DOI: 10.1111/j.1467-8659.2012.03210.x
Pacific Graphics 2012
C. Bregler, P. Sander, and M. Wimmer
(Guest Editors)

Volume 31 (2012), Number 7

Hierarchical Narrative Collage For Digital Photo Album

Lei Zhang Hua Huang†

School of Computer Science and Technology, Beijing Institute of Technology, China

Abstract

Collage can provide a summary form on the collection of photos in an album. In this paper, we introduce a novel approach to constructing photo collage in the hierarchical narrative manner. As opposed to previous methods focusing on spatial coherence in the collage layout, our narrative collage arranges the photos according to the basic narrative elements from literary writings, i.e., character, setting and plot. Face, time and place attributes are exploited to embody those narrative elements in the collage. Then, photos are organized into the hierarchical structure for the multi-level details in the events recorded by the album. Such hierarchical narrative collage can present a visual overview in the chronological order on what happened in the album. Experimental results show that our approach offers a better summarization to browse on the photo album content than previous ones.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms; I.4.9 [Image Processing And Computer Vision]: Applications—

1. Introduction

The last few decades have witnessed a growing popularity of digital cameras and mobile devices, which makes photographing in common use for ordinary people. These advances have directly led to the dizzying increase on the amount of photos, e.g., there are more than 2 million photos uploaded to the famous photo sharing repository *Flickr*, and 100 million to *Facebook* everyday. People like to take pictures of interesting highlights in all kinds of events for memories, such as travels, birthdays, weddings, daily life, and so on. Then, those photos are exported to the destined albums tagged with the corresponding themes as their titles, which help to preserve and share memories across space and time. Typically, there might be hundreds of photos in a single album. Browsing on each photo one by one is a simple yet tedious task for recalling what happened during the events. Whereas a set of representative photos are much helpful to summarize the events, it is necessary to have a visually condensed form on the photos for album navigation and management.

Collage has been a fashionable composite representation for a collection of photos, which gives a condensed summary on the photos. A wealth of methods exists for producing photo collages of different styles, but mainly relying on some spatial rules to establish the coherent layout [RKKB05, WSQ*06, RBHB06]. For example, photos with similar appearance are arranged near to each other, and the ones with sky are placed on the top in the collage (see Figure 1 *left*). Such photo assembly is able to make the whole collage visually compact and appealing, but ignores the sequential occurrence of the events that happened through the photos. People might have no clear clue on recalling the events through the collage. The goal of our paper is to produce a collage form that enables apparent chronological presentation on the attracting events recorded by the photos (see Figure 1 *right*). Inspired by the narrative writing principles, we propose to adopt the structure and organization of narrative to construct the photo collage, named narrative collage. A narrative is a constructive format that describes a sequence of events [Too01]. It has three basic elements, i.e., character, setting and plot, by which a series of events can be told along the direction that happened. Hence, the challenge for constructing narrative collage is to obtain those literary elements from photos, and then arrange them in the appropriate form to fluently bring out the events.

The main contribution of our paper is a fast computational procedure to translate the literary narrative elements from digital photos, thus, creating a novel hierarchical form of narrative collage for story-telling on the events. Technically, a new narrative saliency detection method is therein devised to assist the region of interest retrieval. Experiments

† Corresponding author: huahuang@bit.edu.cn

Figure 1: *The collage result by AutoCollage [RBHB06] (left) and our approach (right).*

show that hierarchical narrative collage can produce more effective summary on the album content.

2. Related Work

Collage has been extensively studied in the field of computer graphics during the past decade. Salient regions of input photos are usually detected, and their visibility is to be maximized in the summary collage. Digital tapestry [RKKB05] determines the location of salient regions by optimizing appearance smoothness between nearby photos. But this method might generate small and isolated fragments. Wang et al. [WSQ*06] use visual attentional model to explicitly detect salient regions. They arrange the photos by adjusting their orientation and layer ordering, such that the salient regions can be most visible in the overlay arrangement. Battiato et al. [BCG*07] subsequently improve the visual effect of collage by using self-adaptive cropping algorithm to exploit more semantic information. However, these two approaches produce collages with apparent boundary artifacts between adjacent regions. Combining the rules of methods above, Rother et al. [RBHB06] propose to optimize a unified collage energy and obtain seamless layout by using α-Poisson blending. Some other methods detect more accurate salient regions to obtain geometry-compact and fascinating collages. Goferman et al. [GTZM10] arrange the salient regions of arbitrary shapes, such that the whole collage is perceived like a puzzle. Huang et al. [HZZ11] assemble the salient cutouts of selected Internet images to resemble a target arbitrary shape, generating the so-called Arcimboldo-like collage. However, those collage methods arrange the photos (salient regions) according to the spatial rules, which is not consistent to the chronological happening of the events recorded by the photos. Our method will make use of the narrative attributes for better story-telling collage.

Video is another medium in recording events, generating more sequential images compared to photo album. Many methods address the problem of video frames summarization in the collage form. Most of them arrange the keyframes using the rules akin to photo collages [KEA06, MYYH08, CLH12]. Barnes et al. [BGSF10] adopt bidirectional similarity to classify frames into multi-scale tapestries, and space-time interpolation to enable continuous zooming. Correa and Ma [CM10] propose the dynamic video narratives that focus on the demonstration of character motions in different backgrounds. Our work is partly inspired by video narrative, but photo album has no such intense continual motion, which needs effortful solution to explore its narrative structure.

Collage is closely related to image saliency detection. Since collage processes a multitude of photos, saliency detection is in favor of fast computation and semantic analysis. Itti et al. [IKN98] build multi-scale visual attention features and compute the saliency map by the local central-surrounded difference, which has been used in the collage like [WSQ*06, RBHB06]. Goferman et al. [GZMT10] propose to use the spatially weighted color contrast to measure the saliency magnitude, which can also detect the context saliency. This method is used in puzzle-like collage [GTZM10], but suffers slow computation. Cheng et al. [CZM*11] employ global color contrast and spatial coherence to evaluate the saliency, which has much fast computation for robust saliency detection. In this paper, we will adapt this global contrast based method to a few of concomitant photos to compute the narrative saliency, which generates regions of interest for narrative collage construction.

3. Preliminaries

Given an album $\mathcal{A}$, we denote the photos belonging to it as $\mathcal{A} = \{P_{i=1...N}\}$, where N is the photo number. Usually,

those photos are stored according to their date taken, and sequentially named with ascending indices in the album (see Figure 2), resulting in a natural presentation narrating the recorded events. However, such narration is not always commendable due to the following observations:

- If there are several cameras recroding the same event, mixing photos from different albums into one might disorder the natural sequence indicated by the photo names. This is normal for photos taken in the group gathering.
- Some photos capture almost the same scene, which slow down the narrating step for the event overview. For example, consecutive photos might be taken by the triple shot with different zooming, exposure or poses.
- The native linear layout of album makes the photo browsing inefficient. People have to look through all the preceding photos for the interested highlight during the events.

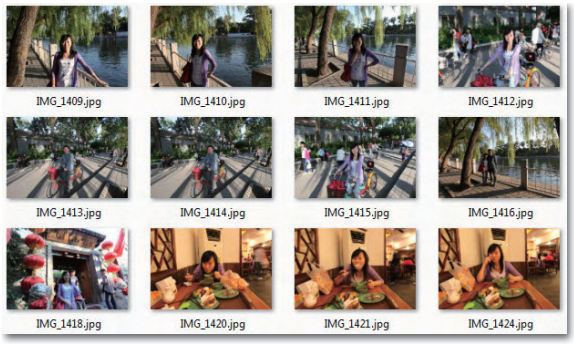

Figure 2: *Photos with the indexed names stored in an album.*

To overcome those limitations, we borrow the notion of narrative elements (character, setting and plot) [Too01] to summarize and arrange photos in the collage. Literarily, characters are the people that a story is about, who appear and join in the events. Setting is the time and place in which the events take place. Plot tells the highlight that happened in an event. These elements can well interpret the substance of activities in the photos, and compose the summarization to address the first and second issues above. So we compute those narrative attributes from photos, by which the photos are further classified into hierarchical structure (Section 4). To keep the informative part of each photo, we further detect the regions of interest using a novel narrative saliency map (Section 5). Finally, the photos are seamlessly blended into a hierarchical collage based on their narrative attributes (Section 6). From the hierarchical collage, people can look through the coarse-to-fine details of the events in a flexible way, instead of linear browsing in the third issue.

4. Construction of hierarchical narration

The essence of narrative collage is to thread the photos in an album according to the forementioned narrative elements. Given the album $\mathcal{A}$, we define the attributes $<\gamma,\sigma,\pi>$ corresponding to the elements of character, setting and plot respectively. Next, we will elaborate the translation of those literary notions to their computational formulation.

4.1. Computation of narrative elements

Character. Character refers to people that appear in the photos. Thus, the computation of character can be cast as the human tracking and recognition. Advanced human behavior analysis could improve the narrative effect of the collage, but imposes the unfavorable computational complexity. So we simply use the classical face detection approach [VJ01] to estimate the existence of faces and their positions. Then, the character attribute is defined as $\gamma = \{(f_i, E_{1,\ldots,f_i})|P_i \in \mathcal{A}\}$, where f_i is the number of faces, and E_j is the bounding ellipse for each face region in the photo P_i.

Setting. The setting consists of time and place attributes, indicating when and where the events happened. Reasoning about the time and place information from a single photo remains challenging due to the large size of solution space. While there are a few methods striving for this daunting problem, it seems unaffordable for the lightweight collage application when incorporating extra 3D structure [SDK07] or geographically calibrated image database [HE08]. Fortunately, providing time and place information is possible by reading EXIF metadata from the head of formatted photo file. Almost all the digital cameras, consumer or professional level, enable the storage of precise photographing date into EXIF. So the time attribute t_i for each photo P_i can be easily obtained. On the contrary, the place information is actually obtained by GPS device equipped with the camera, which is much expensive for common users. As a result, the photos taken by cameras without GPS have no place information stored in the EXIF, like the album photos used in this paper. Besides, the location recorded by GPS of camera is usually rough for classifying photos taken in a small area. So we have to resort to other solution for place inference.

Actually, narrative writing emphasizes the expression of events that happened in the same place [Too01]. Thus, instead of absolutely precise location, we adopt the relative scene similarity between photos to identify the place consistency: the photos with similar scenes are likely to be taken at the close place. Judging the similarity between two images is a fundamental problem in image processing, and different feature descriptors are proposed for scene representation. Here, we employ the GIST descriptor [OT01] and color histogram to measure the scene shape and color respectively, both of fast computation. GIST uses the low dimensional spatial envelope to model the shape of scene in the photo, denoted as $g_i = g(P_i)$. And the histogram, defined by $h_i = [b_{i1}, \ldots, b_{iM}]$, models the RGB color distribution with M bins. Then, we get the setting attribute of the photos as $\sigma = \{(t_i, g_i, h_i)|P_i \in \mathcal{A}\}$.

Plot. Plot is the highlight of a single event, from which we

Figure 3: ***Top:*** *Photos are grouped by clustering on their character and setting attributes.* ***Bottom:*** *Sorted on their time attributes, the cluster centers compose the plots. The arrows denote that there are more photos between the adjacent plots.*

can have an instant overview on the characters and setting involved in the event. So we perform the clustering procedure to concentrate the photos by their character and setting attributes. Then, the cluster centers are expected to be the highlighted plots for different events. Here, we use the unsupervised clustering method based on affinity propagation (AP) [FD07] to classify the photos into different groups, and obtain the representative exemplar (i.e., cluster center).

To cluster the photos using AP, we define the similarity between two photos P_i and P_j based on their setting (time and place) attribute as

$$s_{ij} = -exp((g_i - g_j)^2 + w_1 \cdot \chi(h_i, h_j) + w_2 \cdot (t_i - t_j)^2 / m_t^2) \quad (1)$$

where $m_t = \max\{\|t_a - t_b\|, P_a, P_b \in \mathcal{A}\}$, $\chi(\cdot)$ is the Chi-squared histogram distance [PW10], and $w_{1,2}$ are the empirical coefficients to control the magnitude tradeoff ($w_1 = w_2 = 0.2$ in the experiments). The preference of a photo P_i to be exemplar based on its character attribute is defined as

$$s_{ii} = exp(1/(f_i + 1)) \cdot \sum_{P_j \in \mathcal{A}} s_{ij}/N \quad (2)$$

Based on the similarity matrix $S = [s_{ij}]$, AP clustering partitions the photos into different groups, denoted by $\mathcal{A} = \bigcup \mathcal{C}_i$, of which each one has a cluster center photo P_{c_i}. Then, those photos, sorted by their time attributes, composes the plots for the events from the album, i.e., $\pi = \{P_{c_i}\}$ (see Figure 3).

With the definition of $< \gamma, \sigma, \pi >$ as above, we have translated the literary narrative elements to computational attributes. Next, we wish to condense the photos in accordance with their narrative attributes to achieve the story-telling effect on the events.

4.2. Hierarchical construction of narrative structure

Obviously, the plots $\{P_{c_i}\}$ provide necessary clues for looking through the events, from which people can easily have a summary recall on what happened. But sometimes, people like to look in more details at an interested event, i.e., to see the relevant photos besides the plots. So we further explore the plot attributes of the rest photos iteratively (see Figure 4).

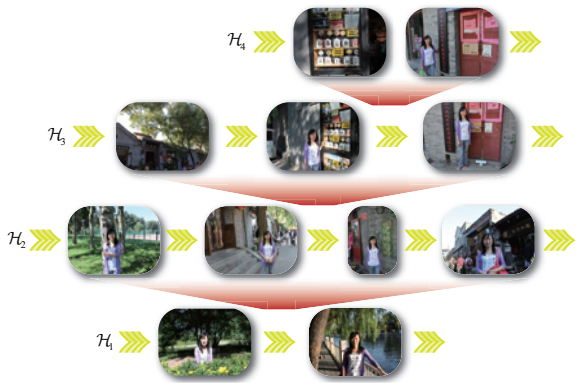

Figure 4: *Hierarchical narrative structure by iteratively clustering on the photos between adjacent plots.*

Assuming the current plots set $\mathcal{H}_k = \{P_{c_j}^k\}_{j=1}^{n_k}$, we denote the set of the rest photos as $\mathcal{T}_k = \{T_j^k\}$, where $T_j^k = \{P_l | c_j < l < c_{j+1}\}$. To be convenient, we add two dummy photos P_{c_0} and $P_{c_{n_k+1}}$ to $\mathcal{H}_k$, with $c_0 = 0$ and $c_{n_k+1} = N+1$. Then, for each subset T_j^k ($j = 0, ..., n_k$), the AP clustering is performed based on the similarity matrix, resulting in new clustering groups $\bigcup \mathcal{C}_j^{k+1}$ and exemplars $\{P_{c_j}^{k+1}\}$ as the plots in the next fine level. We thereafter have the detailed plots sorted by their time attributes, denoted as $\mathcal{H}_{k+1} = \{P_{c_j}^{k+1}\}$. The iteration is performed until all the photos are classified

into the corresponding hierarchical level (see Algorithm **1**). It can be seen that the photos in $\mathcal{H}_{k+1} \setminus \mathcal{H}_k$ provide epitomes on the event that happened between the highlights. Finally, we get the hierarchical structure based on the narrative attributes for the next collage construction task.

Algorithm 1 Construction of narrative hierarchy

Input: Photo album $\mathcal{A}$
Initialize $\mathcal{H}_0 = \{P_0, P_{N+1}\}$, $\Omega = \mathcal{A}$, $k = 0$
while $\Omega \neq \emptyset$
  $\mathcal{H}_{k+1} = \emptyset$
  for each subset $T_j^k \in \mathcal{T}_k$
    applying AP clustering to get new plots $\{P_{c_j}^{k+1}\}$
    sorting $\{P_{c_j}^{k+1}\}$ by time attribute
    $\mathcal{H}_{k+1} = \mathcal{H}_{k+1} \bigcup \{P_{c_j}^{k+1}\}$
  end for
  $\Omega = \Omega \setminus \mathcal{H}_{k+1}$
  $k = k + 1$
end while
Output: Hierarchical narrative $\mathcal{H}_{1,\ldots,K}$

5. Detection of narrative region of interest

Aiming at summarization of the photos in the visually compact form, collage gets used to displaying the interested regions instead of the whole photo [RBHB06]. So we would like to fast compute the saliency that can convey the narrative attributes of the album.

5.1. Narrative saliency detection

Previous methods treat the salient region as the one with acute global/local visual contrast [IKN98, GZMT10, CZM*11], and are applied in many collage works. However, the resulting saliency map lacks the context information of the events from photos, which is clueless to know what happened. To suit the narrative collage construction, we combine the narrative attributes in the saliency computation.

Since plot is the climax expression of characters and setting, it is sufficient to detect the character and setting saliency for narrative region. Given a photo P_i, the character saliency $F(P_i)$ can be simply computed by the face detection and defined for the face region $\bigcup E_{1,\ldots,f_i}$ with value 1, and other region with 0. Considering the setting attribute, we want to detect the salient region that can mark notable place context in the scene, from which people easily recall where the photo is taken. In fact, context saliency is far less explored in image processing, except for the context-aware saliency proposed by Goferman et al. [GZMT10]. They define the context region as the one near the dominant foreground objects. However, this method suffers much slow computation, and fails to identify the significant context region far away from the objects (see Figure 5 (b)). To address this problem, we propose a fast context saliency detection method by combining the concomitant information from a few of relevant photos.

People usually take several photos in the same place for the interested scene, like flowers, statue, and so on. So we have more than one photos $\{P_{ij}\}_{j=1}^{n_r}$ as the references of P_i. Then, the context of a photo can be defined as its salient region as well as possessing inter-similarity in all the reference photos. In other words, region with high context saliency is expected to be distinct in the given photo itself, but frequently appears in other reference photos. The distinction in its own photo can be measured by the global contrast [CZM*11], where the saliency value of pixel $I_k \in P_i$ with color p_i can be defined as

$$Y(I_k) = Y(p_i) = \sum_{l=1}^{n} f_l D(p_l, p_i) \tag{3}$$

where $D(\cdot,\cdot)$ is the color difference in *Lab* space, and f is the frequency of the pixel color occurring in the histogram h_k. Intuitively, $Y(\cdot)$ models the salient region of high contrast to the one with frequent appearance. The frequency with inter-similarity among photos is defined as

$$\tilde{Y}(I_k) = \tilde{Y}(p_i) = \sum_{j=1}^{n_r} \tilde{f}_i^j \log(1 + \tilde{f}_i^j / f_i)/n_r \tag{4}$$

where $\tilde{f}^j$ defines the frequency of the pixel color in the photo P_j. Actually, $\tilde{Y}(\cdot)$ tries to identify the dominant regions frequently appearing in the reference photos, which helps to discover the possible landmark information. Normalizing S and $\tilde{S}$ to $[0,1]$, we get the place context saliency as

$$S(I_k) = S(p_i) = Y(p_i) \cdot \tilde{Y}(p_i) \tag{5}$$

Such combination can suppress the frequent large texture regions, like sky and ground, and interpret the most notable regions that possibly contains important place information.

We use the simplified histogram with $12^3 = 1728$ bins for robust frequency computation like [CZM*11]. Then, we define the narrative saliency (NS) map by combining the face and place context saliency as $F(\cdot) + S(\cdot)$. Figure 5 shows some results by our narrative saliency computation. It is noted that the context saliency might also contain the dominant foreground object that appears frequently in the reference photos (Figure 5 (a)). Compared to previous saliency results [IKN98, GZMT10, CZM*11], narrative saliency can well identify the character and place (setting) attributes of the photo. For example in Figure 5 (b), the Merlion statue can be detected as the place attribute in the saliency map. Besides, NS can be fast computed which favors the processing of a batch of photos. In our implementation, we use the cluster group $\mathcal{C}_i$ as the reference photos of the corresponding photo P_{c_i}, and at the most 6 photos based on the similarity (Equation (1)) for narrative saliency computation.

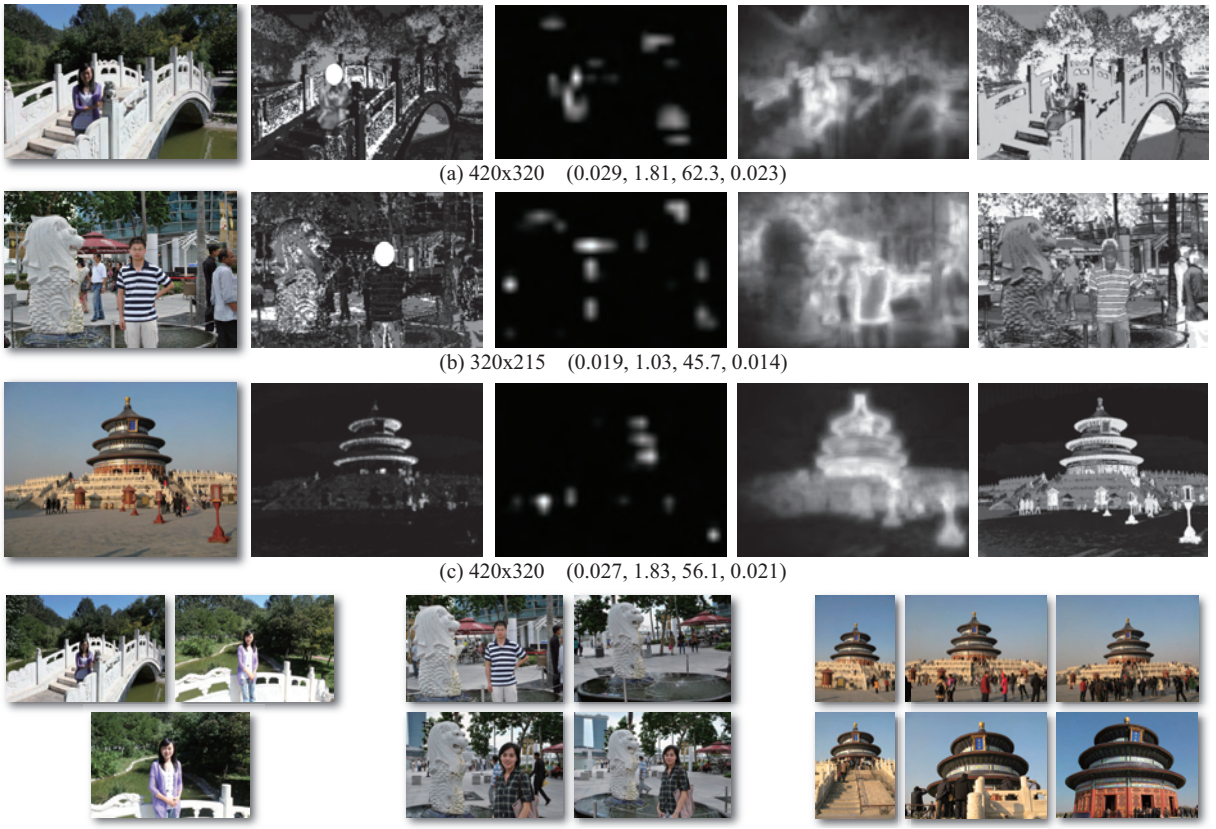

Figure 5: ***Top:*** *From left to right, input photo, saliency map computed by our NS, IT [IKN98], CA [GZMT10] and HC [CZM*11]. The white ellipses in (a) and (b) are the detected face regions. Below each row, the first number denotes the size of photos, and the numbers in the bracket denote running time (in seconds) for each method.* ***Bottom:*** *Reference photos used in our narrative saliency detection for each example.*

5.2. Region of interest cropping

Based on the saliency map, we seek the rectangular region of interest (ROI) that possibly attracts more attention in the photo. Here, we use the dynamic-threshold cropping method [SLBJ03] to search the optimal rectangle that encloses the most saliency with minimal area. Figure 6 shows the ROIs derived from different saliency maps, of which narrative saliency can generate the results that better keep the necessary character and place information after photo cropping. Next, those ROIs will be collected for the narrative collage representation.

6. Hierarchical photo blending

According to the narrative hierarchy $\{\mathcal{H}_k\}$ (Section 4), we now assemble the derived ROIs. To comply with the chronological narration, we align the ROIs one by one in the horizontal direction, rather than stacking them in a rectangular canvas [WSQ*06,RBHB06]. A small overlap δ is prescribed between the adjacent photos, then the start positon of photo

Figure 6: *ROIs are obtained by using different saliency maps, indicated by the four rectangle corners with different colors: NS (red), IT (green), CA (blue) and HC (cyan). The cropping results by NS are shown at the right bottom.*

P_s in the current hierarchy is $x_s = \sum_{j=1}^{s-1} w(P_j) - w(\delta)$, where $w(\cdot)$ is the width, and $w(\delta)$ is set to 50 pixels in our experiments (see Figure 7). Next, we will find the seamless composition for the adjacent photos across the overlap.

There have been many sophisticated approaches to image composition, like the recent hybird image blending [CCT*09]. However in our narrative photo blending, the composition is applied in the simple rectangular overlap, so we adopt the α-Poisson technique [RBHB06] for its practical computation. α-Poisson computes the transparency values based on graph-cut partition by minimizing the image gradient augmented energy, which enables edge sensitive alpha blending in the composition. But instead of graph-cut algorithm, our solution is slightly different in that we compute the shortest path on the graph for partitioning the overlap. Since the overlap is only between two photos, computing the shortest path is $2 \sim 3$ times faster than classical graph-cut algorithm. Then ROIs are blended together with the transparency values to achieve smooth transition.

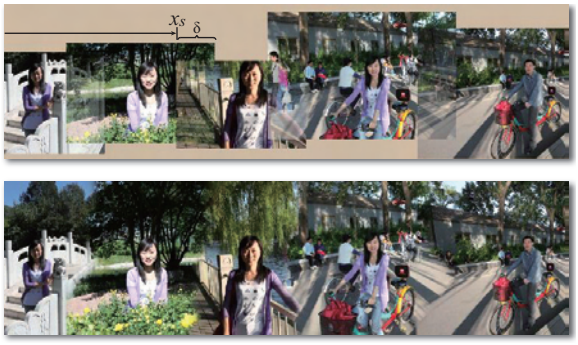

Figure 7: ***Top:*** *Regions of interest are aligned in the horizontal direction with overlap between adjacent photos.* ***Bottom:*** *Seamless collage is obtained by α-Poisson blending.*

7. Experimental results

We have implemented our system and tested it on many photo albums to produce their hierarchical narrative collages. Each album contains hundreds of photos, with the themes including travel, birthday, wedding and so on. We provide a friendly interface to manipulate the hierarchical structure for different levels of plots in the events. User can look into the detailed or abbreviated hierarchical representation by clicking on the navigation bar. For the interested plot, user can also see the corresponding whole photo by clicking on its position in the collage. More details can be seen in the accompanying video. The running time of our algorithm depends on the number of photos in the album (see Table 1) Our experiments were implemented on laptop with 1.44GHz Duo CPU and 4G RAM.

We also made comparison with the reputed AutoCollage [RBHB06] (trial software available online), which arranges the ROIs of photos in accordance with the spatial coherence. This method selects a subset of representative photos in the collage construction, but still suffers slow computation for hundreds of photos. In contrast, our method takes all the photos into the collage, and runs faster to get the summary narration in the hierarchical representation.

Table 1: *Timing statistics (in seconds): #N is number of photos in the collage; #H is the number of hierarchical levels; Ts is total time; Hs is for hierarchical narrative structure construction; Ss is for narrative saliency detection.*

	AutoCollage		Hierarchical narrative collage				
	#N	*Ts*	*#N*	*#H*	*Hs*	*Ss*	*Ts*
Fig. 1	17	23.2	102	6	3.26	5.10	10.9
Fig. 8 (a)	29	42.6	176	7	6.84	8.41	17.1
Fig. 8 (b)	52	85.1	112	7	4.14	7.48	12.7
Fig. 8 (c)	36	54.3	135	5	5.31	8.12	14.9

User study. To informally evaluate the advantage of hierarchical narrative collage, we conducted a user study on the collage effectiveness. We asked 20 participants to judge on the visual quality and summary effect of the collages produced by our system and AutoCollage respectively. Visual quality measures the aesthetic appearance of the whole collage form, scaled from 1 (ugly) to 10 (pretty). Then the participants were encouraged to score on each collage result, and the average score was recorded. Summary effect measures the overview function of collage on presenting the album content. So we asked the participants to narrate the events by only observing the collages in a given time (e.g., 1 minute), i.e., telling who were involved, and when, where and what events happened in the album. Then, we recorded the rate of correct recalling. Figure 9 gives the statistics result on the user study. It can be seen that hierarchical narrative collage can exhibit the appealing form comparable to AutoCollage, while it has a much better command of summarizing the album content. For example in Figure 8 (b), most of participants can correctly tell about one day of a little girl's birthday in the kindergarten, who has a class, and then plays with friends, finally has the celebrating party.

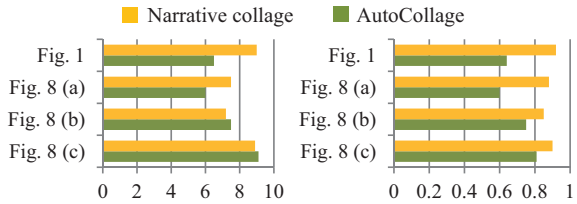

Figure 9: *Statistics of user study on visual quality (left) and summary effect (right).*

Limitations. Although our approach can generate the de-

Figure 8: *Results by AutoCollage [RBHB06] (left) and hierarchical narrative collage (right).*

sired narrative collage, it is not without limitations. Face detection and narrative saliency computation are not always infallible, which would cause the forementioned narrative attributes missing and distort the event recalling through the collage. Because the plot photos are strictly blended in their time order, there might be still transitions artifact between adjacent photos due to their sharp difference in color and texture. Figure 10 shows a failure example with bad narration effect, where the face is missed and visible seam occurs after photo blending in the collage.

Figure 10: *Failure example with missing face and visible seam (enclosed by red rectangles).*

8. Conclusion

We have presented a novel narrative collage that organizes photos in the chronological story-telling way. The literary narrative elements are computed from digital photos, which are used to assemble photos into different hierarchical levels. Our approach compares favorably with previous methods on summarizing the events recorded in the album.

As the future work, we plan to exploit more semantic information to help the narrative organization of photos, like the human pose, emotion, and action. Those high-level information would be helpful for identifying the attributes of photos, which wishes to further improve the narrative effect of collage. Besides, a K-D tree struture [XLJ*09] would be employed to faciliate the hierarchical photo organization.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. Thanks also to Yusha Li and Qi Duan for providing their photo albums in this paper. This work was partly supported by Program for New Century Excellent Talents in University (No. NCET-09-0635).

References

[BCG*07] Battiato S., Ciocca G., Gasparini F., Puglisi G., Schettini R.: Smart photo sticking. In *Adaptive Multimedia Retrieval* (2007), pp. 211–223. 2

[BGSF10] Barnes C., Goldman D. B., Shechtman E., Finkelstein A.: Video tapestries with continuous temporal zoom. *ACM Trans. Graph. 29*, 4 (Jul. 2010), 89:1–89:9. 2

[CCT*09] Chen T., Cheng M. M., Tan P., Shamir A., Hu S. M.: Sketch2photo: internet image montage. *ACM Transactions on Graphics 28*, 5 (Dec. 2009), 124–133. 7

[CLH12] Chen T., Lu A. D., Hu S. M.: Visual storylines: semantic visualization of movie sequence. *Computer & Graphics 36*, 4 (Jun. 2012), 241–249. 2

[CM10] Correa C. D., Ma K. L.: Dynamic video narratives. *ACM Trans. Graph. 29*, 4 (Jul. 2010), 88:1–88:9. 2

[CZM*11] Cheng M. M., Zhang G. X., Mitra N. J., Huang X. L., Hu S. M.: Global contrast based salient region detection. In *Proc. CVPR* (2011), pp. 409–416. 2, 5, 6

[FD07] Frey B. J., Dueck D.: Clustering by passing messages between data points. *Science 315* (Feb. 2007), 972–976. 4

[GTZM10] Goferman S., Tal A., Zelnik-Manor L.: Puzzle-like collage. *Comput. Graph. Forum 29*, 2 (May 2010), 459–468. 2

[GZMT10] Goferman S., Zelnik-Manor L., Tal A.: Context-aware saliency detection. In *Proc. CVPR* (2010), pp. 2376–2383. 2, 5, 6

[HE08] Hays J., Efros A. A.: Im2gps: estimating geographic information from a single image. In *Proc. CVPR* (2008), pp. 1–8. 3

[HZZ11] Huang H., Zhang L., Zhang H. C.: Arcimboldo-like collage using internet images. *ACM Trans. Graph. 30*, 6 (Dec. 2011), 155:1–155:8. 2

[IKN98] Itti L., Koch C., Niebur E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11 (Nov. 1998), 1254–1259. 2, 5, 6

[KEA06] Kim K., Essa I., Abowd G. D.: Interactive mosaic generation for video navigation. In *ACM Multimedia* (2006), pp. 655–658. 2

[MYYH08] Mei T., Yang B., Yang S. Q., Hua X. S.: Video collage: presenting a video sequence using a single image. *Vis. Comput. 25*, 1 (Dec. 2008), 39–51. 2

[OT01] Oliva A., Torralba A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vision 42*, 3 (May 2001), 145–175. 3

[PW10] Pele O., Werman M.: The quadratic-chi histogram distance family. In *ECCV* (2010), pp. 749–762. 4

[RBHB06] Rother C., Bordeaux L., Hamadi Y., Blake A.: AutoCollage. *ACM Trans. Graph. 25*, 3 (Jul. 2006), 847–852. 1, 2, 5, 6, 7, 8

[RKKB05] Rother C., Kumar S., Kolmogorov V., Blake A.: Digital tapestry. In *Proc. CVPR* (2005), pp. 589–596. 1, 2

[SDK07] Schindler G., Dellaert F., Kang S. B.: Inferring temporal order of images from 3D structure. In *Proc. CVPR* (2007), pp. 1–7. 3

[SLBJ03] Suh B., Ling H. B., Bederson B. B., Jacobs D. W.: Automatic thumbnail cropping and its effectiveness. In *ACM symposium on User interface software and technology* (2003), pp. 95–104. 6

[Too01] Toolan M.: *Narrative: a critical linguistic introduction*. Routledge, 2001. 1, 3

[VJ01] Viola P., Jones M.: Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR* (2001), pp. 511–518. 3

[WSQ*06] Wang J. D., Sun J., Quan L., Tang X. O., Shum H. Y.: Picture collage. In *Proc. CVPR* (2006), pp. 347–354. 1, 2, 6

[XLJ*09] Xu K., Li Y., Ju T., Hu S. M., Liu T. Q.: Efficient affinity-based edit propagation using K-D tree. *ACM Trans. Graph. 28*, 5 (Dec. 2009), 118:1–118:6. 9