters and applied it on the ImageNet classification. Their model contains eight learned layers and adopts the Rectified Linear Units as the activation function [27]. The overall architecture of their network is shown in Fig. 2. To ensure that other features are not covered by the high-dimensional deep learning features, we adopt PCA [1] to reduce deep feature from 4096D to 128D with little degradation of performance.

It can handle with the RGB images instead of gray images only. Their neural network consists of five convolutional layers and three full connected layers. The output of the last layer is followed by a 1000-way softmax that produces a distribution over the given 1000 image class labels. Then supervised learning process is conducted on the training set of the ImageNet database. Finally it achieves a breakthrough on this challenging dataset which proves its descriptive power on all kinds of images. In this paper, we firstly implement a deep convolutional neural network that has the same architecture with [12]. Then this neural network is carefully trained on the ILSVRC-2012 training set, which extracts 1.2 million images that cover 1000 categories from the whole ImageNet database. In this way, the convolutional neural network accumulates enough knowledge to understand various images well. For a given image x_i, the last hidden layer of the convolutional neural network produces 4096-dimensional activations, which is the deep learning feature we use to represent the high-level feature of x_i.

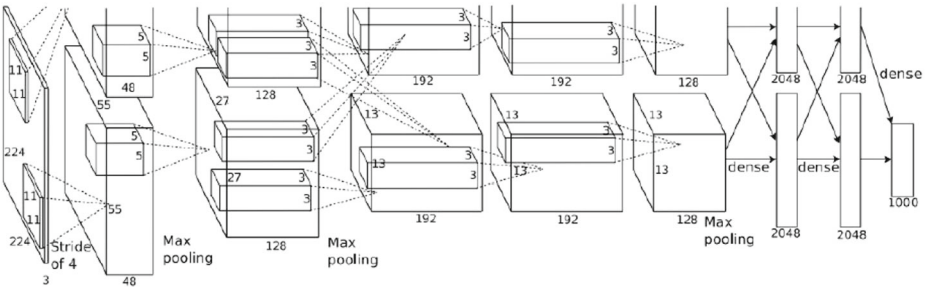

Fig. 2 The architecture of deep convolutional neural network used by Krizhevsky et al. The first five layers of it are convolutional and the following three layers are fully-connected. Input images are resized to be $224 * 224 * 3$

3.2 Event segmentation

Definition Given the Gaussian mixture model, photos in the same event share the same distribution in the multi-modal feature space, i.e., they are close in time, location and content. Different events should differ in concept classes and distributions. That is to say, each photo $x_i \in X = \{x_1, x_2, ..., x_N\}$ corresponds to one latent semantic concept class - event $e_j \in E = \{e_1, e_2, .., e_K\}$, where N and K are the total numbers of photos and events in the photo collection X, respectively. Naturally, the probability of photo x_i generated from event e_j can be formulated as $p(x_i|e_j)$.

Probabilistic model In our model, a photo can be represented by a continuous feature vector consists of time (TM), location (G), color (C), texture (TX), deep learning (DL) feature. If we assume that all these components are independent given the latent concept event e_j, then a priori probability $p(x_i|e_j)$ can be computed by :

$$p(x_i|e_j) = p(TM_i|e_j)p(G_i|e_j)p(C_i|e_j)p(TX_i|e_j)p(DL_i|e_j) = \prod_{l=1}^{L} p(x_{i,l}|e_j) \quad (1)$$

where $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,L})$. $x_{i,l}$ is the lth metadata of photo $x_i (l = 1, 2, ..., L)$, and $L = 5$ is the number of components of the feature vector. Each component $x_{i,l}$ is generated by a single Gaussian distribution:

$$p(x_{i,l}|e_j) = \frac{1}{\sqrt{2\pi\sigma_{i,l,j}^2}} e^{-\frac{(x_{i,l,j}-u_{i,l,j})^2}{2\sigma_{i,l,j}^2}} \quad (2)$$

Learning parameters via EM training We adopt the most widely used principle - maximizing the log-likelihood of the joint distribution to train parameters of each Gaussian distribution. The objective function can be formulated as:

$$l(X;\theta) \triangleq log\left(\prod_{i=1}^{N} p(x_i|\theta)\right) = \sum_{i=1}^{N} log\left(\sum_{j=1}^{K} p(e_j)p(x_i|e_j,\theta)\right) \quad (3)$$

where $p(x_i|e_j, \theta)$ is computed according to (1) with θ given. $p(e_j)$ is the priori probability of event e_j. K and N are the number of events and photos, respectively. To optimize the model, we introduce EM training to tune the parameters. Before the EM training process, values of parameters are initialized by $K - means$ with a given K. The nth iteration of EM training for the given K is presented in Algorithm 1.

The number of events K must be pre-defined to perform EM training. To tackle this problem, we propose to generate a series of candidate segmentations by applying EM training to multiple Ks. Then we select the best K by using the model selection principle in [11] :

$$K^* = \underset{K}{argmax}\{2 \times l(X;\theta) - m_K \times logN\} \quad (4)$$

Where $m_K = (K-1) + K \times N_G$, N_G is the numbers of Gaussian distributions.

Multi-scale event segmentation In order to ensure that user can have great variety in segmentation scale, we introduce a new scale variable s into our model selection principle to adjust the scale of the final segmentation. The new selection principle turns to:

$$K^* = s \times \underset{K}{argmax}\{2 \times l(X;\theta) - m_K \times logN\}, \quad (5)$$

Algorithm 1 The EM training algorithm

E step: Compute the likelihood by (3)
M step:
1. Update posteriori of event e_j by Bayes rule:
$p(e_j|x_i)^{n+1} = \frac{p(e_j)^n p(x_i|e_j)^n}{\sum_{i=1}^{N} p(e_j|x_i)^{n+1}}$
2. Update model parameters of event e_j:
$u_{j,l}^{n+1} = \frac{\sum_{i=1}^{N} p(e_j|x_i)^{n+1} x_{i,l}}{\sum_{i=1}^{N} p(e_j|x_i)^{n+1}}$
$\sigma_{j,l}^{n+1} = \frac{\sum_{i=1}^{N} p(e_j|x_i)^{n+1} (x_{i,l} - u_{j,l}^{n+1})^2}{\sum_{i=1}^{N} p(e_j|x_i)^{n+1}}$
3. Update model of event e_j:
$p(e_j)^{n+1} \approx \frac{1}{N} \sum_{i=1}^{N} p(e_j|x_i)^{n+1}$
$p(x_i|e_j)^{n+1} = \prod_{i=1}^{L} p(x_{i,l}|e_j)^{n+1}$

where $s = 1, 0 < s < 1, s > 1$ correspond to proper scale, coarse scale and fine scale event segmentation, respectively. In our experimental evaluation, we will test the performance of our model in different segmentation scales. We summarize our multi-scale photo stream segmentation algorithm in Algorithm 2.

Algorithm 2 Multi-scale photo stream segmentation

(1) Extract multi-modal features of the photo collection and sort photos by their timestamps.
(2) Initialize K_{min} and K_{max}
(3) For each K from K_{min} to K_{max}
(i) Apply $K - means$ to initialize the model parameters θ
(ii) For event ID ranges in $[1, K]$, update model parameters by EM learning
(4) Select best K based on selection principle and scale s
(5) Assign photos to corresponding events based on $p(e_j|x_i)$

3.3 Key photo selection

In this section, we describe our key photo selection algorithm for photo collection summarization. Figure 3 shows the whole process of our algorithm.

In order to evaluate the importance of different photos, we propose to rank photos by a combination of their quality, representativeness and popularity feature. Details of these features are described as follows:

Quality (Q) – Users generally tend to choose photos with high quality as key photos, therefore we introduce the quality assessment model in [29] to compute the quality score as quality feature. In [29], hand-craft content features($23D$) are extracted from the photo, including hue histogram($8D$), brightness histogram($4D$), sharpness($1D$), depth of field($2D$), hue histogram of the subject region($4D$), average brightness of the subject region($1D$), relative size of the subject region($1D$), contrast between the subject and background($2D$). Then a SVM (Support Vector Machine) classifier is trained on the AVA dataset [19], in which photos are labeled as aesthetic-good(1) or aesthetic-bad(-1). Then

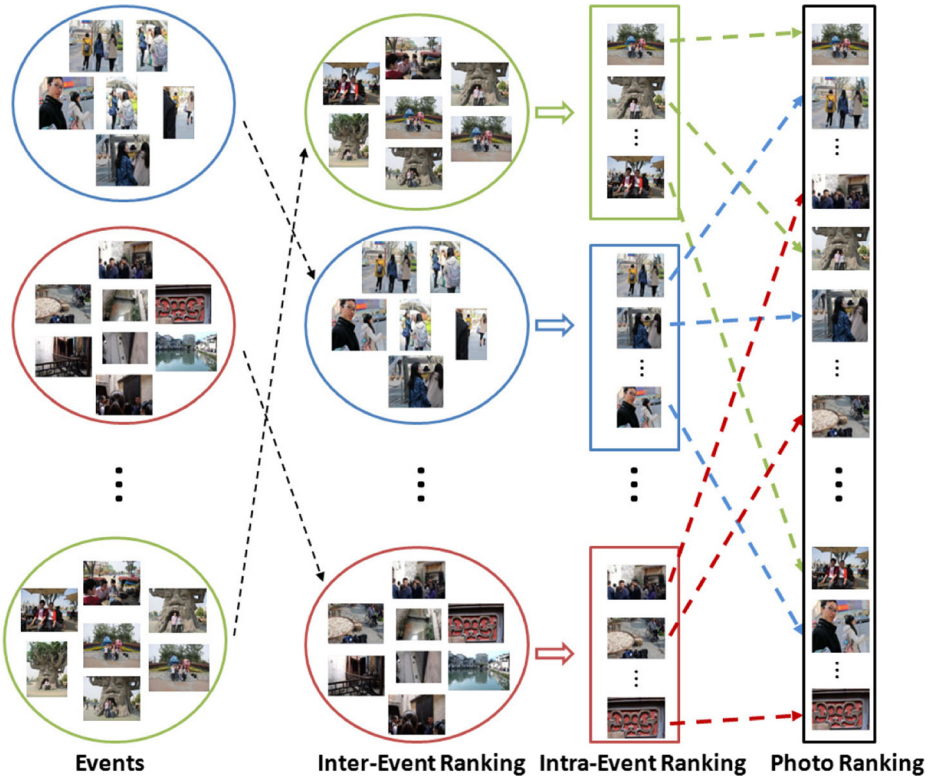

Fig. 3 Our proposed photo ranking algorithm for key photo selection. We first rank events by their popularity, then photos within each event are ranked according to their importance scores. Finally, we subsequently add photos to the final key photo list from top event to bottom event

label of the photo can be predicted as $L_i \in [-1, 1]$ by the classifier. Finally, quality of photo x_i can be computed as:

$$Q(x_i) = \frac{1}{1 + exp(-L_i)} \tag{6}$$

Representativeness (R) — The key photos of a event should represent the visual theme of this event, i.e. they must be highly representative. With the probabilistic model described in the last section, the probability of photo x_i belongs to event e_j is $p(e_j|x_i)$. So every photo can be assigned to a event e_{k^*} with the maximum a posterior probability $k^* = \underset{j}{argmax}\ p(e_j|x_i)$. Thus the representativeness of event x_i to its corresponding event e_{k^*} can be regarded as $p(x_i|e_{k^*})$, i.e.

$$R(x_i) = p(x_i|e_{k^*}) \tag{7}$$

Popularity (P) — When people are interested in a certain event, they tend to take lots of photos about that event, *vice versa*. Therefore within a single event, they tend to take more photos for scenery or object which are more attractive. In addition, a photo with more similar photos in the same event tend to be more attractive. So we use the amount of photos in the event to represent event popularity and amount of near neighbors to

represent the popularity of a photo. To measure the popularity of photos within an event, we propose to first infer the common "visual theme" of the event, then compute the relative strength of each photo to that "visual theme". The stronger the strength is, more popular the photo will be. In order to find the multiple visual themes and their relative strengths in the set of photos in each event, we apply visual rank algorithm in [11] to rank the photos by popularity. Given n images in an event $e = \{x_1, x_2, .., x_n\}$, we first build a similarity matrix S, where $S_{i,j}$ measures the visual similarity between photo i and j.

$$S_{i,j} \propto exp\{-\|x_i - x_j\|^2\} \tag{8}$$

where x_i, x_j are the content feature(color, texture, deep learning feature) of photo i and j, respectively. Then the visual rank VR of photos can be computed by iterating the following equation:

$$VR^{k+1} = dS^* \times VR^k + (1-d)p \tag{9}$$

where S^* is the column normalized adjacent matrix of S, $d \in [0, 1]$ is the damping factor and $p = [\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}]^T_{n\times 1}$ is the initial score of the n photos. Finally, the popularity score of the photo x_i can be presented as:

$$P(x_i) = VR_i^{final} \tag{10}$$

To enable our users to choose arbitrary photos for photo collection summarization as they want, we propose to generate a rank list of key photo candidates, in which relatively better photos will precede other photos. To build this rank, we need to rank events by their importance (inter-event ranking), then rank photos within these events based on photo scores (intra-event ranking).

Inter-event ranking After event segmentation in Section 3.2, all events are ranked in descending order: $E = \{e_1, e_2, ..., e_K\}$ according to the number of photos in each event, based on the assumption that an event contains more photos is more important.

Intra-event ranking After quality, representativeness, popularity of photo x_i are obtained, the importance score of photo x_i can be calculated as:

$$Score(x_i) = a \times R(x_i) + b \times Q(x_i) + c \times P(x_i), \tag{11}$$

where a, b, c are non-negative weights of the three components and $a + b + c = 1$. In our experiment, we simply set $a = b = c = \frac{1}{3}$. Now, we can rank photos within each event according to their scores.

Key photo ranking and selection Finally, if required number of key photos is B, photos to be chosen from each event is $\llcorner B/K \lrcorner$ or $\llcorner B/K \lrcorner + 1$, i.e. for the top $r = B\ \%K$ events, we pick top $\llcorner B/K \lrcorner + 1$ photos as key photos, while $\llcorner B/K \lrcorner$ photos are picked in other events. The key photo ranking and selection process is illustrated in Algorithm 3.

4 Experiments

4.1 Experimental setting

In this section, we evaluate our proposed event segmentation and key photo selection method and compare them with the baseline methods proposed in [5, 17, 21]. We invited four users to share photos in their mobiles and cameras taken during the past two years. All

Algorithm 3 Key photo selection algorithm

(1) Rank K events by descending popularity
(2) Compute number of photos to be chosen from each event: $m = \llcorner B/K \lrcorner, r = B\ \%K$
(3) For photos in event e_j, j from 1 to K
(ii) compute quality of photos by (6)
(i) compute representativeness of photos using (7)
(iii) compute popularity using (10)
(iv) rank photos based on scores obtained by (11)
(v) if $j \leq r$, push top $m+1$ photos into key photo list, else push the top m photos
(4) Return the selected B key photos.

photos have accurate time stamps, while only part of them have GPS information. Users were invited to segment all of their photos to meaningful events and choose $1 \sim 6$ key photos from each event. Both are referred as ground truth for our photo event segmentation and key photo selection algorithm. Table 1 lists the detailed information about the dataset.

We adopt the precision, recall and F-score metric described in [5] to evaluate the performance of event boundary detection and key photo selection. In event segmentation, precision indicates the proportion of correctly detected boundaries:

$$precision_{seg} = \frac{correctly\ detected\ boundaries}{total\ number\ of\ deteced\ boundaries} \tag{12}$$

Recall represents the proportion of true boundaries detected:

$$recall_{seg} = \frac{correctly\ detected\ boundaries}{total\ number\ of\ ground\ truth\ boundaries} \tag{13}$$

The F-score measures the comprehensive performance:

$$F - score = \frac{2 \times precision \times recall}{precision + recall} \tag{14}$$

Similarly, precision of key photo selection is defined as:

$$precision_{select} = \frac{correctly\ selected\ key\ photos}{total\ number\ of\ selected\ photos} \tag{15}$$

4.2 Experimental results and analyses

We first evaluate the event segmentation algorithms. Table 2 shows the segmentation performance of our algorithms as well as the baseline methods proposed in [5, 21]. It shows that our model significantly outperforms the adaptive thresholding algorithm proposed in

Table 1 Detailed information about the dataset

Dataset	Number of photos	Number of events	Number of key photos
User 1	497	32	175
User 2	1086	95	288
User 3	564	107	145
User 4	702	28	140

Table 2 Performance comparison between our event segmentation method and baseline methods

Method	Precision	Recall	F-score
PhotoTOC [21]	0.50	0.71	0.59
TEC [5]	0.39	0.54	0.45
Ours(time)	0.86	0.68	0.76
Ours(time+color+texture)	0.91	0.67	0.77
Ours(time+GPS)	0.85	0.72	0.78
Ours(time+deep)	0.88	0.71	0.79
Ours(time+GPS+deep)	0.93	0.69	0.79

[5] and similarity based algorithm in [21]. It demonstrates the effectiveness of our event segmentation framework. We further conduct experiments to verify the performance of our method with different features. We find that our model with GPS or deep learning feature is better than that with low-level content feature only. It verifies that location and high-level semantic features are helpful for event segmentation. Moreover, when multi-modal features are used in our model (time+GPS+Deep), a better performance can be further achieved. It achieves the highest precision and F-score with recall a little sacrificed.

We also evaluate our algorithm on small dataset (showed in Table 3) to see if our method is effective user who only has a few photos. The results in Tables 4 and 5 reveal that our algorithm still works well though number of user's photo dataset is not so much.

To evaluate the performance of our segmentation algorithm for multiple scales, we invite the 4 users to manually segment their photo collections into different number of events, and the labeled number of events are corresponding to scale $s = \{0.5, 1.0, 1.5, 2.0\}$, respectively. Table 6 shows that our model can provide users multi-scale event segmentation results with robust performance.

In order to evaluate our key photo selection algorithm, we compare the performance of our method with that given by the baseline methods proposed in [5, 17]. For fair comparison, we follow the setting in [5, 17] and choose one photo from every event. Table 7 shows the results of our key photo selection algorithm and that in [5, 17]. The interview with our users indicates that, quality, representativeness as well as popularity are all considered in the process of selecting key photos to summarize each event. Therefore, our photo selection method achieves a higher performance than representativeness or similarity based algorithms [5, 17]. In addition, users suggest that a photo with faces is more likely to be selected than other photos. In future work, we plan to add face features into the key photo selection process.

Furthermore, we also test the key photo selection performance when different number of key photos are required, instead of selecting only one key photo for each event. We select $B = \{10, 50, 100, 150, 200\}$ key photos for each user. The experimental results are given in Table 8. It indicates that our photo selection algorithm is robust on selecting different

Table 3 Information about our small dataset

Dataset	Number of photos	Number of events	Number of key photos
User 5	60	7	15
User 6	76	10	23

Table 4 Performance of Event Segmentation on small Dataset

Method	Precision	Recall	F-score
Ours(time)	0.8	0.86	0.83
Ours(time+GPS+deep)	0.75	0.86	0.82

Table 5 Performance of Key Photo Selection on small dataset

Method	Precision
Ours	0.58

Table 6 Performance of multi-scale event segmentation

Scale	Precision	Recall	F-score
0.5	0.98	0.59	0.74
1.0(Default)	0.93	0.69	0.79
1.5	0.91	0.63	0.75
2.0	0.88	0.66	0.76

Table 7 Performance comparison of key photo selection. Our key photo selection method significantly outperforms the methods proposed in [17] and [5]

Method	Precision
Representative-based [17]	0.54
Similarity-based [5]	0.53
Ours	0.61

Table 8 Performance of selecting different number of key photos

Number of selected key photos B	Precision
10	0.50
50	0.60
100	0.63
150	0.61
200	0.57
Same as number of key photos user selected	0.58

number of key photos from the whole photo collection for users. We can see that when B is small (for example $B = 10$) the precision is a little low. This is because the number of selected key photos is much less than the number of ground-truth events. It is challenging to select a small subset of key photos from those events accurately. When B is close to the number of events, a better performance is achieved.

5 Conclusion

In this paper, we propose a multi-modal and multi-scale photo collection summarization method. Multi-modal features, including GPS, time, low-level visual feature and high-level deep learning feature are introduced for photo representation and event segmentation. We use a Gaussian mixture model for multi-scale segmentation of photo streams and an adaptive selection model for key photo selection. Our approach outperforms all baseline methods and shows robust performance for different segmentation and selection scales.

In the future, we will improve our model from the following directions. First, user feedback and interaction can be utilized to improve the performance of our model. Second, to choose key photos within an event, duplication is also an important issue to be discussed. We will focus on those aspects to improve our model.

Acknowledgments This work is supported by the NSFC under the contract No.61201413 and 61390514, the Specialized Research Fund for the Doctoral Program of Higher Education No. WJ2100060003, the Fundamental Research Funds for the Central Universities No. WK2100060011, WK2100100021.

References

1. Bao B-K, Liu G, Changsheng X, Yan S (2012) Inductive robust principal component analysis. IEEE Trans Image Process 21(8):3794–3800
2. Bao B-K, Zhu G, Shen J, Yan S (2013) Robust image analysis with sparse representation on quantized visual features. IEEE Trans Image Process 22(3):860–871
3. Bengio Y, Courville AC, Pascal V (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
4. Chu W-T, Lin C-H (2008) Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In: ACM Multimedia. ACM, pp 829–832
5. Cooper M, Foote J, Girgensohn A, Wilcox L (2005) Temporal event clustering for digital photo collections. ACM Trans Multimedia Comput Commun Appl 1:269–288
6. Gong B, Jain R (2007) Segmenting photo streams in events based on optical metadata. In: ICSC. IEEE Computer Society, pp 71–78
7. Gozali JP, Kan M-Y, Sundaram H (2012) Hidden markov model for event photo stream segmentation. In: Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops. IEEE Computer Society, pp 25–30
8. Graham A, Garcia-Molina H, Paepcke A, Winograd T (2002) Time as essence for photo browsing through personal digital libraries. In: Proceedings of the second ACM/IEEE-CS joint conference on digital libraries. ACM, pp 326–335
9. Hong R, Tang J, Tan H-K, Ngo C-W, Shuicheng Y, Chua T-S (2011) Beyond search: event-driven summarization for web videos. TOMCCAP 7(4):35
10. Hong R, Wang M, Gao Y, Tao D, Li X, Xindong W (2014) Image annotation by multiple-instance learning with discriminative feature mapping and selection. IEEE T Cybernetics 44(5):669–680
11. Jing Y, Visualrank SB (2008) Applying pagerank to large-scale image search. IEEE Trans Pattern Anal Mach Intell 30:1877–1890
12. Krizhevsky A, Sutskever I, Hinton GE. (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25. Curran Associates Inc, pp 1097–1105

13. Liu H, Mei T, Luo J, Li H, Li S (2012) Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing. In: Proceedings of the 20th ACM international conference on multimedia. ACM, pp 9–18
14. Loui AC (2000) Automatic image event segmentation and quality screening for albuming applications. In: ICME 2000, pp 1125–1128
15. Loui AC, Wood MD (1999) A software system for automatic albuming of consumer pictures. In: Proceedings of the seventh ACM international conference on multimedia (Part 2), MULTIMEDIA '99. ACM, pp 159–162
16. Loui A, Savakis A (2000) Automatic image event segmentation and quality screening for albumin application. In: Proceedings of IEEE international conference on multimedia and expo. IEEE, pp 1125–1128
17. Mei T, Wang B, Hua X-S, Zhou H-Q, Li S (2006) Probabilistic multimodality fusion for event based home photo clustering. In: ICME. IEEE, pp 1757–1760
18. Mei T, Wang B, Hua X-S, Zhou H-Q, Li S (2006) Probabilistic multimodality fusion for event based home photo clustering. In: 2006 IEEE international conference on multimedia and expo. IEEE, pp 1757–1760
19. Murray N, Marchesotti L, Perronnin F (2012) Ava: A large-scale database for aesthetic visual analysis. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2408–2415
20. Platt JC (2000) Autoalbum: clustering digital photographs using probabilistic model merging. Institute of Electrical and Electronics Engineers,Inc
21. Platt JC, Czerwinski M, Field B (2003) Phototoc: Automatic clustering for browsing personal photographs. Institute of Electrical and Electronics Engineers, Inc., p 21
22. Richang H, Bao B-K, Guangcan L (11) General subspace learning with corrupted training data via graph embedding. IEEE Trans Image Process 22:2013
23. Tamura H, Mori S, Yamawaki T Texture features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):1978
24. Tao M, Yong R, Li S, Tian Q (2014) Multimedia search reranking: a literature survey. ACM Comput Surv 46(3)
25. Teng L, Tao M, Kewon I-S, Hua X-S (2009) Multi-video synopsis for video representation. Signal Process 89(13)
26. Ullas G (2003) Modeling and clustering of photo capture streams. In: Proceedings of the 5th ACM SIGMM international workshop on multimedia information retrieval. ACM, pp 47–54
27. Vinod N, Hinton GE, Thorsten J (2010) Rectified linear units improve restricted boltzmann machines. In: Fnkranz J (ed) ICML. Omni press, pp 807–814
28. Yangqing J (2013) Caffe: an open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/
29. Zhiwei L, Wang B, Li M, Wei-Ying M (2005) A probabilistic model for retrospective news event detection. In: SIGIR. ACM, pp 106–113

Xu Shen received his bachelor's degree in electrical engineering (2012) from the University of Science and Technology of China, China. His research interests mainly include photo collection management and deep learning.

Dr. Xinmei Tian is an Associate Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. She received the Ph.D. degree from the University of Science and Technology of China in 2010. Her current research interests include multimedia information retrieval and machine learning. She received the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation award in 2013.