Monet: A System for Reliving Your Memories by Theme-based Photo Storytelling

Yue Wu*, Xu Shen*, Tao Mei†, *Senior Member, IEEE*, Xinmei Tian, *Member, IEEE*,
Nenghai Yu, Yong Rui, *Fellow, IEEE*

***Abstract*—With the ever-increasing use of smartphones and digital cameras, people are now able to take photos anywhere and anytime. Most of these photos simply end up stored in the cloud without further interaction. This occurs because we lack intelligent services to organize these personal photos well. Therefore, there is an urgent need for such a system to enable people to relive their memories by turning their photos into stories. This paper presents a storytelling system named *Monet*, which automatically creates interesting stories from personal photos by mimicking cinematic knowledge based on a set of predesigned editing styles. The system consists of two stages: *photo summarization*, which selects a subset of the "best" photos to represent a photo collection, and *story remixing*, which generates a stylish music video from the selected photos. During photo summarization, photos are grouped into events based on multimodal features (time and location). The "best" photos are then selected according to visual quality, event representativeness, and diversity. The second stage, story remixing, automatically selects an appropriate theme-dependent editing style based on the photo content. Each selected photo is converted to a video clip by applying a virtual camera with appropriate motions. A series of video effects, color filters, shapes, and transitions are then applied to the video clips according to cinematic rules. The generated video is finally multiplexed with a music clip to generate the story. Evaluations show that our system achieves superior performance to state-of-the-art photo event detection and story generation systems.**

***Index Terms*—Personal photos, photo selection, storytelling, cinematic grammar.**

I. Introduction

With the ever-increasing use of mobile phones and cameras, people are now able to take photos anywhere and anytime. It is estimated that about one trillion photos were taken in 2015, and 4.7 trillion photos will be stored in 2017[1]. Browsing or sharing this enormous volume of photos is boring and time consuming, because people lack intelligent media services to help organize them well. Therefore, most personal photos just simply end up with being unused and stored in the cloud. However, these personal photos are records of people's personal experiences and memories from their daily lives. They are produced in a massive scale, but rarely consumed. As a result, there is an urgent need for a system to weave these personal photos into interesting stories to help people relive their memories.

We have witnessed firsthand the challenges in producing interesting and memorable stories from personal photo collections. First, these massive collections are usually disordered, which makes photo browsing and sharing tedious. However, photos are not usually taken randomly, but taken at special moments and cover different events. Being able to group personal photos into events is sorely needed for storytelling purposes. Second, since most people have no professional photography skills, many of the personal photos suffer from quality degradation and content redundancy. Selecting a representative subset of photos is a critical step in storytelling. Third, we need to exhibit the selected and grouped photos in attractive ways to provide a good user experience. In our system, we tell stories in the music video form. Since photos are taken in different types of scenarios - corresponding to different styles - the system should be able to automatically choose appropriate editing styles for the music video. Finally, to make the story more attractive, video editing elements such as effects, shapes, color filters, and transitions, should be considered when rendering the static photos. From the professional video editing perspective, these video editing elements are specific to movies. Therefore, how to discover video editing grammars, design style-specific video editing elements, and apply them to a system are challenging - yet essential - components of storytelling.

Some existing systems already try to solve the aforementioned challenges. Magisto[2] and Animoto[3] are two online services that can generate music videos from user provided photos. However, they do not create stories from personal photos directly, lacking functions such as event segmentation and photo selection. Besides, with these two services, users must specify the editing styles manually. Other services such as Sewing Photos [1], Tiling SlideShow [2], Microsoft OneDrive[4], and Nokia Story Teller[5] can perform event segmentation and photo selection to some degree, but they achieve

∗equal contribution.

†corresponding author.

T. Mei and Y. Rui are with Microsoft Research, Beijing 100080, China (e-mail: {tmei, yongrui}@microsoft.com).

Y. Wu and N. Yu are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: wye@mail.ustc.edu.cn; ynh@ustc.edu.cn).

X. Shen and X. Tian are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, University of Science and Technology of China, Hefei 230027, China (e-mail: shenxu@mail.ustc.edu.cn; xinmei@ustc.edu.cn).

[1]http://resourcemagonline.com/2014/12/infographic-there-will-be-onetrillion-photos-taken-in-2015/45332/

[2]http://www.magisto.com

[3]https://animoto.com/

[4]https://onedrive.live.com

[5]http://www.windowsphone.com/en-us/store/app/lumia-storyteller

only limited performance because they omit many user-selected key photos, leading to an incomplete summarization of personal photos. Photo2Video [3] is one of the pioneering works to generate videos from photos. However, it does not make use of editing styles; its generated videos are more like slide shows of still photos with few camera motions involved. This is also the main drawback of systems such as Sewing Photos [1] and Tiling SlideShow [2]. Moreover, most of the preceding systems do not incorporate cinematic grammars. The generated videos have obvious machine-crafted traces.

To build an end-to-end system that can create attractive stories from personal photos, we propose a system named *Monet*, which automatically summarizes personal photos and narrates them in the form of interesting and memorable music videos based on cinematic grammars and predesigned movie styles. Figure I shows an overview of our entire system. The system comprises two key steps: *personal photo summarization* and *photo story remixing*. In photo summarization, we propose a novel generative multimodal model to first segment personal photos into events. Then, our system selects the best photos from each event to summarize the photo collection, based on a set of criteria, e.g., aesthetic quality and representativeness. In photo story remixing, to narrate these key photos, we use a classification model to assign each of them to an editing style designed by professional video editors, followed by a refinement step to group the photos of the same editing style. A specific camera motion is selected for each key photo to generate a motion photo clip. To improve the level of interest and smoothness, style-based visual effects are applied according to cinematic grammar and the visual content. Finally, the video is synchronized with a music to generate a music video.

Our main contributions can be summarized as follows:

1) We propose a model for automatic movie style assignment. This is a key component for automatic movie story creation yet overlooked in existing systems.
2) We provide predesigned style-based templates based on cinematic rules, which makes our system quite effective at assigning visual effects.
3) Our system is the first to incorporate a mixture of multiple modalities - audio, images, and video - to produce fancy movies.
4) The part of our system that selects the best photos can considerably save users' time when browsing and selecting photos. This feature also makes our system the first fully automatic video generation system for cloud-based storytelling.

The rest of the paper is organized as follows. We discuss related works in Section II. The framework of the Monet system is presented in Section III. The photo summarization capability of our system is described in Section IV. Section V describes details concerning the photo story generation. We conduct experiments to evaluate the effectiveness of our proposed system and report the results in Section VI. We conclude by discussing potential improvements to the system and describing our future work in Section VII.

TABLE I
COMPARISON OF PHOTO STORY TELLING SYSTEMS

	Magisto	Animoto	Google+	Monet
camera-motion analysis	+	-	-	+
video analysis	-	-	-	+
face detection and recognition	+	-	+	+
scene analysis	+	-	+	+
objects recognition	-	-	+	+
music analysis	+	-	-	+
photo grouping and selection	-	-	+	+
designed styles	+	-	-	+
color tuning	+	+	+	+
SNS and cloud storage	-	-	+	+

II. Related Work

Personal photo summarization and storytelling have been popular topics in recent years, including numerous work in the form of papers and tools. We compare the key features of these systems and our Monet in Table I. All these works have three main components: photo event segmentation, key photo selection, and photo storytelling.

A. Photo Event Segmentation

Generally, personal photos are tagged with timestamps and locations when captured. Accordingly, we can use the information to segment photos into different events by choosing appropriate event boundaries. Platt *et al.* proposed setting one hour or adaptive thresholds as the time gap between two adjacent events [4]. Graham *et al.* extended the method, using intra-cluster rates and inter-cluster time gaps to refine the original clusters [5]. Gargi proposed marking sharp local increase in capture frequency as the start of an event and a long interval with no capture as the end [6]. In [7], Matthew *et al.* detected event boundaries by applying confidence scores, dynamic programming or Bayes Information Criterion (BIC) to the photo similarity matrices. More generally, the event segmentation problem can be formulated as a clustering problem. Loui and Svakis proposed grouping photos using a 2-class K-means algorithm and then refining the clusters by checking the color similarity of photos [8]. In [9], photos were assigned to different clusters by hierarchical agglomerative clustering. A Hidden Markov Model with learned parameters was used for clustering in [4]. Mei *et al.* solved this problem by utilizing a Gaussian Mixture Model with time, location and content features [10]. Xu *et al.* further extended this method with texture and deep learning features [11].

B. Key Photo Selection

Many research studies and products that address key photo selection have been produced in recent years. In the research community, key photo selection mainly relies on photos' representativeness [7], [10], [11], [12]. In [7], the first captured photo of an event was considered as a key photo. Mei *et al.* proposed selecting photos with the maximal posterior probabilities as representative photos [10]. In [12], the key photo was determined by the mutual relationship between near-duplicate photo pairs in photo clusters. Xu *et al.* incorporated the popularity of photos according to event importance and

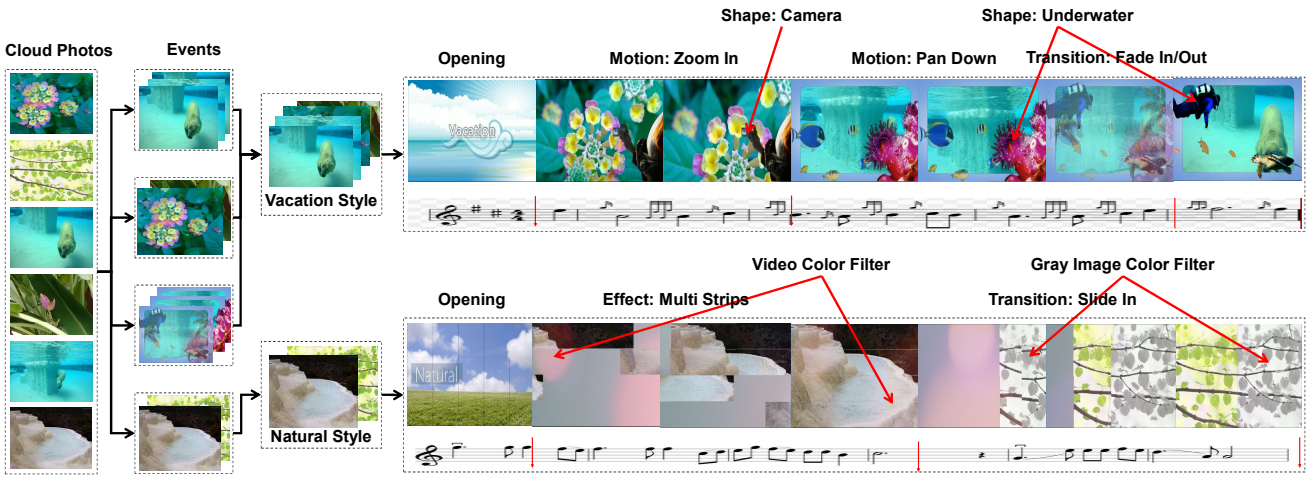


Fig. 1. A brief view of the *Monet* system. The system consists of two stages: *personal photo summarization*, which selects a subset as the "best" photos to represent a photo collection, and *photo story remixing*, which generates a stylish music video from the selected photos.

intra-event similarity for key photo selection [11]. The problem of these approaches is that they tend to select only a small proportion of very high quality and representative photos but omit many photos that users regard as key photos.

C. Photo Storytelling

The most common way to narrate stories in personal photos is through videos. Yang *et al.* presented a way to generate a fascinating layout for photos [13]. Jun-Cheng *et al.* proposed presenting photos in a tile-like slide show, synchronizing them to the pace of background music [2]. Kuo *et al.* focused on assigning smooth transitions between photos in a slide show [1]. Photos were transformed to motion photo clips using camera motion in [3]. The final output video was rendered by connecting these motion clips with transitions and incidental music. All these tools apply few visual effects and fail to take movie styles into account.

III. Framework of Monet

Figure 2 shows the framework of our Monet system. Cloud photos are first segmented into meaningful events. Then, the best photos are selected from these events to achieve a good summarization. Given a selection of best photos, Monet then automatically chooses the most suitable styles for the events. Next, based on predesigned cinematic grammars, Monet combines camera motion, music, and visual effects to generate a video clip for every photo. Finally, it adds transitions between the video clips of each style to connect them into complete fancy movies. The details of each component will be further explored in Sections IV and V.

IV. Personal Photo Summarization

Personal photo summarization is conducted via three steps: event segmentation, photo filtering, and key photo selection. Event segmentation groups photos into events. Photo filtering removes duplicated or low-quality photos. Key photo selection selects high-quality photos with high representativeness and high event uniformity.

A. Event Segmentation

Statistically, people tend to capture photos in bursts. To clearly present the event segmentation, the term "event" is clarified as follows:

- *Event*: An event is a photo taking session where users take photos to record the photo-worthy moments in a specific scene within a relatively short period.

As a result, photos of the same event are typically close in both time and location. Therefore, each photo $x_i \in X = \{x_1, x_2, \cdots, x_N\}$ belongs to one latent semantic event $e_j \in E = \{e_1, e_2, \cdots, e_K\}$, where N and K are the total numbers of photos and events in the photo collection X, respectively. Accordingly, the probability of photo x_i belonging to event e_j can be formulated as $p(x_i|e_j)$, where $x_i = (x_{i,1}, x_{i,2})$, $x_{i,1}$ is the time ($\mathcal{T}$), and $x_{i,2}$ is the GPS ($\mathcal{G}$). Photo x_i is assigned to event e_j if $p(e_j|x_i)$ is the maximum a posteriori probability.

Assuming that distributions of time and GPS are independent given event e_j, the likelihood probability $p(x_i|e_j)$ can be computed as follows:

$$p(x_i|e_j) = \prod_{l=1}^{2} p(x_{i,l}|e_j) = p(\mathcal{T}_i|e_j)p(\mathcal{G}_i|e_j). \quad (1)$$

The probability of each metadata $x_{i,l}$ given event e_j follows a Gaussian distribution, as shown below:

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

To obtain the distribution of each event is to learn the model parameters $\Theta = \{\delta_{j,l}, \mu_{j,l}\}$ in the Gaussian Mixture Model (GMM). We maximize the log-likelihood of the join distribution and formulate the objective function as follows:

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

where $p(x_i|e_j,\Theta)$ is computed by Eq. (1) and $p(e_j)$ is the priori probability of event e_j. We utilize EM to learn the

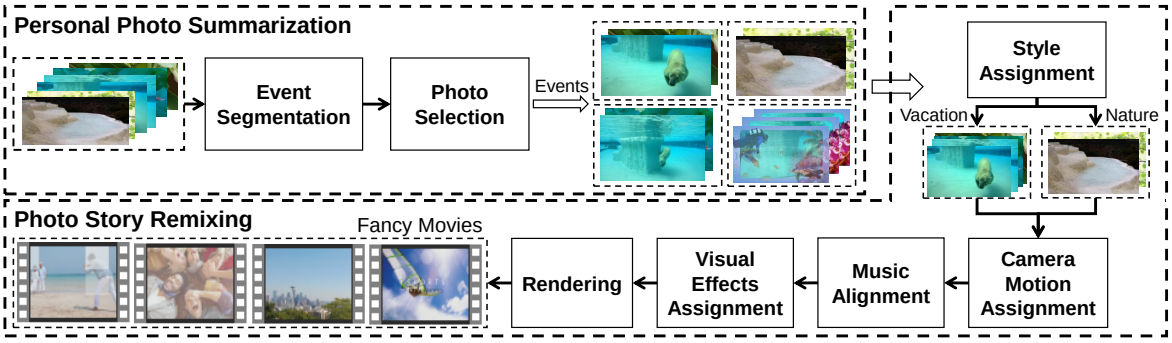


Fig. 2. Framework of our Monet system.

best parameters. Before the EM training process, the model parameters are initialized by the clustering centers of K-means. To determine the number of events K, we generate a series of candidate segmentations by applying EM training to multiple values of K. The best model is selected by minimum description length (MDL) as described in [10]. The most time consuming part in this step is to use EM algorithm to estimate the GMM model. The complexity of each update is $O(2N + KN^2)$.

B. Photo Filtering

Because most personal photos are taken by people without professional photography skills, we must filter out duplicate and low-quality photos to generate high-quality stories.

Quality Filtering. Photo quality can be degraded due to many factors including underexposure, overexposure, homogeneity, blurring, and so on. In our implementation, we evaluate photo quality with 43-dimensional handcrafted features. The features are extracted from the following aspects:

- *Darkness* (1D) and *brightness* (1D) [14]. The proportions of underexposed and overexposed pixels.
- *Blurriness* (1D) [15] and *blurriness difference* (1D). The blurriness of a photo and the difference in blurriness before and after Gaussian blurring.
- *Sharp (1D)* [15], *sharpness* (1D) [16], *simplicity* (1D) [17], *intensity contrast* (1D) [18], *dynamic range* (1D) [18], and *depth of field* (1D) [19]. All these features are commonly used global features (CGF) for photo quality assessment.
- *HSV distribution* (12D) [20]. Firstly, the photo is converted into the HSV color space. Then, nonuniform color quantization is used to quantize "hue" into an 8-bin histogram and "value" into a 4-bin histogram.
- *Best block feature* (7D), *worst block feature* (7D), and *subject block feature* (7D). We find that sometimes only part of a personal photo is bad, causing it to be considered as a bad photo. However, global features are not sufficiently discriminative in this case. So we propose segmenting photos into 5 blocks(top-left, top-right, bottom-left, bottom-right, and the center block). We extract CGF features from the blocks with the highest contrast, lowest contrast and the center block.

We created a dataset with $10,361$ good photos and $3,134$ bad photos. All these were collected from people's personal photos and manually labeled as "good" or "bad" by volunteers. Accordingly, we train a binary SVM classifier using the 43-dimensional features on this dataset. Photo quality can be estimated by the output of this SVM model. Quality filtering removes the photos that have a quality score below a predefined threshold (i.e., the low-quality photos).

Duplication Filtering To better summarize the photo collection, we need to detect duplicate photos and choose only one from the duplicates. We adopted the local image descriptor proposed in [21] to represent each photo by a 64-dimensional feature vector, in which all the elements are integer numbers. The similarity of two photos is estimated by the number of identical integers in the feature vector. If the similarity value exceeds a threshold, they are marked as duplicates and only the one with the highest aesthetic quality will be retained.

C. Key Photo Selection

To select a subset of photos as key photos that are representative of a photo collection, we should consider three factors: *aesthetic quality*, *representativeness*, and *uniformity*.

Aesthetic Quality. To construct a professional movie, photos with high visual aesthetics (quality score, denoted as $\mathcal{Q}$) are preferable. We adopt the model in [22] to assess the aesthetic quality of photos.

Representativeness. The representativeness of photos can be evaluated from two aspects. 1) Event importance. When people are interested in a certain event, they tend to take more photos (and *vise versa*). Therefore, for an event e_i, if the number of photos in this event is n_i and the number of photos in the whole collection is N, the importance of e_i is $\mathcal{EI}_i = \frac{n_i}{N}$. 2) Diversity. We want to select photos from an event that have the most diversity. We extract *time* ($\vec{t} \in R^1$), *location* ($\vec{l} \in R^2$), and *color histogram* ($\vec{c} \in R^{64}$) features from the photo. The distance between photo x_i and photo x_j is defined as follows:

$$d_{ij} = dist(\vec{t}_i, \vec{t}_j) + dist(\vec{l}_i, \vec{l}_j) + dist(\vec{c}_i, \vec{c}_j), \quad (4)$$

where $dist(\vec{a}, \vec{b}) = exp(-\frac{||\vec{a}-\vec{b}||^2}{\sigma^2})$.

For photo x_i, the diversity is defined as: $\mathcal{D}_i = \sum_j d_{ij} I_j$. Here, I_j is an indication function: $I_j = 1$ if photo x_j is selected as the key photo, else $I_j = 0$. Accordingly, the representativeness of photo x_i can be defined as:

$$\mathcal{R}_i = \mathcal{EI}_i + \mathcal{D}_i \quad (5)$$

Uniformity of Events. Quality and representativeness select photos at the event level. To obtain a more comprehensive summarization of the photo collection, event uniformly is also an important concern. Therefore, we propose to utilize event entropy as a measurement of uniformity. If the time gaps between a given photo with the previous and the next selected photo are $t_{i,i-1}$ and $t_{i,i+1}$, respectively, the entropy of photo x_i is defined as follows:

$$\mathcal{E}_i = t_{i,i-1} \log t_{i,i-1} + t_{i,i+1} \log t_{i,i+1} \tag{6}$$

We compute the score of a photo by the linear combination of *quality*, *representativeness*, and *entropy*:

$$S_i = a\mathcal{Q}_i + b\mathcal{R}_i + c\mathcal{E}_i \tag{7}$$

where a, b, c are non-negative weights subject to $a+b+c = 1$, and they are selected to achieve the best accuracy on users' groundtruth best photo selections. Since the calculation of representativeness and entropy are dependent on the overall selection of photos, it's very difficult and time consuming to obtain the global optima for the key photo selection problem. Instead, we use greedy search to select photos with the highest scores as best photos and reduce the complexity to $O(N)$.

V. Photo Story Remixing

The typical filmmaking consists of six steps. 1) Choose an editing style and the corresponding background music. 2) Add some contextual photos based on the style if needed. 3) Design motions for every static photo to make a motion clip. 4) Apply visual effects such as motions, shapes, color filters, texts, and so on to every clip. 5) Select transitions between adjacent clips. 6) Combine all the clips with transitions and create an opening and an ending to generate the final video. Primarily , designers determine the style and visual effects for photo collections based on the semantic content.

Following the same steps involved in professional filmmaking, we first analyze the semantic meaning of photos. Based on the semantic content, we assign a related movie style to them, generate motion clips based on suitable camera motions, and select specific visual effects and transitions.

A. Photo Semantic Analysis

According to the filmmaking process, semantic features or concepts are very important for style selection, motion pattern determination, and visual effects design. In our Monet system, we adopt the photo tagging method in [23] to extract the semantic features. For each photo x_j, we have 112 possible semantic tags (t_i) and a corresponding set of probabilities $P(t_i|x_j)$. In our implementation, these probabilities will be used to predict the likelihood that a photo belongs to a specific movie style, as described in Section V-B.

The tags are grouped into 20 categories that represent the most frequent features in personal photos, including *animal, building, dark, food, indoor, outdoor, object, people, group people, crowd people, plant, sky, text, mountain & rock, ocean & beach, flower, grass, road, wheel*, and *sculpture*. Besides, faces are highly important for personal photos. Therefore, we use a face detector to detect the number, gender, size, and location of faces in photos. Then, the faces in a photo are classified as "face profile", "1 or 2 large faces", "1 or 2 small faces", "3 to 5 large faces", "3 to 5 small faces", "group of small faces", and "group of large faces". Gender information in photos is classified into "single", "single male", "single female", "two females", "two males", "couple", and "crowd" categories. All these semantic features will be applied during subsequent steps including the generation of motion patterns, addition of visual effects, and determination of color filters (these steps are detailed in Section V-E).

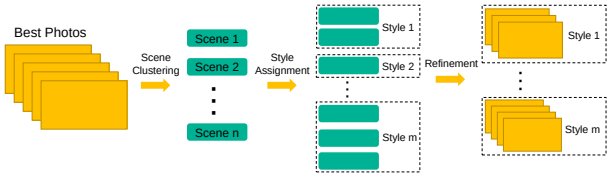


Fig. 3. Workflow of movie style assignment. Photos are first clustered into different scenes. Then a SVM classifier is used to assign these scenes into appropriate styles based on their semantic features. In the refinement step, scenes with the same style are merged to generate a single fancy movie.

B. Automatic Style Assignment

This section describes how to automatically cluster photos to different scenes and how to assign appropriate movie styles to these scenes, as shown in Figure 3.

Scene Clustering. As we talked with designers, people take photos to record the photo-worthy moments in a scene. In other words, the photos are not only explicitly organized by timestamps and locations, but also implicitly correlated by the scenes. Our system is to create music videos for the scenes to summarize personal photo collections. To bridge the correlation between photos and scenes, each photo is represented by a semantic probability vector $\mathbf{p}_i = (P(t_1|x_i), P(t_2|x_i), \cdots, P(t_N|x_i))$ as described in Section V-A. We use the Affinity Propagation algorithm [24] to cluster the photos, whose complexity of $O(N^2)$. Each cluster is regarded as a scene.

Style Assignment. In our system, different scenes are presented in different styles to make the storytelling appealing and smart. To determine which style should be assigned to a scene, we propose to train a multi-class SVM model on all the styles based on the semantic features of photos. We invite designers to define a set of most commonly appeared semantic terms to represent the styles in personal photos. For example, *love, couple, sweetheart, wedding*, and *honey* are terms related to the "love" style. To make these semantic terms computable and representable, we use these terms to search for at least $5,000$ related personal photos from Flickr for each style. The photos are represented as probability vectors as in *Scene Clustering*. We train a multi-class SVM classifier to distinguish styles. Accordingly, the probability or confidence (which is the output of the multi-class classification model) that a particular photo belongs to style $\mathcal{S}_j$ is denoted by $P(\mathcal{S}_j|\mathbf{p}_k)$.

If there are M photos in scene i, i.e. $scene_i = \{\mathbf{p}_1^i, \mathbf{p}_2^i, \cdots, \mathbf{p}_M^i\}$, the style probability of scene i can be

calculated by:

$$P(scene_i, \mathcal{S}_j) = \sum_{k=0}^{M-1} P(\mathcal{S}_j|\mathbf{p}_k^i)/M. \quad (8)$$

Consequently, scene i is assigned to the style with the maximum probability.

Refinement. In the style assignment step, different scenes may be attributed with the same style. We merge photos of these scenes to create a single music video.

C. Photo Clip Generation

There are three key steps involved in generating a clip from a photo [3].

- Step 1: Key-frame selection. To simulate camera motions, we need to select key frames to be used as full shots, medium shots, and close-ups, using the method described in [3].
- Step 2: Key-frame sequencing. We need to determine the order of key frames selected in Step 1. Based on cinematic rules, we adopt the predefined 14 framing schemes described in [3] to generate the order of selected key frames.
- Step 3: Motion generation. The output clip is generated by applying specific camera motions to the key frame sequence selected in Step 2. We build a suitability matrix as detailed in [3]. For a series of photos, we assign motion patterns by maximizing both the overall framing scheme suitability and the motion pattern distribution uniformity.

D. Music Analysis

In our Monet system, music is sampled at 8 kHz. We detect "strong" onsets of music to analyze the rhythm of the music, as described in [3]. The duration and transition of the generated video clips are then aligned with the music by detecting cut points according to the rhythm and amplitude of the music. A cut point denotes the time point where the video should transit from one clip to another. The length of each video clip will be adjusted according to the cut points. Based on the onsets, we adopt the method proposed in [25] to detect cut points for better visual-aural relevance [26], so that the switching frequency of video clips is compatible with the rhythm of the music. Besides, the cut point detection strategy also avoids video clip switching at speaking or singing intervals of the music, which is quite different from existing methods which switch at the strong onsets.

E. Video Composition

To generate a professional movie, a motion clip accompanied by music is not sufficient. We need to add visual effects (VE) including *effects*, *shapes*, *color filters*, and *transitions* to clips to make them visually appealing and smooth [26]. To achieve this goal, we first design visual effect templates for every movie style. Then we assign these effects to the generated clips based on some computable filmmaking grammar. The final movie is generated based on the video clips and the selected visual effects.

1) Template Design: For each style, we first design many template effects, shapes, color filters, and transitions appropriate to images with different content. Each VE is only suitable for images with specific content. Figure 4 shows some examples of the designed visual effects.

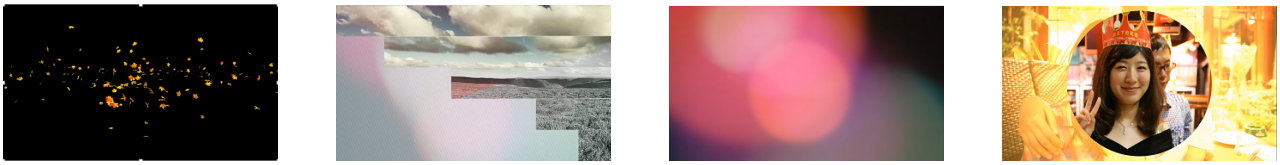
2) Constructing the Grammar for Styles: This paragraph describes how to assign appropriate VEs for video clips. Different VEs may express different feelings and may only be suitable for specific semantic features. For example, the multiple stripe effect in the nature style (shown in Figure 4 (b)) is suitable for photos taken outdoors with plants or sky, but is not suitable for photos containing people because no one wants his/her face to be split into stripes. Similarly, the leaf shape in the "*Original*" style (Figure 4 (a)) is only suitable for photos that contain leaves or grass. The circular transition in the "*Party*" style (Figure 4 (d)) is admirably suited for photos with a single item of focus in the center of the photo - especially a single large face. Moreover, some VEs are only suitable for video clips with specific motion patterns, especially the transitions. In summary, different VEs have different "suitability" levels for different semantic features and motion patterns. We define the *suitability grammar* for all the effects, shapes, color filters, and transitions for each style. The grammar of effects follows the syntax below (grammar of shapes and transitions use similar syntaxes).

- **Root Element <Grammar>**: The root element contains style information and child elements of effects, shapes, and transitions.
- **Effect Element <Effect>**: The effect element describes the suitability score of an effect with respect to different semantic features and motions. Each effect element contains one or more "*condition*" sub-elements and an optional "*percent*" sub-element. The "*condition*" sub-element can contain two types of sub-elements: feature and score. If a "*feature*" sub-element starts with a plus ("+"), it is suitable for this semantic feature. Starting with a minus ("-") means that this effect is not suitable for this semantic feature. The "*score*" sub-element indicates the suitability score when the feature starts with "+". If a "*percent*" sub-element is present, it indicates the expected proportion of occurrence of this effect in all selected effects.

If a clip contains unsuitable features, the suitability score is set to 0; otherwise, the suitability score is the sum of the scores of all the suitable features. According to the VE grammar, we can obtain a suitability score of every VE with respect to every clip and the expected proportion of occurrence of some of the VEs.

Grammar for color filters. Even for the same clip, varying changes in light and color distributions express different feelings. Therefore, we design different color filters for different styles. Our grammar for color filters uses the syntax below.

- **Video Filter Element <ColorFilter>**: This element describes information about a video color filter. The "path" attribute shows the location of the filter video. The "opacity" controls opacity of the filter when blended with video clips. "overlay" shows the way to blend the filter



(a) leaf shape (b) multi stripes effect (c) sunshine color filter (d) circular transition

Fig. 4. Examples of the designed shapes, effects, transitions, and color filters. In the leaf shape, yellow leaves float in the video slowly, which awakes people's memory about original and old time. The multi stripes effect presents each stripe of the photo sequentially, which is an interesting way to show the contents of nature. Blending photos with sunshine color filter help people to experience the feeling of sunshine in the "*natural*" style. The circular transition gradually shows people's face, which is a very effective way to attract viewer's attention to the main person or object in photo.

video with video clips. "minPercent" and "maxPercent" control the minimum and maximum percentage of video clips this filer can be applied to.

- **Image Filter Element <ImageFilter>**: This element describes information about an image color filter. All the settings are the same as in video filter, except that "path" is replaced by the "name" attribute, which indicates the type of predefined image color filters. Besides, this element contains the "*condition*" nodes as in the effect element.

The suitability scores between video clips and color filters are calculated as follows. We first extract the saliency map [27]. The suitability scores are calculated as the negative correlation of saliency maps between video clips and color filters. In other words, color filters should not distract viewers' attention from the major subject of photos.

3) Optimize the Assignment Problem: The selection of VE and color filters can be formulated as optimization problems. Suppose there are N_c clips with each clip denoted as c_i. There are also N_e VEs or color filters, each of which is represented by VE_j. The selection of VE and color filters is to find the selection matrix $X = (\vec{x_1}, \vec{x_2}, \cdots, \vec{x_{N_C}})$, where $\vec{x_j}$ is a N_e dimensional binary vector with only one element equal to 1. Because different VEs and color filters have different expected occurrence rate (percent), we need to obtain their expected occurrence times (n_j^*) first. If the expected percent of VE_j is p_j, it is obvious that $n_j^* = p_j N_c$. When p_j is missing, we calculate the percentage according to their suitability with the video clips. Suppose the suitability of video clip c_i and VE_j is S_{ij}. The overall suitability of VE_j is calculated as:

$$S_j = \sum_{i=0}^{N_c-1} S_{ij}(1 - I_j), \tag{9}$$

where I_j is an indication vector. I_j equals 1 when p_j is specified, otherwise 0. The expected occurrence rate of VE_j is computed as follows:

$$p_j = \frac{S_j}{\sum_{k=0}^{N_e-1} S_k(1 - I_k)} p^*, \tag{10}$$

where $p^* = 1 - \sum_{k=0}^{N_e-1} p_k I_k$ is the remaining percent of VEs not defined in the grammar.

The selection of visual effects can be formulated as maximizing the overall suitability subject to occurrence rates:

$$X^* = arg \max_{X_{ij}} \sum_{j=0}^{N_e-1} d_j \frac{\sum_{i=0}^{N_c-1} X_{ij} S_{ij}}{n_j}, \tag{11}$$

where $d_j = exp(-\frac{(n_j - n_j^*)^2}{2})$ is the *percentage* score to evaluate the gap between the assigned occurrence rate and the expected rate for VE_j.

However, only maximizing the suitability often leads to boring results when the same VE is applied to adjacent clips. To avoid this problem and to keep the order of photos for "telling" stories, we add constraints on the distribution uniformity of visual effects.

Suppose that VE_j appears n_j times in the N_c clips. We expect VE_j to appear uniformly. Therefore, the expected interval between two adjacent clips to which VE_j is assigned is $\frac{N_c - n_j}{n_j}$. If the interval between the kth and $(k+1)$th assigned clips of VE_j is δ_k, the uniformity score u_{jk} is defined as:

$$u_{jk} = exp(-\frac{(\delta_k - \frac{N_c - n_j}{n_j})^2}{2}). \tag{12}$$

Accordingly, the overall uniformity score of all VEs is

$$U = \sum_{j=0}^{N_e-1} \frac{\sum_{k=0}^{n_j-1} u_{jk}}{n_j^* - 1}. \tag{13}$$

Consequently, considering *suitability*, *occurrence rate* and *uniformity*, the VE and color filter selection problem is formulated as solving the following objective function:

$$X^* = arg \max_{X_{ij}} \sum_{j=0}^{N_e-1} (d_j \frac{\sum_{i=0}^{N_c-1} X_{ij} S_{ij}}{n_j} + \lambda \frac{\sum_{k=0}^{n_j-1} u_{jk}}{n_j^* - 1}). \tag{14}$$

We use backtracking to find the optimal selection of VEs for the above optimization problem. Though the complexity of backtracking is $O(N_e^{N_c})$ in theory, the computation can be significantly accelerated by effective pruning of search paths.

4) Rendering: After the motion patterns, music clips, effects, shapes, color filters, and transitions have been determined for the entire photo collection, we are able to construct the final movie containing a series of professional music videos, telling stories summarized from the photo collection.

VI. Evaluation

As far as we know, there is no other systems that can both automatically conduct photo summarization and create a movie from the summarized photos. To evaluate the effectiveness of Monet, we compare the personal photo summarization and the photo story remixing of Monet separately.

In our experiments, all the photos are collected from photo albums uploaded to our Monet cloud by users. To evaluate the

TABLE II
INFORMATION OF PERSONAL PHOTO COLLECTIONS

Dataset	User 1	User 2	User 3	User 4	User 5	User 6
#Photos	1080	481	496	564	702	866
#Events	95	32	40	107	28	58
#Best Photos	206	145	108	375	66	285

performance of the photo summarization algorithms, we ask the uploaders to label the groundtruth of their own photos. We compare the segmentation and photo selection results by uploading the same photos to existing systems/applications. For video generation, users are invited to recommend photos for each movie style from their own album. Then, we randomly select one suggestion for each style and upload these photos and the corresponding music to each competitor. The generated videos are used to perform a subjective user study.

A. Event Segmentation & Best Photo Selection

We invited six users to share photos from their mobile phones and cameras taken during the past two years. All the photos had accurate time stamps but only some had GPS information. Users were asked to group all the photos into meaningful events. For each event, they were asked to choose 1 to 6 photos which they thought can best represent the event. These groups and best photos are used as groundtruth to evaluate personal photo segmentation and photo selection. Detailed information of the photos is listed in Table II.

As described in [7], the measures of precision, recall, and F-score are adopted to evaluate the event segmentation performance. Here, precision indicates the proportion of correctly detected event boundaries:

$$\text{Precision}_{seg} = \frac{\text{\#correctly detected boundaries}}{\text{\#deteced boundaries}}. \quad (15)$$

Recall represents the proportion of true boundaries detected over the groundtruth:

$$\text{Recall}_{seg} = \frac{\text{\#correctly detected boundaries}}{\text{\#groundtruth boundaries}}. \quad (16)$$

The F-score measures the comprehensive performance:

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (17)$$

Accuracy of best photo selection is defined as:

$$\text{Accuracy}_{best} = \frac{\text{\#correctly selected key photos}}{\text{\#groundtruth selected photos}}. \quad (18)$$

The comparison of event segmentation is presented in Table III. It's easy to observe that Monet works better than PhotoTOC and TEC on both Precision and Recall, thus a higher F-score. The high precision and high recall indicate that Monet not only detects obvious event boundaries, but also tries to detect those hard event boundaries. As to the best photo selection, Monet achieves an accuracy of 0.68, which is higher than existing applications (e.g., Google+, OneDrive, and Nokia StoryTeller). In other words, Monet can find out more accurate key photos that can best represent the events.

TABLE III
EVENT SEGMENTATION PERFORMANCE

Method	Precision	Recall	F-score
PhotoTOC[4]	0.50	0.71	0.59
TEC[7]	0.39	0.54	0.45
Monet	0.85	0.72	0.78

TABLE IV
SUBJECTIVE EVALUATION OF PHOTO STORY PRESENTATION

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Average
Magisto	**5.83**	**5.67**	5.33	**5.78**	5.67	**5.22**	5.05	5.54	5.51
Animoto	4.5	5	4.33	5.11	4.28	3.8	4.17	4.56	4.47
Monet	5.78	5.60	**5.83**	5.67	**6**	4.94	**5.28**	**5.72**	**5.60**
TS Show	-	3.62	3.22	3.05	2.9	2.5	2.56	2.83	2.59

B. Photo Story Remixing

Objective evaluation of photo story presentation is difficult. We conduct subjective user study to evaluate our work. There are 10 movie styles in the Monet system. We ask users to recommend their photos for each style from their photo albums. As a result, different styles may have photos collected from different number of users. For comparison, we randomly select photos from one user for each movie style. Then, we upload these photos and the corresponding style music to Monet, Animoto, Magisto, and Tiling Slide Show [2]. Thus, there are 40 resulted music videos in total. In Animoto, photos are displayed one by one with incidental music in the background and fancy effects. Animoto adds transitions between each photo but does not simulate camera motion. In Magisto, photos are displayed with background music, fancy effects, transitions, and camera motions. In Tiling Slide Show, photos are displayed in a tile-like manner, coordinated with the pace of background music.

Twenty users (12 males and 8 females) were asked to rate all the 40 videos. The users ages ranged from 22 to 28. None of the users were professional movie makers or designers. Videos of the same style were presented in the same web page in random order. Users were asked to provide rating scores ranging from 1 to 7 to show how satisfying each video was (higher scores are better) from the following aspects:

- Question 1: Are the camera motions professional (like a professional movie)?
- Question 2: How smooth are transitions between clips?
- Question 3: How attractive are the visual effects over the entire video?
- Question 4: Do you think the transitions between video clips match well with the music tempo?
- Question 5: Does the entire video look like a professional movie?
- Question 6: How well do you think the video tells an interesting story?
- Question 7: How likely are you to share the generated video to your friends in your social networks?
- Question 8: What is your overall level of satisfaction with this video?

The average satisfaction scores of all the questions above are listed in Table IV.

The first four questions are about the detailed aspects of story remixing. Compared with the slide show of static photos produced by the Tiling Slide Show and the very simple motion videos from Animoto, Monet and Magisto are both rated as superior because they apply professional camera motions, various effects, shapes, and transitions. Monet performs slightly worse than Magisto in camera motions and in the selection of transitions. One reason is that Monet chooses motions based on the content of the current photo. Transitions are selected based on the previous and the following clip. The cut points of video clips are also selected with a Markov assumption [25]. All the aforementioned elements are determined by local information only. We suspect Magisto utilizes some global optimization algorithms or global refine steps to create a fluent and consistent video, so that the whole video is more unified. Nevertheless, Monet stills gets very similar rating compared with Magisto. For the attractiveness of visual effects, Monet gets a much better score, which verifies the effectiveness of our style templates and our video composition algorithms.

The latter four questions are about overall ratings of different systems. Even though Magisto gets slightly higher scores on motions and transitions, Monet still looks the most professional. This again indicates the superiority of the design and selection of visual effects of Monet. Since Magisto creates more fluent videos with smoother transitions and motions, it gets better scores on telling stories, which requires more strictly on the fluency. Nevertheless, viewers are more willing to share Monet videos with their friends. We also get the highest scores on the overall satisfaction evaluation and the highest average score. This phenomenon shows that: 1) visual effects are the most key aspect to generate a professional and satisfying video; 2) on other aspects, Monet works comparably well with competitors like Magisto. Generally, Monet obtains the best degree of satisfaction in person photo storytelling.

VII. Conclusion

In this paper, we present a fully automatic personal photo storytelling system that works in the cloud to generate fancy movies. The system achieves user experience ratings superior to those of state-of-the-art storytelling systems. It generates movie videos based on cinematic grammars and predesigned movie styles. We first segment cloud photos into meaningful events and select the best photos based on photo content and time distributions. Following the process and principles of filmmaking, photos are assigned to specific styles. Then, based on these assigned styles, we select specific motion patterns, music, and visual effects for each photo to generate video clips accompanied by music. The final movie output is constructed by connecting all the individual video clips, applying transitions between them.

There are a number of possible improvements to this personal photo storytelling system. The current system aligns music cut points only with motion clips in time sequence. However, if we could align more important clips with more intense portions of the music, the user experience could be further improved. Moreover, we could search for all the photos uploaded by different users for a particular event or that match a given theme to compose a more comprehensive and detailed story [28]. Based on this shared global story, we could create personalized stories for each user.

Acknowledgment

This work is supported by NSFC No.61572451, National Natural Science Foundation of China (No.61371192), Youth Innovation Promotion Association CAS CX2100060016, and Fok Ying Tung Education Foundation. We would like to give our special thanks to Guilong Hu, Junjie Yu, Ye Yang, and Yunzi Qian for their kind help on discussing the cinematic grammar and designing the style templates.

References

[1] T.-H. Kuo, C.-Y. Tsai, K.-Y. Cheng, and B.-Y. Chen, "Sewing photos: smooth transition between photos," in *Advances in Multimedia Modeling*. Springer, 2011, pp. 73–83.

[2] J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J.-L. Wu, "Tiling slideshow," in *ACM Multimedia*, 2006, pp. 25–34.

[3] X.-S. Hua, L. Lu, and H.-J. Zhang, "Photo2Video: A System for Automatically Converting Photographic Series Into Video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 7, pp. 803–819, 2006.

[4] J. C. Platt, M. Czerwinski, and B. A. Field, "PhotoTOC: automatic clustering for browsing personal photographs," in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 1, 2003, pp. 6–10.

[5] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, "Time as essence for photo browsing through personal digital libraries," in *Proceedings of the second ACM/IEEE-CS Joint Conference on Digital libraries*, 2002, pp. 326–335.

[6] U. Gargi, "Modeling and Clustering of Photo Capture Streams," in *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2003, pp. 47–54.

[7] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal Event Clustering for Digital Photo Collections," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, pp. 269–288, 2005.

[8] A. Loui and A. Savakis, "Automatic image event segmentation and quality screening for albuming application," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2000, pp. 1125–1128.

[9] B. Gong and R. Jain, "Segmenting Photo Streams in Events Based on Optical Metadata," in *International Conference on Semantic Computing (ICSC)*, 2007, pp. 71–78.

[10] T. Mei, B. Wang, X.-S. Hua, H.-Q. Zhou, and S. Li, "Probabilistic Multimodality Fusion for Event based Home Photo Clustering," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2006, pp. 1757–1760.

[11] X. Shen and X. Tian, "Multi-modal and multi-scale photo collection summarization," *Multimedia Tools and Applications*, 2015.

[12] W.-T. Chu and C.-H. Lin, "Automatic selection of representative photo and smart thumbnailing using near-duplicate detection," in *ACM Multimedia*, 2008, pp. 829–832.

[13] X. Yang, T. Mei, Y. Xu, Y. Rui, and S. Li, "Automatic Generation of Visual-Textual Presentation Layout," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 2, p. 33, 2016.

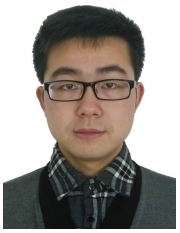
[14] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox, "A Semi-automatic Approach to Home Video Editing," in *the 13th Annual ACM Symposium on User Interface Software and Technology*, 2000, pp. 81–89.

[15] H. Tong, "Blur detection for digital images using wavelet transform," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2004, pp. 17–20.

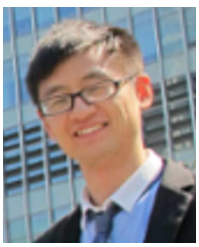
[16] Y. Ke, X. Tang, and F. Jing, "The Design of High-Level Features for Photo Quality Assessment," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE Computer Society, 2006, pp. 419–426.

[17] Y. Luo and X. Tang, "Photo and Video Quality Evaluation: Focusing on the Subject," in *Proceedings of the 10th European Conference on Computer Vision: Part III*, ser. ECCV, 2008, pp. 386–399.

[18] T. Xia, T. Mei, G. Hua, Y.-D. Zhang, and X.-S. Hua, "Visual quality assessment for web videos," *Journal of Visual Communication and Image Representation*, vol. 21, pp. 826–837, 2010.
[19] Z. Dong and X. Tian, "Effective and efficient photo quality assessment," in *2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC*, 2014, pp. 2859–2864.
[20] Ch.Kavitha, D. Rao, and Dr.A.Govardhan, "Image Retrieval Based On Color and Texture Features of the Image Sub-blocks," *International Journal of Computer Applications*, vol. 15, pp. 33–37, 2011.
[21] S. A. Winder and M. Brown, "Learning local image descriptors," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
[22] Z. Dong and X. Tian, "Effective and efficient photo quality assessment," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2014, pp. 2859–2864.
[23] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui, "Relaxing From Vocabulary: Robust Weakly-Supervised Deep Learning for Vocabulary-Free Image Tagging," in *2015 IEEE International Conference on Computer Vision (ICCV)*, December 2015, pp. 1985–1993.
[24] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[25] Y. Wu, T. Mei, Y.-Q. Xu, N. Yu, and S. Li, "MoVieUp: Automatic Mobile Video Mashup," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 1941–1954, Dec 2015.
[26] T. Mei, X.-S. Hua, L. Yang, and S. Li, "Videosense: towards effective online video advertising," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 1075–1084.
[27] Y.-F. Ma and H.-J. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing," in *ACM Multimedia*, 2003, pp. 374–381.
[28] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 38, 2014.



Yue Wu received his B.E. degree in 2012 from the University of Science and Technology of China (USTC), Hefei, China. He is currently a Ph.D. candidate in the Department of Electronic Engineering (EEIS), USTC. He worked as an intern in Microsoft Research, Beijing, China, from July 2012 to June 2012 and from February 2014 to March 2015, respectively. His research interests include multimedia, computer vision, machine learning, and data mining.



Xu Shen received his bachelors degree in electrical engineering (2012) from the University of Science and Technology of China, China. He is currently a Ph.D. candidate in the Department of Electronic Engineering (EEIS), USTC. His research interests mainly include multimedia, computer vision and deep learning.

Tao Mei (M07-SM'11) is a Senior Researcher with Microsoft Research, Beijing, China. His current research interests include multimedia analysis and retrieval, and computer vision. He has authored or co-authored over 100 papers in journals and conferences, 10 book chapters, and edited four books. He holds over 15 U.S. granted patents and 20+ in pending. Tao was the recipient of several paper awards from prestigious multimedia journals and conferences, including IEEE Communications Society MMTC Best Journal Paper Award in 2015, IEEE Circuits and Systems Society Circuits and Systems for Video Technology Best Paper Award in 2014, IEEE Trans. on Multimedia Prize Paper Award in 2013, and Best Paper Awards at ACM Multimedia in 2009 and 2007, etc. He was the principle designer of the automatic video search system that achieved the best performance in the worldwide TRECVID evaluation in 2007. He is an Editorial Board Member of IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing, Communications, and Applications, Machine Vision and Applications, and Multimedia Systems, and was an Associate Editor of Neurocomputing, a Guest Editor of eight international journals. He is the General Co-chair of ACM ICIMCS 2013, the Program Co-chair of ACM Multimedia 2018, IEEE ICME 2015, IEEE MMSP 2015 and MMM 2013, and the Area Chair for a dozen international conferences. Tao received B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a Senior Member of the IEEE and the ACM, and a Fellow of IAPR.

Xinmei Tian (M'13) is an Associate Professor in the CAS Key Laboratory of Technology in Geospatial Information Processing and Application System, University of Science and Technology of China. She received the B.E. degree and Ph.D. degree from the University of Science and Technology of China in 2005 and 2010, respectively. 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

Nenghai Yu received the B.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1987, the M.E. degree from Tsinghua University, Beijing, China, in 1992, and the Ph.D. degree from University of Science and Technology of China, Hefei, China, in 2004. He has been on the faculty of the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) since 1992, where he is currently a professor. He is the executive director of the Department of Electronic Engineering and Information Science, and the director of the Information Processing Center at USTC. His research interests include multimedia security, multimedia information retrieval, video processing and information hiding. He has authored or co-authored over 130 papers in journals and international conferences. He has been responsible for many national research projects. Prof. Yu and his research group won the Excellent Person Award and the Excellent Collectivity Award simultaneously from the National Hi-tech Development Project of China in 2004. He was the co-author of the Best Paper Candidate at ACM Multimedia 2008.

Yong Rui is currently Deputy Managing Director of Microsoft Research Asia (MSRA). A Fellow of IEEE, IAPR and SPIE, and a Distinguished Scientist of ACM, Rui is recognized as a leading expert in his research areas. He is the recipient of the IEEE Computer Society 2016 Technical Achievement Award, IEEE Trans. Multimedia 2015 Best Paper Award, and ACM Multimedia 2009 Best Paper Award. He holds 60 US and international patents. He has published 16 books and book chapters, and 200+ referred journal and conference papers. Ruis publications are among the most cited 17,000+ citations and his h-Index = 55. Dr. Rui is the Editor-in-Chief of IEEE Multimedia Magazine, an Associate Editor of ACM Trans. on Multimedia Computing, Communication and Applications (TOMM), and a founding Editor of International Journal of Multimedia Information Retrieval (IJMIR). He was an Associate Editor of IEEE Trans. on Multimedia (2004-2008), IEEE Trans. on Circuits and Systems for Video Technologies (2006-2010), ACM/Springer Multimedia Systems Journal (2004-2006), and International Journal of Multimedia Tools and Applications (2004-2006). He also serves on the Advisory Board of IEEE Trans. on Automation Science and Engineering. He is an Executive Member of ACM SIGMM, and the founding Chair of its China Chapter. Dr. Rui received his BS from Southeast University, his MS from Tsinghua University, and his PhD from University of Illinois at Urbana-Champaign (UIUC).