PhotoFeel: Feeling Your Photo Collection with Graph-based Audiovisual Flocking

Cheng-Te Li, Hsun-Ping Hsieh, Shou-De Lin
Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan
{d98944005, d98944006, sdlin}@csie.ntu.edu.tw

ABSTRACT
This paper proposes an audiovisual presentation system, called *PhotoFeel*, to not only give users pleasant browsing atmosphere but also deliver a quick sense conveyed by given photo collection. While conventional photo display systems aim at improving the styles of presentation, we explore the visual and semantic feelings to create a space exhibiting the feelings from photographers. This is achieved by simulating the interactions among photos to emerge some flocking behaviors, where each photo is regarded as a simulated agent. To present the feelings of photos on the flocking, we construct two graphs by investigating the visual contents and tag semantics respectively. In addition, to enhance the diverse feelings of a photo collection, three audiovisual effects are composed to have rich presentation of feelings. Experimental results show that our *PhotoFeel* truly exhibits potential feelings for given photos and people comparatively favor our system.

Categories and Subject Descriptors
H.5.1 [**Information Interface and Presentation**]: Multimedia Information Systems – *Animation*; I.6 [**Simulation and Modeling**]: Types of Simulation – *Visual*.

General Terms
Design, Human Factors.

Keywords
Photo feelings, audiovisual presentation, flocking simulation, visualization, photo-based artwork.

1. INTRODUCTION
Nowadays, taking photos has become a popular activity for people to record what they experience in daily life. At the instant of pressing the shutter of camera, the photograph captures things ranging from physical impressions to personal feelings and moods, called "a picture is worth a thousand words." And thus we browse photo albums as if tasting our stories. This motivates us to devise an application which not only provides rich presentation for users to browse the visual contents but also allows users to experience the diverse feelings conveyed by the given collection of photos (e.g. album) from photographers.

--
Area Chair: Susanne Boll

MM'11, November 28--December 1, 2011, Scottsdale, Arizona, USA.

There are lots of tools and studies on photo browsing or exploring. A common used technique is the photo browser, like ACDSee[1] and Picasa[2] which provides a thumbnail manner to facilitate management of photo collections. Recently, some intelligent systems are devised to express rich browsing experience. Photo Tourism [12] presented an interactive navigation manner to explore photos in 3D scene. Photo Navigator [4] exploited the spatial relations among photos to create a feeling of space for users. In [13], a 2D virtual canvas was developed to easily browse desired photos based on visual similarities. On the other hand, the photo slideshow with music accompaniment is another prevalent form to experience photos. Photo2Video [5] applied "Ken Burns" effect to produce motion clips with incidental music to highlight some attentive regions. Tiling Slideshow [2] generated a music-driven slideshow where each frame is tiled by multiple photos. These techniques, however, mainly focus on improving the style of presentation for those who see the photos. The diverse feelings conveyed from the photos are not considered. Although the works [7][8] exploited music generation and accompaniment to present the feelings expressed in visual contents, they primarily riveted on illustrations and paintings, respectively.

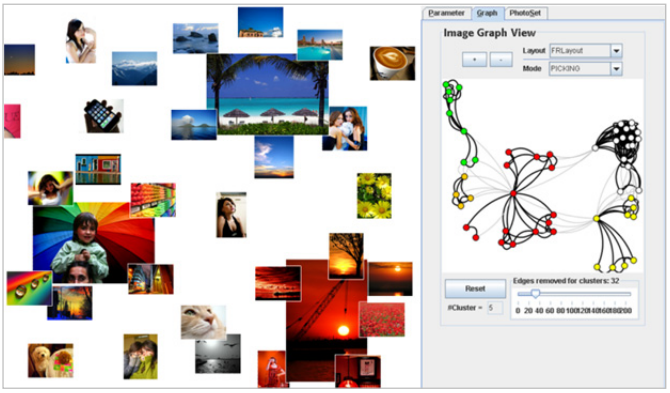

Figure 1. A snapshot (partial) for the proposed *PhotoFeel* audiovisual flocking simulation.

In this paper, we propose a novel audiovisual system, *PhotoFeel*, for users to not only browse the visual contents easily but also taste the feelings conveyed by the given collection of photos. The central idea of *PhotoFeel* is that by regarding each photo as a simulated character, termed photo agent, we create the physical interactions among photos to emerge some flocking behaviors, which exhibits a kind of movement process of "birds of a feather flock together" for characters. An atmosphere of feeling or mental picture will be created by viewing the emerging behaviors among photo agents. Since the feelings are physically characterized by

[1] ACDSee: http://www.acdsystems.com
[2] Picasa: http://picasa.google.com/

visual contents and tag semantics of photos, we model the feelings by constructing two *graphs* using visual and semantic similarities among photos. Such graphs are used to guide the flocking simulations and provide the feelings of photos from two diverse perspectives. We further compute three audiovisual effects to enhance the browsing and feeling experience as the simulation proceeds: (1) detecting photo clusters to direct the interactions among photo agents, (2) highlighting the representative photo for each flock of photo agents, and (3) calculating the affinity values between photos to generate real-time sound effects based on a series of psychological color-music mappings. Figure 1 shows a snapshot of *PhotoFeel* audiovisual flocking system.

2. SYSTEM OVERIEW

There are three components in the proposed *PhotoFeel* system, including content processing, graph-based computation, and audiovisual flocking simulation, as shown in Figure 2. First, given a collection of photos (e.g. album), where each photo is associated with some tags, we extract both some global and local visual features from them. And then we construct two graphs to capture the hidden feelings from two aspects: visual contents and tag semantics. Second, to present the feeling of photos during the simulation, based on the visual and semantic graphs, we develop three audiovisual effects, (a) photo clustering, (b) representative photo determination, and (c) feeling affinity calculation, which are integrated into the following flocking simulation. Finally, by considering each photo as a simulated agent, we combine item (a) and (b) to simulate and guide the interactions among photo agents to emerge flocking behaviors of photos for visualization, and we also use item (c) to generate real-time sound effects conforming to the captured feelings based on the two graphs. People can view and listen into the flocking behaviors to taste the feelings or atmosphere created by the given collection of photos.

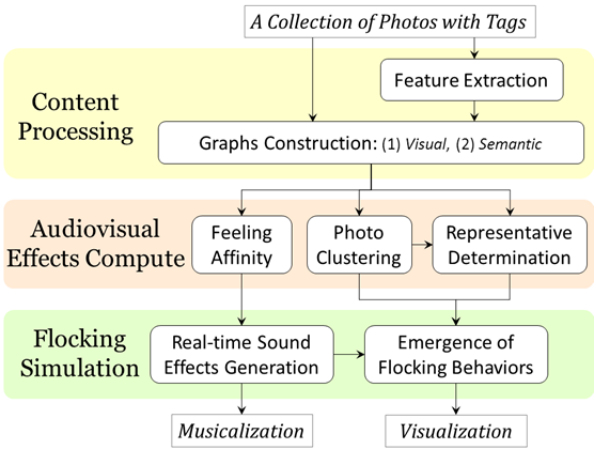

Figure 2. System flowchart of the proposed *PhotoFeel*.

3. CONTENT PROCESSING

In this section, we build the foundation of our system using the graph representation to capture the feelings among photos from the visual and semantic points of view. We first elaborate the visual features we employ for the constructions of graphs.

3.1 Visual Feature Extraction

We combine the global features of colors and texture and local geometric features to model the visual contents. The global features capture the photographic visual composition while the local ones identify the actual structural elements of objects. Such mixed global and local features have been shown to provide significant complementary information in recognition tasks [1].

- **Global Features.** We extract color histograms and color moments as global features of photos to capture theirs visual and spatial color distributions. We also utilize the Gabor textures to model the textural information. A feature vector will be produced for each photo by concatenating these features. The similarity between two photos is calculated according to normalized Euclidean distance.

- **Local Features.** We also represent the photos by local interest point descriptors, given by the scale-invariant feature transform (SIFT) with a Difference of Gaussian (DoG) process. Given two photos p_i and p_j with the descriptor vectors, $D_i=(d^1_i, d^2_i, ..., d^m_i)$ and $D_j=(d^1_j, d^2_j, ..., d^n_j)$, we define the similarity between two photos as the number interest points shared between two photos divided by their average number of interest points.

3.2 Graph Constructions

Based on the extracted visual features and the associated semantic tags on photos, we intend to model the feelings conveyed from photos by means of constructing a visual graph and a semantic graph. Both visual and semantic graphs contain photos as nodes, and the links between two photos are connected if their defined similarity exceeds a given threshold. The first one is the *visual graph*, which is designed to capture the visual feelings and impressions conveyed by the given photo collection. If two photos exhibit similar visual contents and resemble each other in appearance, they tend to express similar visual feelings, and thus we gather and connect them together in the graph. The second one is the *semantic graph*, which considers the photo feelings from the viewpoint of semantic tags. If two photos possess more common conceptual tags (e.g. object names, emotional terms, events, persons, and locations), they have high potential to convey similar meanings or reflect the same feelings, and thus we link them together in the sematic graph. In next section, we will elaborate how to exploit the visual and semantic graphs to compute diverse audiovisual effects for audiovisual flocking simulation. What follows describes how to construct these graphs in our system.

- **Visual Graph.** Given a collection of photos, the visual graph is constructed by estimating the similarity between photos using the extracted visual features. Since there are global and local features, we define the *visual similarity* between two photos p_i and p_j by using a convex function as

$$VisualSim(p_i, p_j)=\alpha \times GlobalSim(p_i, p_j) + (1-\alpha) \times LocalSim(p_i, p_j). \quad (1)$$

  This similarity definition provides a flexible preference when the given photo collection is characterized by either global or local part. Two photos are linked if their similarity exceeds a given *visual threshold* δ_v which controls the strength of visual feelings.

- **Semantic Graph.** We regard the tags associated on each photo as a kind of semantic features to construct the semantic graph. We modify the *Normalized Google Distance* (*NGD*) [3], which is a measure of semantic interrelatedness derived by calculating the correlation from Google search results, to define the *semantic similarity* between photos. Given two photos p_i and p_j with tag sets $T(p_i)$ and $T(p_j)$, we define the semantic similarity

$$SemanticSim(p_i, p_j) = \sum_{t_a \in T(p_i), t_b \in T(p_j)} \frac{1}{1+NGD(t_a, t_b)} \quad (2)$$

where $NGD(t_a, t_b) = \frac{\max\{\log f(t_a), \log f(t_b)\} - \log f(t_a, t_b)}{\log N - \min(\log f(t_a), \log f(t_b))}$, in which $f(t_a)$, $f(t_b)$, and $f(t_a, t_b)$ denotes the number of photos containing t_a, t_b, both t_a and t_b, separately in a photo corpus. N is the total number of photos in our corpus. Likewise, a *semantic threshold* δ_s is given to allow the extent of semantic feelings.

4. AUDIOVISUAL EFFECTS

To present the feelings conveyed by the given photos collection, in this section, based on the visual and semantic graphs, we attempt to create three audiovisual effects, including photo clustering, representative photo determination, and feeling affinity calculation. The first two parts are used to guide the interactions among photo agents to emerge some flocking behaviors from the visual aspect while the last one is used to generate real-time sound effects to create a kind of atmosphere consistent with the feelings behind given photos.

4.1 Photo Clustering

The first part is to group the photos according to visual and semantic graphs. Since a photo collection could deliver some different impressions, we gather photos with similar visual feelings to a cluster for further simulation so that the mental image of each flock can appear clearly during simulation. We perform the photo cluster by employing the *Fast Newman algorithm* [9], which a graph-based clustering method finding some tightly intra-connected and loosely inter-connected subgraphs as final clusters.

4.2 Representative Photos Determination

To enhance the feelings of each detected photo cluster, we find a representative photo for each cluster to be the central impression. The representative photos will be highlighted and enlarged during flocking simulation so that users can not only browse the diverse visual topics but also easily understand the feeling of each cluster. Based on the induced subgraph of each photo cluster, we provide three diverse importance measures to determine the representative photos from different aspects, including *degree*, *closeness* [15], and *PageRank* centrality. Degree is a local measure and thus served as a first glance or feeling for the feeling of a cluster. Both closeness and PageRank are global measures to find essential ones relatively in such subgraph of photos.

4.3 Feeling Affinity Estimation

While the above two parts are used to guide the flocking effects visually, we exploit this feeling affinity to generate sound effects in auditory perception. Specifically, if two photo agents are too close or collide during simulation, our system will produce some harmonious or discord sound effects according to their affinity of feelings. Two photos in the visual or semantic graph could deliver different feelings and the corresponding photo agents could meet during flocking simulation. We propose the feeling affinity to estimate their resemblance of mental impressions for sound effects generation. The basic idea of feeling affinity is that if two photos have higher structural proximity to one another in the graph, they tend to obtain higher affinity score. We employ the *Random Walk with Restart* (*RWR*) [14] to compute the feeling affinity. Given two photos p_i and p_j in the graph, the feeling affinity is defined as

$$FeelingAffinity(p_i, p_j) = RWR(p_i, p_j) + RWR(p_j, p_i). \quad (3)$$

where $RWR(p_i, p_j)$ is the steady-state probability that a random surfer walks from p_i and finally stay at p_j in the graph.

5. AUDIOVISUAL FLOCKING

Equipped with the three audiovisual effects capturing different mental impressions carried by the given photo collection, now we intend to present the audiovisual presentation system for feeling photos. We realize the idea of "birds of a feature flock together" to create a space of feelings for these photos. By taking each photo as an agent, we exploit the technique of flocking simulation to exhibit the sentimental interactions among photos for sensing. We employ a typical flocking model, *Reynolds' model* [11], in our system. This flocking model consists of three basic rules to lead agents to produce flocking behaviors. The first is *separation* rule steering agents to avoid crowding local flockmates, the second is *alignment* steering agents towards the average heading of local flockmates, and the third is *cohesion* steering agents to move forward the average position of local flockmates. On the other hand, by using the audiovisual effect of photo clusters, we devise a probability P_c to guide the interactions among photos agents to produce sentimental flocks during simulation process. That is, if P_c is set to higher, photos with similar feelings will have higher potential flock together. Moreover, we enhance the central feeling of each photo cluster by highlighting the corresponding representative photo to display. Some sound effects will be generated based on the audiovisual effect of feeling affinities among photo agents. What follows elaborates the emergent flocking behaviors exhibited in visual aspect, and describes how to generate sound effects to enhance the feeling in auditory aspect.

5.1 Emergence of Flocking Behaviors

We utilize two produced emergent flocking behaviors (i.e., visual and semantic) to illustrate and demonstrate the photo feelings delivered by our system. Figure 3 shows an example of flocking behaviors using the visual graph. That is, agents of photos with similar visual impressions tend to flock and move together. We can see there are four diverse appearances, including red, green, blue, and black-white. Besides, photo agents with no visual attributions wander among these four crews. Note that though the wandering agents could move toward some visual flock, they tend to leave the flock since they are affected by such cluster. Figure 4 shows the emergent flocking behaviors by the semantic graph. Likewise, photo agents with similar semantic feelings or concepts, including sunset, flower, cup, girl, and colorful, flock together controlled by P_c. Some agents belonging to no major clusters wander in the simulating space.

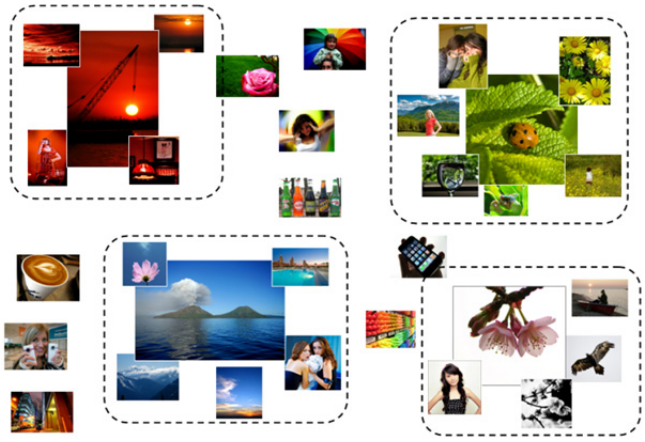

Figure 3. Flocking behavior using the *visual graph*.

Figure 4. Flocking behavior using the *semantic graph*.

6. EVALUATIONS

We evaluate our *PhotoFeel* system by subjective experiments, as the evaluation methods in [2][4][5][7][8]. We compare the flocking simulation produced by our *PhotoFeel* to the slideshows generated by ACDSee[1] and Photo Story[3] in terms of satisfaction. And the slideshows are accompanied with sentimental music. The dataset we use is the MIR Flickr 25000 collection [6]. We manually select two photo sets with number of photos = 100 for the evaluation.

Table 1. List of five criteria for evaluation.

Feeling (Sen)	Does the presentation convey feelings for photos?
Fun (*Fun*)	Is it a funny presentation?
Experience (*Exp*)	Does the presentation reach appreciation aesthetics?
Atmosphere (*Atm*)	How do you feel the audiovisual effects?
Acceptance (*Acc*)	Do you want to use it to taste your photos?

We invited 25 persons to participate this subjective experiment. Each person is asked to view the slideshows and our audiovisual flocking simulation. Then they are asked to give scores from 1 to 10 to express their satisfaction for the questions listed in Table 1. Higher score indicates better satisfaction for that criterion. Note that since our *PhotoFeel* provides two feelings (i.e., visual and semantic) to display, we compute average score of the two aspects for each criterion.

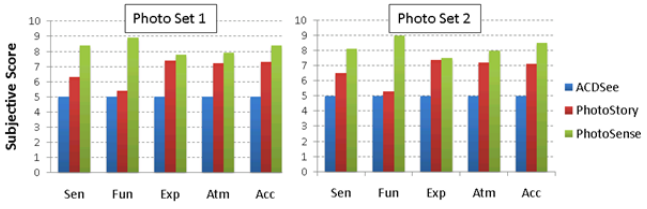

Figure 5. Results of subjective evaluation for two photo sets.

We set the score of ACDSee to 5 as the baseline. Figure 5 shows the results of subjective evaluation. We can observe our system outperform other two on average, especially for the criterion of feelings and fun. This comes from the exhibition of the space of central feelings by the emergent flocking behaviors among photo agents with audiovisual and animated effects. However, we nearly have no difference on the criterion of experience. We think it is because sometimes the movement of photo agents becomes a little complicated and not intuitive for users to explore.

[3] Photo Story: http://www.microsoft.com/athome/morefun/photostory.mspx

7. CONCLUSIONS

We present an audiovisual presentation system, *PhotoFeel*, to capture and exhibit the feelings conveyed by the given photos. The central idea is presenting flocking behaviors emerged from the interactions among photos to create a space of photo feelings. The feelings are modeled by visual and semantic graphs from visual contents and tag semantics. And then we embed the impressions of photos on our system by computing three audiovisual effects and applying them to guide the simulation in both visual and auditory aspects. Experimental results demonstrate the satisfaction of this new kind of photo presentation.

8. REFERENCES

[1] S. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. *NIST TRECVID Workshop*, 2005.

[2] J. C. Chen, W. T. Chu, J. H. Kuo, C. Y. Weng, and J. L. Wu, Tiling Slideshow. In *Proc. of ACM Intl. Conference on Multimedia (MM'06)*, 25–34, 2006.

[3] R. L. Cilibrasi, and P. B. M. Vitanyi. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(3), 370–383, 2007.

[4] C. C. Hsieh, W. H. Cheng, C. H. Chang, Y. Y. Chuang, and J. L. Wu. Photo Navigator. In *Proc. of ACM International Conference on Multimedia (MM'08)*, 419–428, 2008.

[5] X. S. Hua , L. Lu, and H. J. Zhang. Photo2Video - A System for Automatically Converting Photographic Series into Video. In *Proc. of ACM Intl. Conference on Multimedia (MM'04)*, 708–715, 2004.

[6] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *Proc. of ACM Intl. Conf. on Multimedia Information Retrieval (MIR'08)*, 39–43, 2008.

[7] K. Ishizuka and T. Onisawa. Generation of Variations on Theme Music Based on Impressions of Story Scenes. In *Proc. of ACM International Conference on Game Research and Development*, 129–136, 2006.

[8] C. T. Li and M. K. Shan. Emotion-based Impressionism Slideshow with automatic music accompaniment. In *Proc. of ACM International Conference on Multimedia (MM'07)*, 839–842, 2007.

[9] M. E. J. Newman. Fast Algorithm for Detecting Community Structure in Networks. *Physical Review*, E 69, 066133, 2004.

[10] R. W. Pridmore. Music and Color: Relations in the Psychophysical Perspective. *Colour Research and Application*, 17(1), 57–61, 1992

[11] C. W. Reynolds. Flocks, Herds and Schools: A Distributed Behavior Model. *ACM SIGGRAPH*, 25–34, 1987.

[12] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. *ACM Transactions on Graphics*, 25(3), 835–846, 2006.

[13] G. Strong and M. Gong. Organizing and Browsing Photos using Different Feature Vectors and Their Evaluations. In *Proc. of ACM Intl. Conference on Image and Video Retrieval (CIVR'08)*, No.3, 2008.

[14] H. Tong, C. Faloutsos, and J. Y. Pan. Fast Random Walk with Restart and Its Application. In *Proc. of IEEE Intl. Conference on Data Mining (ICDM'06)*, 613–622, 2006.

[15] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. *Cambridge University Press*, 1994.