Photo Sequences of Varying Emotion: Optimization with a Valence-Arousal Annotated Dataset

CHRISTOS MOUSAS, CLAUDIA KROGMEIER, and ZHIQUAN WANG, Department of Computer Graphics Technology, Purdue University

Synthesizing photo products such as photo strips and slideshows using a database of images is a time-consuming and tedious process that requires significant manual work. To overcome this limitation, we developed a method that automatically synthesizes photo sequences based on several design parameters. Our method considers the valence and arousal ratings of images in conjunction with parameters related to both the visual consistency of the synthesized photo sequence and the progression of valence and arousal throughout the photo sequence. Our method encodes valence, arousal, and visual consistency parameters as cost terms into a total cost function while applying a Markov chain Monte Carlo optimization techniques called simulated annealing to synthesize the photo sequence based on user-defined target objectives in a few seconds. As our method was developed for the synthesis of photo sequences using the valence-arousal emotional model, a user study was conducted to evaluate the efficacy of the synthesized photo sequences in triggering valence-arousal ratings as expected. Our results indicate that the proposed method synthesizes photo sequences in which valence and arousal dimensions are perceived as expected by participants; however, valence may be more appropriately perceived than arousal.

CCS Concepts: • **Information systems** → **Image search**; *Search interfaces*; **Search interfaces**; • **Multimedia and multimodal retrieval** → Visualization toolkits; • **Human-centered computing** → **Visualization toolkits;**

Additional Key Words and Phrases: Valence, arousal, photo sequence, visual consistency, optimization, simulated annealing

ACM Reference format:
Christos Mousas, Claudia Krogmeier, and Zhiquan Wang. 2021. Photo Sequences of Varying Emotion: Optimization with a Valence-Arousal Annotated Dataset. *ACM Trans. Interact. Intell. Syst.* 11, 2, Article 16 (June 2021), 19 pages.
https://doi.org/10.1145/3458844

1 INTRODUCTION

Digital cameras can be found in many devices such as laptops, tablets, and smartphones. The availability of such devices has dramatically increased the number of pictures taken by users on a daily basis. Additionally, we now have access to vast collections of digital photos, which makes

All authors contributed equally to this research.
Authors' address: C. Mousas, C. Krogmeier, and Z. Wang, Department of Computer Graphics Technology, Purdue University, West Lafayette, IN 47907, USA; emails: {cmousas, ckrogmei, wang4490}@purdue.edu.

2160-6455/2021/06-ART16 $15.00
https://doi.org/10.1145/3458844

the process of browsing and creating photo sequences (including photo stories) ever more time-consuming and increasingly difficult. Accessing large image datasets can make image selection and sequencing a daunting task for those interested in generating products such as photo albums, photo stories, slideshows, calendars, and more.

Unsurprisingly, there are various commercial solutions (e.g., Smilebox,[1] Magisto,[2] NCH,[3] Adobe Spark,[4] Pholody,[5] Kizoa,[6] Clideo[7]) that take a set of user-defined photos as input and either automatically or semi-automatically create photo slideshows. Synthesizing photo sequences to create slideshows is based primarily on rules defined by the developer [74], the metadata-based systems [64, 83], or the visual features analyses [1, 80]. However, the selection of the most appropriate images based on predefined rules has several limitations [52]. Among them, rule-based methods often fail to select appropriate photos for the desired collection and, in most cases, group photos in a product-appropriate manner [52, 75]. This drawback can be eliminated using methods that process annotated data, metadata, and visual features. Specifically, by representing images through a number of variables, the system can organize and automatically synthesize photo sequences based on those variables and, therefore, create a vibrant and meaningful visual product [11, 17, 40, 43, 72, 75].

In most existing methods, only limited information extracted from images (e.g., chronology, upload order) is used to synthesize the photo sequence. Therefore, even if the system can propose a photo sequence, a user most likely needs additional time to edit and correct the photo sequence due to the poor quality of the system. Users may want to create photo sequences that: (1) follow a loose chronological order, (2) track specific emotional patterns, and (3) maintain visual consistency. For example, in a given image dataset, one might want to gradually change the emotional content of the photo sequence from pleasant to unpleasant while also maintaining visual consistency across these images.

In this article, we use a dimensional model of affect called the valence-arousal model [61]. Measured with scales, valence represents the level of pleasure while arousal represents the level of excitement. Considering that the valence-arousal model accounts for much of the variance found in self-reported affective states [61], we consider this model acceptable for identifying the emotional attribute of a photo in a concrete manner. While there exist other models of affect, such as the Pleasure-Arousal-Dominance model [48], which incorporates valence and arousal as well as a dominance dimension, we chose the two-dimension valence-arousal model, as we thought providing two affective dimensions might be simpler for users to understand to rate photos in a meaningful way.

Given a valence-arousal annotated image dataset, our system automatically synthesized photo sequences based on valence and arousal targets as well as progression towards those targets throughout the photo sequence. In achieving a visually consistent photo sequence, we extracted visual information (dominant color and brightness) from the images, and then, based on color objectives, our system optimized photo sequences that fulfilled user-defined targets. Thus, our system selects images from a dataset and automatically synthesizes the requested photo sequence that satisfies both emotional and visual goals (see Figure 1). Therefore, it can be said that the developed method provides direct control over the synthesized photo sequence in a quantitatively measurable fashion. Our process expands on previously published work in the area of image selection and

[1] https://www.smilebox.com/.
[2] https://www.magisto.com/.
[3] https://www.nchsoftware.com/.
[4] https://spark.adobe.com/.
[5] https://www.pholody.com/.
[6] https://www.movavi.com/.
[7] https://clideo.com/.

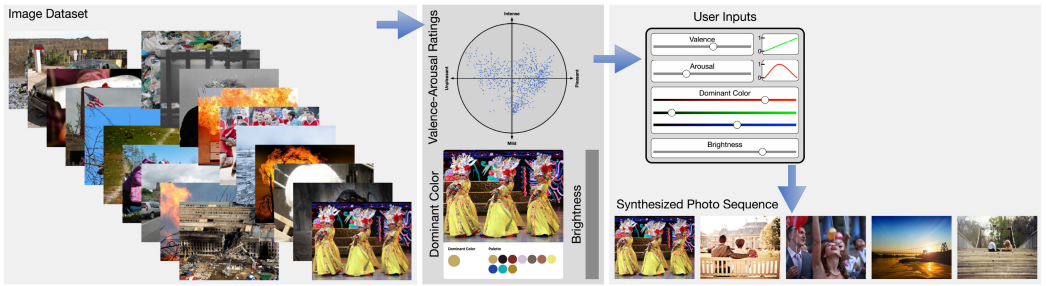

Fig. 1. Using a dataset of annotated images with valence and arousal ratings, our system optimizes a photo sequence that fulfills user-defined targets in terms of valence, arousal, dominant color, image brightness, and valence and arousal progression throughout the photo sequence.

synthesis of photo sequences using both annotated emotional rating and visual information extracted from the images. Our method primarily applies to annotated photos in the user's collection.

The contributions of this project include:

- A method to automatically synthesize a photo sequence, giving control to the user over the emotional values elicited from the photo sequence.
- Introduction of the valence-arousal model for optimizing the photo sequence.
- Validation of the method through a user study.

The remainder of this article is organized as follows. Section 2 provides an overview of previous research. Section 3 introduces the proposed method, the photo dataset used in the study, and the preprocessing of the dataset. Then, Section 4 presents the problem formulation. Section 5 explains the terms of the developed costs, and Section 6 describes the process of optimization. The user study conducted to evaluate our method is presented in Section 7. The limitations of our method are discussed in Section 8. Finally, conclusions and recommended directions for future research are presented in Section 9.

2 RELATED WORK

This article concerns methods of visualization in the creation of photo sequences. In the past, several methods have been developed in the hopes of improving the process of synthesizing photo sequences. Among them, techniques for photo organization and presentation were used extensively to cluster images based on time, faces, and background features [31, 32, 41]. These features were then used to semantically extract images with similar characteristics from a large dataset to synthesize photo sequences.

Chu and Lin [9] developed a method for image summarization that uses temporal characteristics of images to create photobooks. Sinha et al. [67] also focused on a summarization technique that uses a method called "duplicate detection" to synthesize photobooks. A multidimensional scaling algorithm has also been used in previous projects [10, 60]. These algorithms have treated the dissimilarities across images in a high dimensional space and then attempted to estimate them in a low dimensional output space. Other methods have explored the use of ranking-based techniques. Jing et al. [27] developed a technique that relies on examining the structure of visual links across images and text queries and then ranks the returned images based on input text queries. Liu et al. [42] used similarities between the tags and the image content to rank the tag queries, after which their system recommends and synthesizes a set of images based on user queries. Finally, Gao et al.

[18] developed a semi-automatic user-controlled method that incorporates image selection and theme-based image grouping to create photobooks.

Other recent research has focused on providing methods that enable automatic or semi-automatic storytelling using media collections based on data extracted from social networks [54, 58, 64]. Social networks provide an abundant source of metadata and contextual information related to the content. Automated methods require either minimal or no effort from the user. For example, by using photo albums that are available online, Obrador et al. [54] developed a method that can learn social context associated with users to help users synthesize photo albums for sharing. Among other advantages, their work has taken into consideration the aesthetics of facial and image characteristics. By ranking the photos, it was possible to identify the best one to use. The method developed by Saini et al. [64] attempted to create meaningful photo stories by extracting metadata retrieved from images as well as extracting information from a user's social network of friends. Sadeghi et al. [63] developed a method of image selection that utilized context to synthesize photo collections for vacations.

Semi-automatic methods require the user to interact with the system to provide additional help in the process of photo sequence synthesis. Specifically, such semi-automated methods focus on providing a supportive user interface to help the user find and select assets that will be used in the final photo sequence. Among others, Chen et al. [6] describes a human-centered interface for creating story narratives. This system uses storytelling as a way to group photos in a theme-based clustering algorithm, and then seamlessly tile multiple photos in each theme. Another cluster-based method was developed by Karlsson et al. [29]. This method first groups photos into clusters, and then helps the user view the assets in the collection based on a user-controllable time scale. Previous research has distinguished relevant from irrelevant images based on keywords. The above methods are fundamentally different from the photo sequence synthesis method introduced in this article, in which the photo sequence is synthesized based on user-defined emotional and visual targets.

Previous studies have shown that the sequential effects resulting from visual stimuli affect one's perception of the photographs [5, 30, 55, 70, 71]. Moreover, extensive familiarization with either complexity or simplicity has resulted in a heightened visual preference for that attribute [69]. Therefore, when synthesizing sequences of images, visual consistency of the photo sequence should be considered [26, 78, 82]. Locher et al. [44] stated that there should be visual balance across photo sequences, and Hübner et al. [25] indicated that the aesthetics of an image depend significantly on the perceptual balance of the elements that compose the image. Although there are numerous definitions of visual consistency, aesthetics, and balance due to the subjective nature of such elements, it is not always easy to process such factors when synthesizing photo sequences.

Methods that evaluate visual consistency across images have also been explored in several recent studies. Among others, machine learning techniques that train neural networks have been used extensively [35, 46, 47]. Other methods also include models for measuring visual consistency based on a number of parameters. Ngo et al. [53] introduced a model to estimate the balance, equilibrium, symmetry, and sequence of a visual image, while Geigal et al. [19] introduced additional factors to measure visual quality including balance, emphasis, chronology, and unity. Yang et al.'s [81] propositions are based on the principle that the synthesized output should maintain a symmetrical balance and conform to the golden ratio. In assuming that images that correspond to a specific search query within a specific search engine are often visually similar, some methods have considered visual consistency based on data retrieved via search engine results [14, 23]. Metrics for consistency have also been proposed. In the past, Huang et al. [24] developed a metric for visual consistency that considered several image-related features such as color, texture, and edges. Last, Gonzalez and Lapeer [21] developed a visual consistency metric based on image brightness.

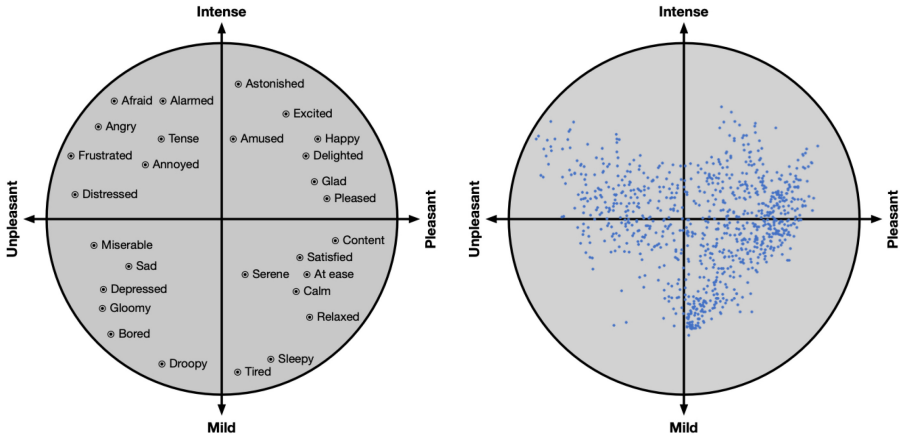

Fig. 2. The valence-arousal circumplex model of affect, as defined by Russell [61]. The horizontal axis represents valence, and arousal is represented by the vertical axis (left). The distribution of the data used in this study was based on Russell's circumplex model (right). In our system, the user is able to choose a valence-arousal combination, and the system will synthesize a photo sequence that fulfills the user's desired valence-arousal combination.

Although this work focuses primarily on synthesizing image sequences using the valence-arousal emotional model, our work has also considered the use of dominant color and image brightness to improve visual consistency in the synthesized output. We assigned a set of emotional and visual consistency cost terms to a total cost function, and we optimized the photo sequence based on user-defined emotional and visual consistency targets. While we understand users may also like to create photo sequences with semantic meaning, we leave this area for future work, allowing our current system's primary contribution to be the optimization of target affect. Emotions are a central part of storytelling [2]; therefore, we believe emotion is a logical first step for our system. To the best of our knowledge, no other published study has focused on the synthesis of a photo sequence using emotionally annotated data and visual features extracted from images.

3 PRELIMINARIES

In this section, we present the valence-arousal emotional model, the dataset that was used for the implementation and testing procedure, and the process for extracting visual information from the images that were later used to form the total cost terms.

3.1 The Valence-Arousal Emotional Model

Many models describe the emotional or affective states of humans [15]. Among them is the emotional model proposed by Russell [61], the emotional diagram model developed by Plutchik [57], the **Pleasure, Arousal, Dominance (PAD)** emotional state model proposed by Mehrabian [49] and finally, the three-dimensional cube model developed by Lövheim [45]. In our study, we used Russell's circumplex model of affect as a point of reference, from which emotions in the valence and arousal axes are distributed (see Figure 2). This valence-arousal model [61] has been widely used in the field of psychology. Valence represents the pleasantness of an emotional stimulus, ranging from unpleasant to pleasant. Arousal is the intensity of emotion provoked by a stimulus, ranging from low to high. Our intent in using the model was to allow users to easily choose valence and arousal targets, as a means of asking our system to optimize photo sequences based on their predefined preferences.

3.2 Image Dataset

There are many publicly available datasets with emotional ratings that could have been used. Among them is EmoMadrid [3], an open-access database consisting of 813 photos. CAP-D [51] contains 526 photos from four databases in which discrete emotions categorize images. Affectnet [50] contains more than one million facial images drawn from three major search engines using 1,250 emotion-related keywords in six languages. Our study used the **open affective standardized image set (OASIS)** [39], an open-access dataset containing 900 color images that depict a broad spectrum of themes including humans, animals, objects, and scenes. The dataset also contained normative ratings obtained from an online study of 822 participants, based on the two affective dimensions (valence and arousal). Both the broad spectrum of themes, much like those found in a personal photo collection, and the valence-arousal ratings made the OASIS dataset ideal for our study. The distribution of valence and arousal ratings of the OASIS dataset is shown in Figure 2.

3.3 Extraction of Visual Information

The OASIS dataset provided the valence and arousal ratings for each image. However, we needed additional information to form our method for rating visual consistency. Therefore, we preprocessed the image dataset to extract additional information (the dominant color and brightness of the images) deemed to be critical for optimizing the visual consistency of a synthesized photo sequence. The dominant color and brightness of the images were chosen, since prior psychology and graphic design research [12, 59, 65, 68, 73] has shown that such parameters express consistency across visual information.

Dominant Color. In synthesizing visually consistent photo sequences, we used the color features of the images. The dominant color of the image is a component that could help us establish the visual connection across images. A number of dominant color extraction techniques have been developed in the past years [33, 66, 79] that take into account different color spaces (e.g., RBG, HSL, CIELAB). We considered an RGB-based method, since the RGB-based dominant color extraction can benefit from the fast computation [4, 20, 76]. To extract the dominant color (the RGB values of a color), we used the median cut [22] algorithm from the largest cluster. Specifically, the algorithm first defines the number of clusters (dominant colors that should be obtained) and then the method estimates and returns the dominant color to be used during the optimization process. Note that for all examples presented in this article, a single dominant color was used. However, additional dominant colors could also be utilized based on the proportion of each dominant color in the image.

Image Brightness. We also considered image brightness an essential factor in visual consistency. We wanted to synthesize image sequences that did not differ significantly in intensity. To compute the mean brightness of an image, we first converted all images to grayscale. Then, for each image, we calculated the average grayscale value of all pixels.

4 PROBLEM FORMULATION

In using a dataset of images annotated with valence and arousal ratings, the objective of our method was to synthesize a photo sequence that fulfilled user-defined targets for emotional and visual consistency. We started by defining a photo sequence as $S = [p_1, p_2, \ldots, p_N]$, consisting of a user-defined number of pictures (N) generated by our system and assembled in sequential order, in which a picture $p_i \in S$ corresponded to any possible image. The method synthesized photo sequences based on emotional and visual consistency, and prior costs assigned to a total cost function. In synthesizing a photo sequence S, we defined an overall cost function $C_{Total}(S)$ as

$$C_{Total}(S) = C_{Emo} w_{Emo}^T + C_{Vis} w_{Vis}^T + C_{Prior} w_{Prior}^T, \quad (1)$$

where $C_{Emo} = [C^V_{Emo}, C^A_{Emo}]$ is a vector of emotional costs, and $w^T_{Emo} = [w^V_{Emo}, w^A_{Emo}]$ are the corresponding weights to the emotional costs. The C^V_{Emo} and C^A_{Emo} terms, respectively, were used for evaluating the valence and arousal of the synthesized photo sequence. The $C_{Vis} = [C^C_{Vis}, C^B_{Vis}]$ is a vector of visual consistency costs, and $w^T_{Vis} = [w^C_{Vis}, w^B_{Vis}]$ are weights that correspond to the visual consistency cost terms. The term C^C_{Vis} evaluates the dominant color in the synthesized photo sequence, and the C^B_{Vis} is the brightness of the synthesized photo sequence. The term $C_{Prior} = [C^{VP}_{Prior}, C^{AP}_{Prior}, C^R_{Prior}]$ is prior costs associated with the synthesized photo sequence S, and $w^T_{Prior} = [w^{VP}_{Prior}, w^{AP}_{Prior}, w^R_{Prior}]$ are weights assigned to the prior cost terms. C^{VP}_{Prior} and C^{AP}_{Prior} denote the progression of valence and arousal throughout the photo sequence. The objective of these terms is to evaluate how the synthesized photo sequence fits a user-defined progression of valence and arousal. The term C^R_{Prior} constrains the photo sequence to not repeat an image if it already exists.

We apply a Gaussian model in our C^V_{Emo}, C^A_{Emo}, C^C_{Vis}, and C^B_{Vis} cost functions to penalize deviation and large variation from the desired targets. Essentially, Equations (2)–(5) evaluate how close the valence, arousal, color, and brightness of the current photo sequence is compared to the user-defined target valence, arousal, color, and brightness based on a Gaussian distribution. Note that in its current form, a user can easily control the targets through changing the corresponding variables in our software and also change the progress of valence and arousal through choosing from a list of pre-defined curves and lines. No additional control (e.g., draw a progression line or curve) is given to a user.

Note that aside from the proposed cost terms, other cost terms can be examined and implemented depending on the characteristics of the target domain that requires the automatic synthesis of photo sequences. In the next sections, we present the developed cost terms (Section 5) and the process for optimizing the photo sequence based on user-defined targets (Section 6).

5 COST TERMS

Details on all cost terms implemented in our software are presented in this section. Note that the valence and arousal ratings, as well as the dominant color and image brightness values, were normalized. Normalization ensured that all components in our optimization had a similar range and so would be easy to manage.

5.1 Emotional Terms

The emotional terms that are responsible for generating a new photo sequence S are defined in this section.

Valence Cost. Valence indicates averseness (negative valence) or intrinsic attractiveness (positive valence) [15]. To evaluate the synthesized photo sequence S as a whole and not as individual images that should follow specific target values, we defined a cost term to evaluate the average amount of valence that is included in the synthesized photo sequence:

$$C^V_{Emo}(S) = 1 - \exp\left(- \frac{\left(\frac{1}{|S|} \sum_{p_i} V(p_i) - \tau_V \right)^2}{2\sigma_V^2} \right), \tag{2}$$

where $V(p_i)$ denotes the valence rating of the p_i photo, and $\tau_V \in [0, 1]$ is the average target valence that is requested by a user in synthesizing a photo sequence S. Note that a low target valence value τ_V close to 0 denotes unpleasant, and a high valence value τ_V close to 1 denotes pleasant. σ_V controls the spread of the Gaussian penalty function and is empirically set as τ_V.

Arousal Cost. Arousal is the physiological and psychological state of being awoken or having organs stimulated to the point of perception [15]. Low arousal means calm and inactive, and

positive arousal means excited and active. Similar to the valence cost term, we evaluate the synthesized photo sequence S as a whole; therefore, we defined a cost term that encodes the average amount of arousal that is included in the synthesized photo sequence. The arousal cost function is expressed as

$$C^{A}_{Emo}(S) = 1 - \exp\left(- \frac{\left(\frac{1}{|S|} \sum_{p_i} A(p_i) - \tau_A \right)^2}{2\sigma_A^2} \right), \tag{3}$$

where $A(p_i)$ denotes the arousal rating of the p_i photo, and $\tau_A \in [0, 1]$ is the average target arousal requested by a user in synthesizing a photo sequence S. Finally, σ_A controls the spread of the Gaussian function and is set as τ_A.

5.2 Visual Consistency Terms

We developed cost terms that are responsible for synthesizing a photo sequence to fulfill user-defined visual consistency. Therefore, the synthesized photo sequence could be considered as visually-consistent. The developed costs are represented below.

Dominant Color Cost. Our dominant color cost synthesized the photo sequence by minimizing the color differences across images. Therefore, the synthesized results could be considered visually consistent. Moreover, as the user could control the target values of the dominant color, the system was also able to synthesize photos that fulfilled this user parameter. To avoid observing dominant color discrepancies across images, our dominant color cost term forces each image to follow the user define target dominant color. Thus, the dominant color cost is defined as

$$C^{C}_{Vis}(S) = 1 - \exp\left(- \frac{\frac{1}{|S|} \sum_{p_i} \left(C(p_i) - \tau_C \right)^2}{2\sigma_C^2} \right), \tag{4}$$

where $C(p_i)$ denotes the dominant color of the p_i image, and $\tau_C \in [0, 1]$ represents a user-defined target dominant color of the photo sequence S. Note that both $C(p_i)$ and τ_C represent the RGB color space so that the cost term is computed for each color channel individually. The normalized values of each individual channel of the RGB color space were used for this cost term. Thus, this cost term evaluates the dominant color $C(p_i)$ of picture p_i and the user-defined dominant color τ_C. Finally, we set σ_C as τ_C.

Brightness Cost. We have included brightness cost in our visual consistency term, which encoded the brightness of all pixels of the image. This term helped us synthesize a photo sequence in which the overall brightness did not deviate across images in the sequence. This is achieved by forcing each image that is evaluated by the brightness cost term to follow the user define target brightness value. The image brightness cost is defined as

$$C^{B}_{Vis}(S) = 1 - \exp\left(- \frac{\frac{1}{|S|} \sum_{p_i} \left(B(p_i) - \tau_B \right)^2}{2\sigma_B^2} \right), \tag{5}$$

where $B(p_i)$ denotes the average brightness of the p_i image, and $\tau_B \in [0, 1]$ denotes a user-defined target brightness of an image that belongs to the synthesized photo sequence S. This cost term evaluates the brightness $B(p_i)$ of picture p_i against the user target brightness τ_B. We also set σ_B as τ_B. Note that low τ_B values correspond to dark images and high τ_B values to bright images.

5.3 Prior Terms

Besides the ability to control the average amount of valence and arousal that should be included in the synthesized photo sequence, the prior terms were developed to provide the user with additional control. This control allowed the user to specify the progression of valence and arousal throughout

the photo sequence as well as to specify whether a photo should appear more than once in the synthesized photo sequence. At this point, we would like to briefly explain the terms of valence and arousal progression. In particular, a user specifies a target average valence amount of 0.5. The valence progression term asks the user to choose how that valence should be distributed in the synthesized photo sequence. The user could then input a curve or line that defines the valence amount for each image in the photo sequence. For example, the user might choose a Gaussian-like bell curve. Thus, the images at the beginning of the synthesized photo sequence would have a low valence, the images in the middle of the photo sequence would have a high valence, and the images at the end of the photo sequence would have a low valence. In this way, we ensured that the user-specified target valence level would be distributed according to the user-defined input. Note that the valence and arousal progression could have also been controlled directly from the valence and arousal cost; however, it would have been more difficult for the user to specify an average target level for valence and express it through an input curve or line.

Valence Progression Cost. The valence progression cost takes the valence value $V(p_i)$ of the ith picture of the photo sequence S, and evaluates this valence amount with a user's defined ith amount input $\tau_{V(p_i)}$, using the following:

$$C^{VP}_{Prior}(S) = \frac{1}{|S|} \sum_{p_i} \left(\mathcal{N}\left(V(p_i)\right) - \mathcal{N}(\tau_{V(p_i)}) \right)^2, \tag{6}$$

where $\mathcal{N}(\cdot)$ denotes a function that normalizes the valence amount $V(p_i)$ of all images that belong to the photo sequence S, while also normalizing all user-defined target valence values $\tau_{V(p_i)}$ before computing $C^{VP}_{Prior}(S)$. In this way, it is ensured that significant discrepancies between the input curve/line and the synthesized photo sequence are eliminated when computing the valence progression cost.

Arousal Progression Cost. In controlling the arousal progression throughout the photo sequence S, the arousal progression cost evaluates the arousal value $A(p_i)$ against the user-defined arousal amount $\tau_{A(p_i)}$ for the ith image of the photo sequence. The arousal progression cost is then represented as

$$C^{AP}_{Prior}(S) = \frac{1}{|S|} \sum_{p_i} \left(\mathcal{N}\left(A(p_i)\right) - \mathcal{N}(\tau_{A(p_i)}) \right)^2, \tag{7}$$

where, as before, $\mathcal{N}(\cdot)$ denotes a normalization function.

Repetition Cost. In synthesizing photo sequences and avoiding tedious repetition of images, a repetition term was added as a prior cost. Thus, the repetition cost ensured that the generated photo sequence would not include duplicate photos. This cost is represented as

$$C^{R}_{Prior}(S) = \frac{1}{\frac{|S|!}{\left(2!(|S|-2)!\right)}} \sum_{(p_i,p_j)} \Gamma(p_i, p_j), \tag{8}$$

where $\frac{|S|!}{(2!-(|S|-2)!)}$ returns the total number of combination between (p_i, p_j) that are two of the pictures of the photo sequence S. $\Gamma(p_i, p_j)$ is then computed based on the following condition:

$$\Gamma(p_i, p_j) = \begin{cases} 1 & \text{if the picture } p_i \text{ is identical to picture } p_j, \\ 0 & \text{otherwise.} \end{cases}$$

This ensures that the $C^{R}_{Prior}(S)$ cost term returns a high value when the picture p_i is identical to picture p_j. Conversely, the $C^{R}_{Prior}(S)$ cost term returns a low value when the picture p_i is different from picture p_j.

6 OPTIMIZATION

To automatically synthesize the photo sequence based on the user-defined targets, we employed an optimization method. Considering that a photo sequence could be generated with a collection of images that could be retrieved from an image dataset, an optimal solution for the user-defined target cost was searched in the solution space. For the optimization process, we used a Markov chain Monte Carlo technique, which is known as simulated annealing [34]. We then employed a Metropolis-Hastings state searching step [8]. Simulated annealing was chosen because: (1) its ability to avoid small local minima on route to a global minimum; (2) it works well with problems with a large number of parameters; (3) it generally outperforms other graph-based methods [28, 62]; and (4) the computational burden in working with the OASIS image dataset was acceptable (for $N = 10$ images, simulated annealing needed < 2 s to provide an optimal solution).

To employ the technique, we started by implementing a Boltzmann-like objective function:

$$f(S) = \exp\left(-\frac{1}{t} C_{Total}(S) \right), \tag{9}$$

In Equation (9), t denotes the temperature parameter of the simulated annealing process [34], which was set to decrease gradually throughout the optimization process. As the optimization evolved, the optimizer randomly chose and applied a move to the current photo sequence S to propose a new photo sequence S'. For the optimization process, two moves were considered by the optimizer: (1) replacing a randomly chosen image from photo sequence S, and (2) swapping images by randomly choosing two images from the existing photo sequence S.

At the beginning of the optimization process, a random photo sequence S that contained a user-defined N number of images was initialized. At each iteration of the optimization, one of the moves was chosen randomly and applied to a randomly selected image (or images when the swap images move is chosen) of the current photo sequence S to create a new photo sequence S'.

In our implementation, we defined the selection probabilities of moves as follows. For the first 500 iterations, the selection probabilities of the moves were set to $Pr_{replace} = 0.50$ for "replace an image" and $Pr_{swap} = 0.50$ for "swap images." For the second 500 iterations, the selection probabilities of the moves were set to $Pr_{replace} = 0.75$ for "replace an image" and $Pr_{swap} = 0.25$ for "swap images." For the rest of the iterations needed, the selection probabilities of the moves were set to $Pr_{replace} = 1.0$ for "replace an image" and $Pr_{swap} = 0.0$ for "swap images." As the optimization evolved, it favored the "replace image" move.

To decide if the system should accept the photo sequence S', our method compared the proposed total cost $C_{Total}(S')$ to the current total cost of $C_{Total}(S)$. The developed method accepts a proposed photo sequence S' based on the Metropolis criterion [8] denoted as

$$Pr(S'|S) = \min\left(1, \frac{f(S')}{f(S)}\right). \tag{10}$$

The simulated annealing method was employed to optimize different photo sequences. A temperature parameter t was first defined. When optimization began, the temperature parameter t was represented by a high value, allowing the optimizer to explore the optimized results aggressively. As the iterations of the optimization evolved, the temperature parameter was reduced until it reached zero. An initial temperature $t = 1.0$ was used at the beginning of the optimization and was reduced by 0.10 every 100 iterations. As the temperature parameter decreased, the optimizer became greedier in finding optimal solutions. The optimization process was complete when the total cost change was less than 5% in the past 100 iterations.

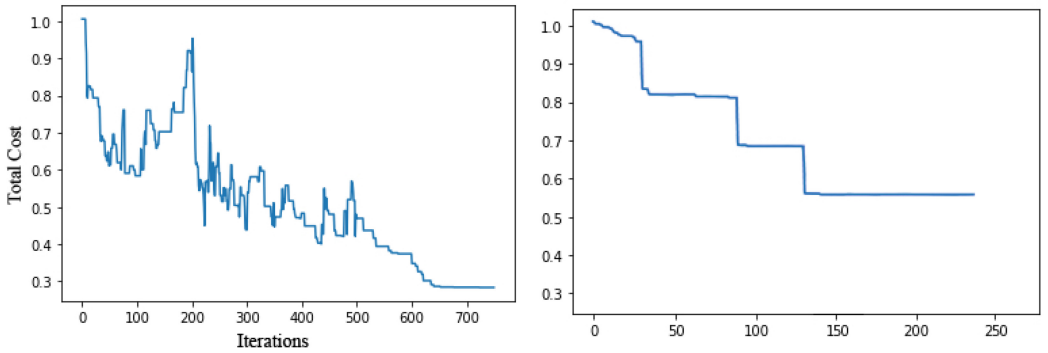

Fig. 3. A comparison between optimizing the photo sequence using the Markov chain Monte Carlo (left) and the greedy (right) algorithm. The Markov chain Monte Carlo algorithm achieves lower minima compared to the greedy algorithm.

A comparison of how the total cost $C_{Total}(S)$ changed over several iterations of the photo sequence optimization between the Markov chain Monte Carlo and the greedy algorithm is shown in Figure 3. The Markov chain Monte Carlo algorithm obtains a solution with a lower total cost value ($\approx$0.30) compared to the greedy approach ($\approx$0.55). The total cost value of the greedy algorithm experiment did not change from about iteration 140 to about iteration 240. Thus, the greedy optimization stopped at about iteration 240. Since the Markov chain Monte Carlo algorithm can accept a solution with a cost higher than that of the current solution with a certain acceptance probability, the sampling is capable of jumping out from a locally optimal solution. This prevents the sampling from being performed locally, and eventually locating a more optimal solution with a lower total cost value. Thus, the Markov chain Monte Carlo optimization stopped at about iteration 770.

Unless otherwise specified by the user, the weights assigned to the cost terms responsible for the emotional cost terms were set to $w^A_{Emo} = 1.00$ and $w^V_{Emo} = 1.00$, the weights of the visual consistency cost terms were set to $w^C_{Vis} = 0.50$ and $w^B_{Vis} = 0.50$, and the weights of the prior cost terms were set to $w^{VP}_{Prior} = 0.50$, $w^{AP}_{Prior} = 0.50$, and $w^R_{Prior} = 1.00$. Note that the user is able to control these weights and synthesize photo sequences by individually prioritizing the importance of each cost term. Figure 4 shows examples of synthesized photo sequences in which a higher priority was requested for a different color. Additionally, examples of synthesized sequences based on different user-defined valence and arousal target values are shown in Figure 5. Finally, examples of a synthesized photo sequence using different valence and arousal progression inputs are shown in Figure 6. We would like to note that in Figure 6 the photo sequences only follow the sequential logic of the valence-arousal targets. We decided to not prioritize the visual consistency cost terms, since we realized that there are not enough images in the dataset that could be used to synthesize a photo sequence that fulfills the sequential logic of the valence-arousal targets and visual consistency targets simultaneously.

7 USER STUDY

To evaluate the efficacy of our method in synthesizing photo sequences that fulfill specific valence and arousal targets, we conducted a user study to determine if valence and arousal targets achieved by the method could be appropriately perceived by general users.Because a developed photo sequence of a target valence and target arousal may include singular photos with varying valence and arousal values, it was important to test the perceived valence and arousal of the

Fig. 4. Photo sequences ($N = 10$) synthesized by prioritizing the visual consistency terms. From top to bottom, we have prioritized red, green, blue, yellow, and black color.

photo sequence in its entirety. For example, a photo sequence that meets a 0.50 arousal target may include photos with both 0.30 and 0.70 arousal values. Therefore, we evaluated the overall perceived target values from the photo sequence as a whole with our user study. We wanted to evaluate perceived valence and arousal from the photo sequence rather than from singular images, which are provided in the OASIS dataset. As this work is proof-of-concept, a user study with general users would be effective in determining if users could adequately perceive valence and arousal properties of our generated image sequences. We chose not to evaluate the visual consistency, as both the dominant color and brightness of the photo sequences could easily be inspected by the researchers after the photo sequences were generated.

7.1 Participants

Participants were recruited via emails sent to our department's undergraduate and graduate students. We recruited 53 participants in total. All participants were volunteers, and no compensation was involved. From the participant pool, 25 were female and 28 were male. Participants ranged in age from 18 to 33 ($M = 23.06$, $SD = 3.67$). Participants provided written consent that was approved by the Institutional Review Board of our university.

7.2 Stimulus and Procedure

We synthesized three photo sequences that included images with low, medium, and high arousal annotations at a constant valence of 0.50, as well as three photo sequences that included images with low, medium and high valence annotations at a constant arousal of 0.50. The six photo sequences were synthesized with target valence and arousal values of 0.20 for low, 0.50 for medium, and 0.80 for high. The target number of images per synthesized photo sequence was set at $N = 10$. For each of the developed valence-arousal variations, we created an image slideshow that consisted of ten images each. The images that belong to each synthesized slideshows are shown in Figure 5. The duration of the slideshow was 40 s, which meant that each image appeared for 4 s; the lower duration used by Chen et al. [7]. The ten images played during the 40-s period are considered the developed photo sequence.

After each 40-s photo sequence was displayed, users rated arousal, and then valence on separate pages of our survey, to avoid confusion between the two terms. Users rated the photo sequence in its entirety. After completing arousal and valence ratings for one photo sequence, another photo

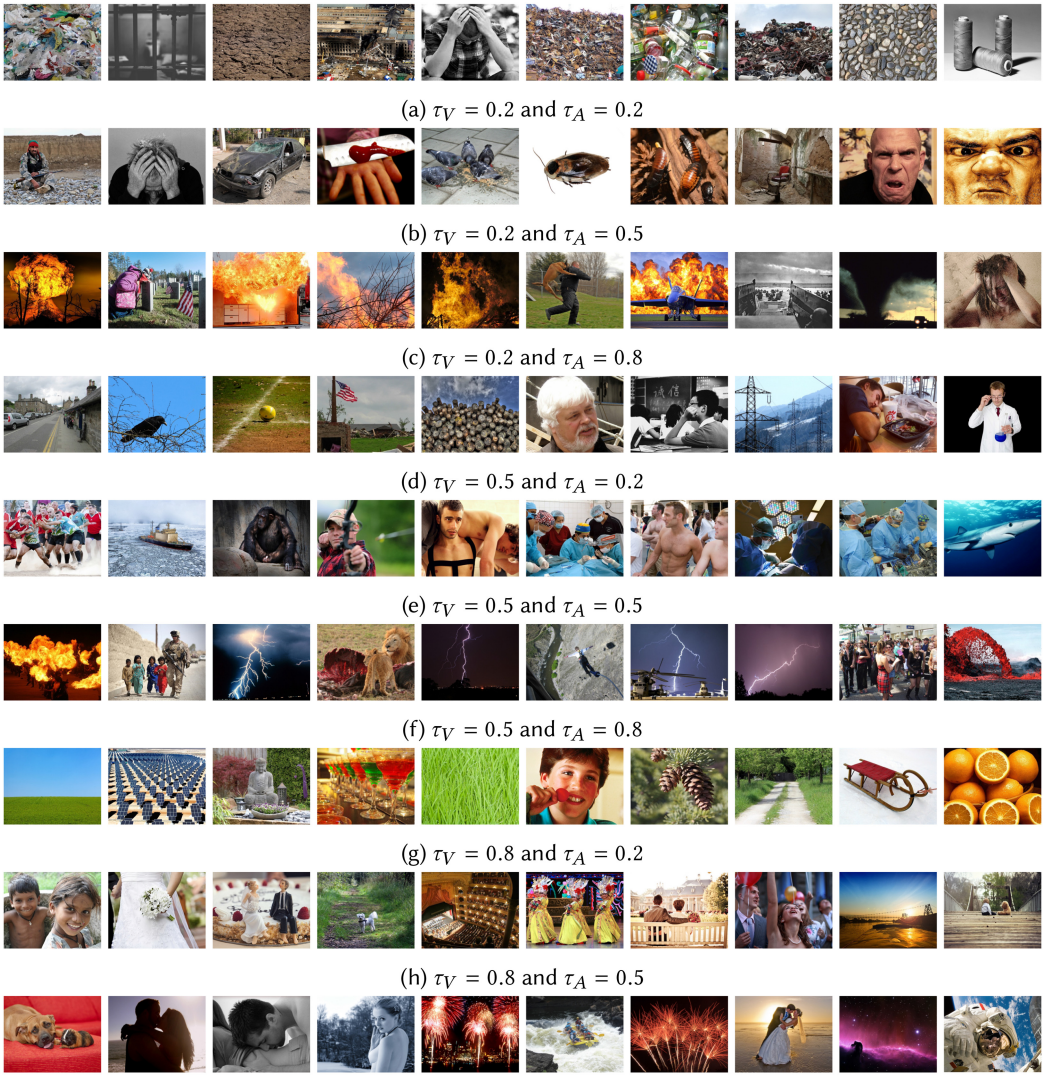

Fig. 5. Variation of synthesized photo sequences ($N = 10$) based on the developed method by varying the target values of valence and arousal.

sequence was displayed, and the participant then rated it. The total time for the viewing and rating of all photo sequences took less than 15 min per participant.

7.3 Results

We used a repeated measures Analysis of Variance with Huynh-Feldt correction to explore differences in arousal and valence ratings across low, medium and high arousal and valence photo sequences. Q-Q plots of the residuals were used to determine that the data was normally distributed. We used Bonferroni corrections for our post hoc comparisons.

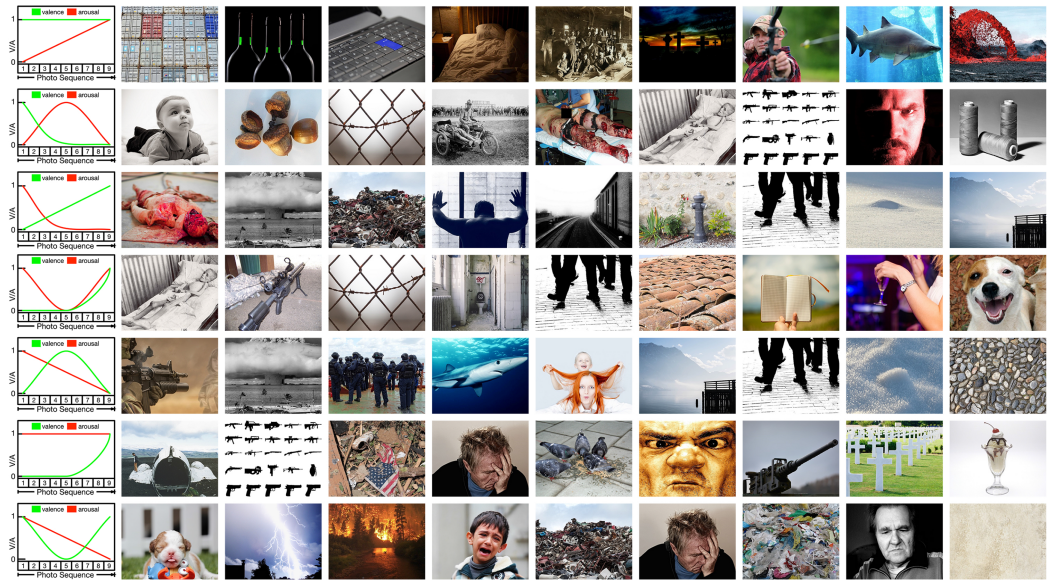

Fig. 6. Variation of synthesized photo sequences ($N = 9$) based on the developed method by defining different valence and arousal progress inputs (left column) and keeping both the average target valence ($\tau_V = 0.5$) and arousal ($\tau_A = 0.5$) constant. The weights assigned to the emotional cost terms were set to $w^A_{Emo} = 1.00$ and $w^V_{Emo} = 1.00$, and the weights of the visual consistency cost terms were set to $w^C_{Vis} = 0.50$ and $w^B_{Vis} = 0.50$; therefore, only the emotional cost terms were prioritized for the optimization process.

Our results indicate that arousal ratings across low, medium, and high arousal photo sequences were significantly different [$F(1.722, 89.532) = 18.683$, $p = 0.000$, $\eta^2_p = 0.264$]. Post hoc comparisons revealed that low arousal photo sequence ratings ($M = 3.53$, $SD = 0.19$) were significantly lower than both medium arousal photo sequence ratings ($M = 4.47$, $SD = 0.14$) and high arousal photo sequence ratings ($M = 5.02$, $SD = 0.18$). While post hoc comparisons revealed that medium arousal photo sequence ratings ($M = 4.47$, $SD = 0.14$) were significantly higher than ratings for the low arousal photo sequence ($M = 3.53$, $SD = 0.19$), the medium arousal photo sequence ratings ($M = 4.47$, $SD = 0.14$) did not differ significantly from the high arousal photo sequence ratings ($M = 5.02$, $SD = 0.18$). While our results demonstrate that users were able to perceive differences in low and high arousal and between low and medium arousal within our photo sequences, we were unable to determine significant differences between medium and high arousal photo sequence ratings.

We also examined differences in valence ratings of our photo sequences. We determined that valence ratings were significantly different across low, medium and high valence photo sequences [$F(1.784, 92.759) = 474.209$, $p = 0.000$, $\eta^2_p = 0.901$]. Post hoc comparisons showed that low valence photo sequence ratings ($M = 1.85$, $SD = 0.69$) were significantly lower than both medium valence photo sequence ratings ($M = 4.08$, $SD = 0.78$), and high valence photo sequence ratings ($M = 6.36$, $SD = 0.76$). Additionally, post hoc comparisons revealed that medium valence photo sequence ratings ($M = 4.08$, $SD = 0.78$) were significantly lower than high valence photo sequence ratings ($M = 6.36$, $SD = 0.76$). Our results indicate that users were able to perceive low, medium and high valence at a constant arousal within our synthesized photo sequences, as seen through ratings.

7.4 Discussion

According to Kurdi et al. [39], who developed the OASIS dataset, it is possible for arousal to be influenced by gender when viewing sexually explicit images. Although none of our photo sequences contained sexually explicit images, it is possible that the inclusion of a few suggestive images (see Figure 5) may have influenced our results concerning arousal ratings, in that women may have responded differently than men when rating the arousal for photo sequences with these images. Additionally, it has been shown that some individuals are prone to providing responses that are socially desirable, rather than personally felt [16]. Therefore, it is also possible that our results concerning arousal may have been impacted by the user's desire to respond in a socially acceptable manner.

Considering that arousal and valence can be influenced by mood or focus of attention, and that arousal focus may be more influential in defining emotional experiences for certain individuals more so than for other individuals [13], perhaps a lack of attention to the images' arousal level contributed to similar ratings for medium and high arousal photo sequences. Interesting next steps with this work might include having users rate their attention to both arousal and valence content, as well as having users report their state affect at the time of the study.

Kurdi et al. [39] found that differences in ratings arose from a user's liberal or conservative political stance concerning images with sexuality and violence. Considering that several images in our photo sequences depict scenes of violence, it is possible that we could not determine significant differences in arousal ratings between medium and high arousal photo sequences due to perception of violence, perhaps based on political feelings, as suggested by Kurdi et al. [39]. Differences in culture can also influence perception of affect; however, because we collected neither political nor cultural demographics from our users, we cannot make these claims, and can only suggest them as potential influences for arousal ratings.

In a review by Kuppens et al. [38], it was found that as people feel more positive or as they feel more negative, arousal ratings tend to be higher, suggesting a relationship between valence and arousal. Not only has it been shown that relationships between arousal and valence are likely, but it has been shown that individual differences may significantly impact reports of affect, so much so that arousal may not be an appropriate equivalent of affect intensity [38]. Despite our imperfect ratings of arousal, users generally could distinguish arousal levels as well as appropriately perceive variations in the valence of the photo sequences, as shown by significant differences in low, medium, and high valence photo sequence ratings.

8 LIMITATIONS

As with any photo sequence synthesis technique, the limited number of images contained in a database might affect the quality of the final photo sequence. Moreover, the difficulty in finding emotionally annotated images might also be considered a limitation, especially in cases where our method is used for personal photo collections and not for publicly available image datasets. We assume that with recent developments in deep learning, techniques that automatically predict valence and arousal levels [36, 50] could contribute to our method by allowing a user's photo collections to be annotated automatically. Therefore, our approach could become even more appropriate for personal photo collections with the potential inclusion of machine learning techniques.

Another limitation of our method, concerning the extraction process of visual information from the images, includes our lack of semantic analysis and inter-segment semantics such as time, season, and content. We believe that deep learning approaches [37, 77] could be used to extract information such as event identification [56], which could be helpful in assigning images to discrete groups. In this way, photos could be synthesized in a sequence by retrieving images from specific

groups. Given these limitations, we may want to include semantics, image features, and information grouping in our current total cost function to create more targeted photo sequences in the future.

9 CONCLUSION

We developed a method for synthesizing photo sequences in which images are rated based on valence and arousal, and in which visual information is extracted from the photos. Our method, which is considered a proof of concept, provides considerable flexibility to users, allowing them to synthesize photo sequences in fast and intuitive ways. In evaluating our method, we conducted a user study to determine if valence and arousal levels of a synthesized photo sequence are perceived as expected by the user. Based on our results, we conclude that participants were able to rate photo sequences appropriately, despite minor discrepancies in arousal ratings. Therefore, our method can be used for synthesizing photo sequences that trigger the desired valence and arousal levels of users.

Our current application dealt only with images. However, we hope to expand our system's functionality to handle sounds and short videos. Specifically, we plan to explore methods of enhancing and optimizing a photo sequence while considering features extracted from audio files, thus, enabling audiovisual slide shows. Moreover, we want to explore the synthesis of storyboards (e.g., weddings, graduations) using the valence-arousal emotional model. In future work, we would like to extract semantic meaning from photos, to explore photo sequences that tell stories in addition to evoking target emotions. Finally, we also plan to create an annotated dataset with short video clips, which once combined, will make it possible to synthesize a short story. Based on this dataset, we plan to explore the possibility of synthesizing short films in which the user will be able to control the emotional content and evolution of the story based on multiple objectives. By expanding our current method, we hope to develop an application that will prove useful to both public and private users who require fast and automated synthesized media sequences for a variety of purposes.

REFERENCES

[1] Edoardo Ardizzone, Marco La Cascia, and Filippo Vella. 2008. A novel approach to personal photo album representation and management. In *Multimedia Content Access: Algorithms and Systems II*, Vol. 6820. International Society for Optics and Photonics, 682007.

[2] Kristopher Blom and Steffi Beckhaus. 2005. Emotional storytelling. In *Proceedings of the IEEE Virtual Reality Conference*. 23–27.

[3] L. Carretié, M. Tapia, S. López-Martín, and J. Albert. 2019. EmoMadrid: An emotional pictures database for affect research. *Motiv. Emot.* 43, 6 (2019), 929–939.

[4] Tony F. Chan, Patrick Ciarlet Jr., and W. K. Szeto. 1997. On the optimality of the median cut spectral bisection graph partitioning method. *SIAM J. Sci. Comput.* 18, 3 (1997), 943–948.

[5] Seah Chang, Chai-Youn Kim, and Yang Seok Cho. 2017. Sequential effects in preference decision: Prior preference assimilates current preference. *PloS One* 12, 8 (2017).

[6] Jiajian Chen, Jun Xiao, and Yuli Gao. 2010. iSlideShow: a content-aware slideshow system. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*. 293–296.

[7] Jun-Cheng Chen, Wei-Ta Chu, Jin-Hau Kuo, Chung-Yi Weng, and Ja-Ling Wu. 2006. Tiling slideshow. In *Proceedings of the 14th ACM International Conference on Multimedia*. 25–34.

[8] Siddhartha Chib and Edward Greenberg. 1995. Understanding the metropolis-hastings algorithm. *Amer. Stat.* 49, 4 (1995), 327–335.

[9] Wei-Ta Chu and Chia-Hung Lin. 2009. Automatic summarization of travel photos using near-duplication detection and feature filtering. In *Proceedings of the 17th ACM International Conference on Multimedia*. 1129–1130.

[10] Tammara T. A. Combs and Benjamin B. Bederson. 1999. Does zooming improve image browsing? In *Proceedings of the 4th ACM Conference on Digital Libraries*. 130–137.

[11] Jeffrey Dalton, James Allan, and Pranav Mirajkar. 2013. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. 1857–1860.

[12] Andrew J. Elliot, Mark D. Fairchild, and Anna Franklin. 2015. *Handbook of Color Psychology*. Cambridge University Press.
[13] Lisa A. Feldman. 1995. Valence focus and arousal focus: Individual differences in the structure of affective experience. *J. Personal. Soc. Psychol.* 69, 1 (1995), 153.
[14] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from google's image search. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, Vol. 2. IEEE, 1816–1823.
[15] Nico H. Frijda et al. 1986. *The Emotions*. Cambridge University Press.
[16] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personal. Individ. Differ.* 7, 3 (1986), 385–400.
[17] Yuan Gan, Yan Zhang, Zhengxing Sun, and Hao Zhang. 2020. Qualitative photo collage by quartet analysis and active learning. *Comput. Graph.* 38 (2020), 35–44.
[18] Yuli Gao, Clayton Brian Atkins, Phil Cheatle, Jun Xiao, Xuemei Zhang, Hui Chao, Peng Wu, Daniel Tretter, David Slatter, Andrew Carter et al. 2009. MagicPhotobook: Designer inspired, user perfected photo albums. In *Proceedings of the 17th ACM international conference on Multimedia*. 979–980.
[19] Joe Geigel and Alexander C. P. Loui. 2000. Automatic page layout using genetic algorithms for electronic albuming. In *Internet Imaging II*, Vol. 4311. International Society for Optics and Photonics, 79–90.
[20] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. 2011. Computational color constancy: Survey and experiments. *IEEE Trans. Image Process.* 20, 9 (2011), 2475–2489.
[21] Gerardo Gonzalez Garcia and Rudy Lapeer. 2009. An evaluation of photo-consistency for intra-operative registration in an image enhanced surgical navigation (IESN) System. *Proceedings of Medical Image Understanding and Analysis (MIUA'09)*. 229–233.
[22] Paul Heckbert. 1982. Color image quantization for frame buffer display. *ACM SIGGRAPH Comput. Graph.* 16, 3 (1982), 297–307.
[23] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. 2007. Reranking methods for visual search. *IEEE MultiMedia* 14, 3 (2007), 14–22.
[24] Jun Huang, Xiaokang Yang, Xiangzhong Fang, Weiyao Lin, and Rui Zhang. 2011. Integrating visual saliency and consistency for re-ranking image search results. *IEEE Trans. Multimedia* 13, 4 (2011), 653–661.
[25] Ronald Hübner and Martin G. Fillinger. 2016. Comparison of objective measures for predicting perceptual balance and visual aesthetic preference. *Front. Psychol.* 7 (2016), 335.
[26] Alejandro Jaimes and Shih-Fu Chang. 1998. Model-based classification of visual information for content-based retrieval. In *Storage and Retrieval for Image and Video Databases VII*, Vol. 3656. International Society for Optics and Photonics, 402–414.
[27] Yushi Jing and Shumeet Baluja. 2008. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 11 (2008), 1877–1890.
[28] David S. Johnson, Cecilia R. Aragon, Lyle A. McGeoch, and Catherine Schevon. 1989. Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning. *Operat. Res.* 37, 6 (1989), 865–892.
[29] Kolbeinn Karlsson, Wei Jiang, and Dong-Qing Zhang. 2014. Mobile photo album management with multiscale timeline. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 1061–1064.
[30] Mel W. Khaw and David Freedberg. 2018. Continuous aesthetic judgment of image sequences. *Acta Psychol.* 188 (2018), 213–219.
[31] Jinho Kim, Suan Lee, Ji-Seop Won, and Yang-Sae Moon. 2011. Photo cube: an automatic management and search for photos using mobile smartphones. In *Proceedings of the IEEE 9th International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 1228–1234.
[32] Kwanghwi Kim, Sora Kim, and Hwan-Gue Cho. 2012. A compact photo browser for smartphone imaging system with content-sensitive overlapping layout. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*. 1–8.
[33] Serkan Kiranyaz, Stefan Uhlmann, and Moncef Gabbouj. 2009. Dominant color extraction based on dynamic clustering by multi-dimensional particle swarm optimization. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*. IEEE, 181–188.
[34] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220, 4598 (1983), 671–680.
[35] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision*. Springer, 662–679.
[36] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vision Comput.* 65 (2017), 23–36.
[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. MIT Press, 1097–1105.

[38] Peter Kuppens, Francis Tuerlinckx, James A. Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychol. Bull.* 139, 4 (2013), 917.
[39] Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. 2017. Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* 49, 2 (2017), 457–470.
[40] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch. 2018. Image selection in photo albums. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 397–404.
[41] Marco La Cascia, Marco Morana, and Salvatore Sorce. 2010. Mobile interface for content-based image management. In *Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE, 718–723.
[42] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proceedings of the 18th International Conference on World Wide Web*. 351–360.
[43] Guang-Hai Liu, Zuo-Yong Li, Lei Zhang, and Yong Xu. 2011. Image retrieval based on micro-structure descriptor. *Pattern Recogn.* 44, 9 (2011), 2123–2133.
[44] Paul J. Locher, Pieter Jan Stappers, and Kees Overbeeke. 1998. The role of balance as an organizing design principle underlying adults' compositional strategies for creating visual displays. *Acta Psychol.* 99, 2 (1998), 141–161.
[45] Hugo Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Med. Hypoth.* 78, 2 (2012), 341–348.
[46] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 457–466.
[47] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4535–4544.
[48] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 4 (1996), 261–292.
[49] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 4 (1996), 261–292.
[50] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* 10, 1 (2017), 18–31.
[51] Natali Moyal, Avishai Henik, and Gideon E. Anholt. 2018. Categorized affective pictures database (CAP-D). *J. Cogn.* 1, 1 (2018).
[52] Wolfgang Nejdl and Claudia Niederee. 2015. Photos to remember, photos to forget. *IEEE MultiMedia* 22, 1 (2015), 6–11.
[53] David Chek Ling Ngo, Azman Samsudin, and Rosni Abdullah. 2000. Aesthetic measures for assessing graphic screens. *J. Info. Sci. Eng.* 16, 1 (2000), 97–116.
[54] Pere Obrador, Rodrigo De Oliveira, and Nuria Oliver. 2010. Supporting personal photo storytelling for social albums. In *Proceedings of the 18th ACM international conference on Multimedia*. 561–570.
[55] Teresa K. Pegors, Marcelo G. Mattar, Peter B. Bryan, and Russell A. Epstein. 2015. Simultaneous perceptual and response biases on sequential face attractiveness judgments. *J. Exper. Psychol.: Gen.* 144, 3 (2015), 664.
[56] Sang Phan, Duy-Dinh Le, and Shin'ichi Satoh. 2015. Multimedia event detection using event-driven multiple instance learning. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1255–1258.
[57] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Amer. Sci.* 89, 4 (2001), 344–350.
[58] Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011. Multimedia retrieval in social networks for photo book creation. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. 1–2.
[59] Rose M. Rider. 2010. Color psychology and graphic design applications. Senior Honors Theses 111. Liberty University. Retrieved From https://digitalcommons.liberty.edu/honors/111.
[60] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. 2001. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 190–197.
[61] James A. Russell. 1980. A circumplex model of affect. *J. Personal. Soc. Psychol.* 39, 6 (1980), 1161.
[62] Rob A. Rutenbar. 1989. Simulated annealing algorithms: An overview. *IEEE Circ. Devices Mag.* 5, 1 (1989), 19–26.
[63] Fereshteh Sadeghi, J. Rafael Tena, Ali Farhadi, and Leonid Sigal. 2015. Learning to select and order vacation photographs. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 510–517.
[64] Mukesh Kumar Saini, Fatimah Al-Zamzami, and Abdulmotaleb El Saddik. 2014. Towards storytelling by extracting social information from OSN photo's metadata. In *Proceedings of the 1st International Workshop on Internet-Scale Multimedia Management*. 15–20.
[65] Carl Emil Seashore. 1908. *Elementary Experiments in Psychology*. Holt.
[66] Feng Shao, Mei Yu, and Gangyi Jiang. 2007. Dominant color extraction based color correction for multi-view images. *Chinese Optics Lett.* 5, 8 (2007), 449–451.

[67] Pinaki Sinha, Hamed Pirsiavash, and Ramesh Jain. 2009. Personal photo album summarization. In *Proceedings of the 17th ACM international conference on Multimedia*. 1131–1132.
[68] Terry Lee Stone, Sean Adams, and Noreen Morioka. 2008. *Color Design Workbook: A Real world Guide to Using Color in Graphic Design*. Rockport Pub.
[69] Pablo P. L. Tinio and Helmut Leder. 2009. Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychol.* 130, 3 (2009), 241–250.
[70] Cody Tousignant and Glen E. Bodner. 2014. Context effects on beauty ratings of photos: Building contrast effects that erode but cannot be knocked down. *Psychol. Aesthet. Creat. Arts* 8, 1 (2014), 81.
[71] Cody Tousignant and Glen E. Bodner. 2018. Context effects on beauty ratings of abstract paintings: Contrast, contrast, everywhere! *Psychol. Aesthet. Creat. Arts* 12, 3 (2018), 369.
[72] Tiberio Uricchio, Marco Bertini, Lorenzo Seidenari, and Alberto Bimbo. 2015. Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 9–15.
[73] Lujin Wang, Joachim Giesen, Kevin T. McDonnell, Peter Zolliker, and Klaus Mueller. 2008. Color design for illustrative visualization. *IEEE Trans. Visual. Comput. Graph.* 14, 6 (2008), 1739–1754.
[74] Mark D. Wood. 2008. Exploiting semantics for personalized story creation. In *Proceedings of the IEEE International Conference on Semantic Computing*. IEEE, 402–409.
[75] Mark D. Wood, Madirakshi Das, Peter O. Stubler, and Alexander C. Loui. 2016. Event-enabled intelligent asset selection and grouping for photobook creation. *Image Vision Comput.* 53 (2016), 57–67.
[76] Xiaolin Wu. 1991. Efficient statistical computations for optimal color quantization. In *Graphics Gems II*. Elsevier, 126–133.
[77] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. 2015. Recognize complex events from static images by fusing deep channels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1600–1609.
[78] Shao-Fu Xue. 2015. *Aesthetics of photographs, photobooks, and magazine covers: Tools for autonomous quality evaluation and content creation*. Ph.D. Dissertation. Purdue University.
[79] Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo, and Tsia-Hsing Li. 2008. A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval. *J. Vis. Commun. Image Represent.* 19, 2 (2008), 92–105.
[80] Seungji Yang, Sihyoung Lee, Yong Man Ro, and Sang-Kyun Kim. 2007. Semantic photo album based on MPEG-4 compatible application format. In *Proceedings of the International Conference on Consumer Electronics*. IEEE, 1–2.
[81] Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. 2016. Automatic generation of visual-textual presentation layout. *ACM Trans. Multimedia Comput. Commun. Appl.* 12, 2 (2016), 1–22.
[82] Jun Yu, Xiaokang Yang, Fei Gao, and Dacheng Tao. 2016. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans. Cybernet.* 47, 12 (2016), 4014–4024.
[83] Lei Zhang, Le Chen, Feng Jing, Kefeng Deng, and Wei-Ying Ma. 2006. EnjoyPhoto: A vertical image search engine for enjoying high-quality photos. In *Proceedings of the 14th ACM international conference on Multimedia*. 367–376.

Received April 2020; revised November 2020; accepted March 2021