CONTENT-AWARE AUTO-SOUNDTRACKS FOR PERSONAL PHOTO MUSIC SLIDESHOWS

Peter Dunker, Phillip Popp, and Randall Cook

Media Technology Lab (MTL)
Gracenote Inc, Emeryville, CA, USA
peter.dunker@ieee.org, {ppopp,recook}@gracenote.com

ABSTRACT

We present a novel slideshow generation concept based on content-aware photo-music mapping. Current technologies for automated personal photo slideshow generation primarily focus on photo presentations and visual effects. These solutions utilize either a manually chosen single song or preset slideshow songs. Our technology focuses on an automatic soundtrack generation process that attempts to comprehend what the photos depict and choose music accordingly. The process comprises analyzing a photo album, composing an auto-soundtrack by dynamically mapping song segments to photo events, and automatically choosing content-aware transitions and visual effects appropriate for the presentation.

Index Terms— Personal photo slideshows, music information retrieval, photo information retrieval, emotion recognition

1. INTRODUCTION

Along with the ubiquity of affordable digital cameras, the creation and sharing of digital slideshows has become a pervasive activity. While the term *slideshow* might harken back to the sentimental old days of the carousel slide projector, today we find new tools for creating exciting and entertaining slideshows that allow us to easily share them with friends and families over the web. These tools add exciting and eye-catching photo transitions, accompany the slideshow with music, and package this multimedia experience in a form that can be shared with others. Though these tools enliven the slideshow, they are largely agnostic towards the actual content of the photos and the interplay between photos, transitions and musical events. If we consider different types of photo albums such as photos of parties, day trips, or multi-week vacation trips, we can improve the overall quality of the slideshow by pairing it with music that relates to the album and conforms to personal preferences.

For personal photo albums shared with friends and family, it is essential that the photos appear in the order they were taken. The importance of the order of photos is demonstrated in several successful commercial tools which employ timeline user interfaces for organizing, presenting, and grouping photo collections [1]. By retaining the chronological order of the photos, we preserve the individual events within the album, such as surfing at the beach in the morning, having lunch in the city midday, and attending a party in the evening. This motivates us to select different songs to accentuate the content within specific photo events.

2. RELATED WORK

Several publications describe methods for content-aware audio/visual slideshow generation. Various papers focus on photo analysis in order to present multiple related photos on a single slide [2]. In [3], music and photo analysis techniques were applied to gather high-level information to map photos to music. Other methods analyze the mood of music at high granularity and select photos with a similar mood or emotional expression for the presentation [4].

All of these approaches utilize a single song for the complete presentation. They either map the entire photo album to a single song, or choose subsets of photos to match the music. Approaches based on assigning selected photos to certain music positions cannot be applied to our targeted scenario because the chronological order of the photos would not be preserved. A second challenge arises when the mood in a series of photos changes, but the mood of music does not. This requires a different, more sophisticated, approach to music and photo album analysis and mapping, as well as to the overall slideshow authoring process. To the knowledge of the authors, the challenge of conserving the chronological order of photos while also performing a content-aware mapping of photos to music has not yet been addressed.

In the following chapters we describe how we overcome these hurdles by utilizing several information retrieval algorithms, creating a novel concept of mapping, and using those tools to author a personalized, content-aware slideshow presentation.

3. PRESENTED TECHNOLOGY

The presented system is realized in a web-based environment where users can upload their photo albums, specify sev-

978-1-61284-350-6/11/$26.00 ©2011 IEEE

eral music preferences, and watch automatically generated slideshow videos. The overall generation process is depicted in Fig. 1 and is based on three major steps. First, the music collection is indexed and dynamic music metadata are generated for each song in the collection. The music metadata are stored in a database to enable easy selection of music segments according to different criteria. Second, the user's photo album is analyzed by extracting facial emotions, location, and visual sensual features for each image. Third, the photo-music slideshow is generated by first estimating photo-event boundaries by analyzing the photo collection's temporal distribution, and then mapping photo criteria into music criteria in order to select song segments to accompany each photo event. Then, appropriate photo transitions are chosen and aligned to musical beats. Finally, the authored slideshow is rendered as an MPEG-4 video.

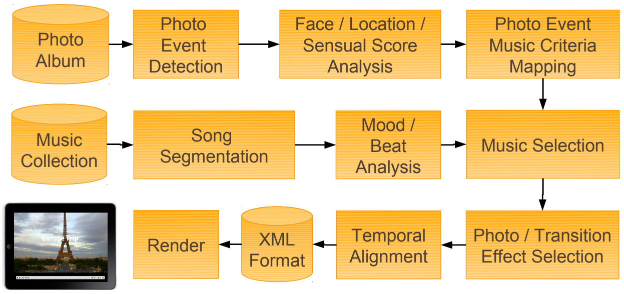

Fig. 1. System process overview

3.1. Information Retrieval

Several metadata attributes are derived from the user's photos and music. These are determined primarily by automatic analysis of the photo or song, though some music metadata rely on external web-services to retrieve e.g. genre information. While certain metadata such as the timestamp of the photo and geographical location can be embedded within the photo's technical metadata (as in EXIF), information extracted via automatic analysis of music and photos contribute significantly to the subsequent photo-music mapping process. Combined, these pieces of information allow characteristics of the images to be mapped to characteristics of the songs.

3.1.1. Music Analysis

The music analysis process extracts three distinct pieces of metadata: segments, moods, and beats. Additionally, editorial information about a particular song, such as the genre and artist's origin, is looked up using a commercial music information web service [5].

Segmentation groups a single song into useful temporal elements. First, the beginning and ending locations of sections within a song are pinpointed using auditory segmentation techniques [6]. This results in timestamps of basic elements of a song such as the introduction, verse, and/or chorus.

Second, within each section we analyze the audio to determine its mood. The mood of music describes universal characteristics that relate to a human's emotional response to hearing it. Several models exist for describing the mood of music such as hierarchical labeling, multi-labeling, and dimensional representations such as valence/arousal [7]. We chose the valence/arousal model (see Fig. 2) because it supports a simple mapping between music valence/arousal and the information we extract from photo events. Third, we perform beat analysis to automatically produce timestamps of beat locations within a song. The beat timestamps are organized on a grid and correspond to the tempo and meter of the originating song. Auditory beat analysis involves first estimating the tempo of a song and then applying a dynamic programming approach to determine the exact locations of beats based on the song's spectral flux and tempo [8].

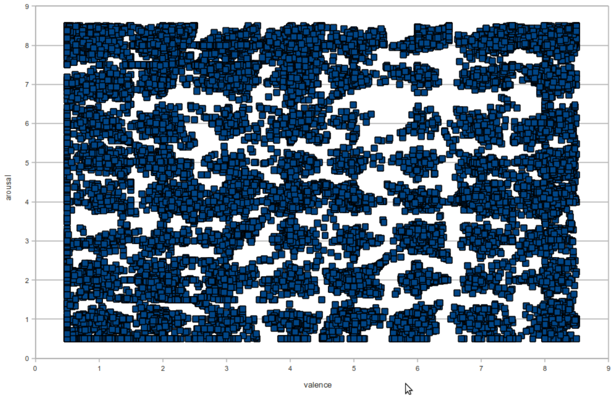

Fig. 2. Valence and arousal distribution of 25000 segments.

3.1.2. Photo Album Analysis

Photo album analysis is comprised of three modules: facial analysis, location analysis, and image sensation score estimation.

Some of the most important objects in consumer photos are the faces of friends and relatives, and are perhaps the highlight of the photo. Furthermore, faces express emotions which typically dominate the emotional impression of other objects in a photo. For example, a photo of a rainy day with dark clouds, but bright smiling faces in the foreground, will lead to a positive happy impression rather than the dark brooding impression implied by the rainy clouds. We utilized an approach similar to [9] for face analysis, which performs face detection as well as face categorization. We exploit the face positions and sizes as well as the confidence scores for the categories smile (cf_{smile}) and child (cf_{child}) in later stages.

In addition to face information, the location where the photos were taken is important to understand the complete

Fig. 3. Five consecutive photo events depict the difference in visual appearance and emotional impression.

photo event and album. In our system we utilize GPS values and keywords stored in the photo metadata. Geographical metadata is looked up via the reverse geo-location web service GeoNames [10]. Later, this information is used to choose musical artists who originate from the locations where the photos were taken.

Besides the high level semantic information about faces and locations, we extract visual features directly from pixel data which reflect the overall visual appearance. We estimate the overall darkness and color intensity based on highly saturated and bright pixels. Two features, a weighted darkness and a weighted color intensity score, are derived from a Gaussian weighted $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ HSV color space. The Gaussian weighting gives a smoother transition between pixels considered as dark and non-dark or high vs. low color intensity. For the darkness score we accumulate the Gaussian weighted value V with $\mu = 0$ and $\sigma = 30/256$. The Gaussian weighted pixel values V_w and S_w are estimated with $\mu = 1$ and $\sigma = 64/256$ from V and S. The color intensity scores are estimated by accumulating $\sqrt{S_w \times V_w}$. The μ and σ values were empirically determined. In Fig. 3, photos of five events are depicted, while Fig. 4 shows the image sensation scores of the same photo events.

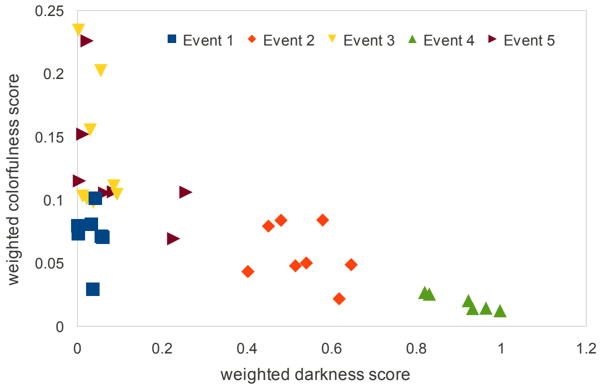

Fig. 4. The scatter plot depicts the distribution of two image score values across different photo events. Each point represents a single photo in an event.

The scatter plot in Fig. 4 substantiates our assumption that photo albums consist of photo events with differing sensual perspectives. This further supports our method for generating slideshows by selecting different music segments for individual photo events.

3.2. Slideshow Generation

Slideshow generation, the main contribution of the presented technology, combines the information gathered in the previous steps to create a context aware slideshow. In Fig. 5 we depict the structure of a slideshow. While preserving the chronological ordering of the images within an album, we group them into distinct photo events. These events are analyzed for semantic commonalities. Based on these criteria, matching music segments are chosen to accompany the individual photo events.

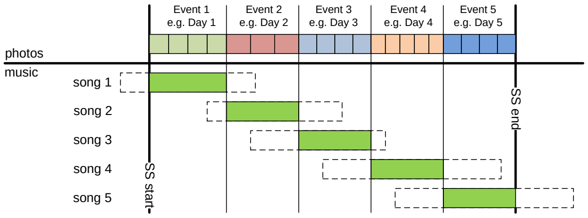

Fig. 5. The typical structure of the generated slideshows is depicted. Suitable song segments are chosen and presented together with previous estimated photo events.

3.2.1. Photo Event Detection

Photo event detection focuses on temporal analysis of the EXIF photo timestamps, as well as visual impressions derived from the presented image scores. A k-means algorithm similar to [11] was deployed, which clusters photo events based on the timestamps. We estimate the time differences $d_{0..n-1}$ in seconds for n consecutive photos, and cluster these differences. With $k = 2$, the cluster centroid with the bigger difference represents photos at the borders of photo events. The photo album is then split into events between this particular photo pair.

To enable more flexibility and allow user preferences regarding the number of photos within a photo event, we adapted this algorithm to use $k = 10$ and applied additional logic. After clustering, we iterate from the centroid with the highest difference value to the smallest while computing the average number of photos in an event until the user's preference is achieved. This results in a list of photo events consisting of images taken in connected periods of time. If the average photo event size at the end of this process is still larger than the user's preference, we split the photo events further by estimating the biggest euclidean distance of image scores across multiple neighboring photos using a Haar-like sliding window.

3.2.2. *Photo Event / Music Mapping*

To map music to photo events, we first determine a description of the photo event based on various metadata perviously extracted from the individual photos. We then map this description into music metadata search criteria in order to find a segment of music that best fits the photo event. To algorithmically describe the photo event, we create a *child score* from the amount and confidence values of children's faces cf_{child} in the photo event, a *smile score* from the amount and confidence values of smiling faces cf_{smile}, the *location* where the majority of photos were taken, and the *average image scores* across all photos. The color intensity and inverted blackness scores are mapped to valence and arousal of musical moods. From location information of the photo event we create a filter for editorial metadata, especially artist origin. The relevance of the child and smile scores are gated by a threshold, which defines whether children or smiling faces are important to the photo event. If the smile score passes the threshold it is mapped onto valence and arousal values, which means smiling faces lead to positive energetic mood values. If the child score passes the threshold it is used as valence and inverted as arousal value, which relates to positive calm mood values. All mapped valence and arousal values are then merged by averaging.

3.2.3. *Music Selection*

The music lookup focuses mainly on the search for song segments matching the previous estimated criteria. Beside these automatically determined criteria, additional user preferences, e.g. for genre or artists, are used to filter the search. Combinations of these filters are leveraged to provide music themes to the user in order to ease the music selection. Additional filters are introduced, e.g. to prevent choosing the same song for multiple photo events, or selecting a song start segment for the first photo event. Chorus song segments are preferred as they are more prominent to most users. Aside from the valence and arousal values all criteria are considered equally as search filters, whereas the valence and arousal values are used for a nearest neighbor search in the emotional space. The first song segment of the sorted result list is used for the corresponding photo event.

3.2.4. *Photo Effects, Transitions and Alignment*

Image transitions and effects improve the overall quality of the slideshow, yet applying them haphazardly can impact the result negatively. Typical photo presentation transition effects are "Ken Burns", zooming to and from faces, tiling of photos, cross fades between photos, or displaying multiple photos at the same time in single motif. To ensure seamless application of transitions and effects, we designed several patterns, each pattern consisting of eight transitions and effects that work well with each other. Arousal scores are assigned to these patterns so they can be chosen based on a song segment's arousal score. We selected arousal instead of valence because the visual effects mainly express impressions between calm and energetic. In addition, we use the music arousal score, rather than the photo event arousal score, because the musical moods are more consistent and dominant in the presentation.

After determining the transition/effect pattern, facial zoom effects are integrated to highlight people, and finally all transitions and effects are aligned so that they begin and end on a musical beat. We store a complete description of photos, photo events, music segments, photo and transition effects, presentation timestamps, as well as further textual descriptions, in a generic XML slideshow description format.

3.2.5. *Slideshow Rendering*

Conclusively, the slideshow description must be rendered into a sharable format so that the user can enjoy it on their home computer or internet connected device. The technology presented here was developed to be agnostic to the video rendering engine, though it easily can support a variety of renderers via an XML transformation, e.g SMIL [12].

For the developed web system we utilize a web rendering SDK from Stupeflix [13], which allows for submitting a slideshow description in XML format and then downloading the rendered video.

4. RESULTS AND DISCUSSION

We tested our system with photo albums provided by users who consider themselves consumer-grade photographers. We asked them to provide typical photo albums, such as a vacation trip, which they would share with friends (see Fig. 6). The overall development was embedded in an iterative system design and quality-in-use evaluation process, which is known from usability research [14]. This method enables benchmarking through continuous user feedback on incremental versions of the algorithm. Selected improvements identified

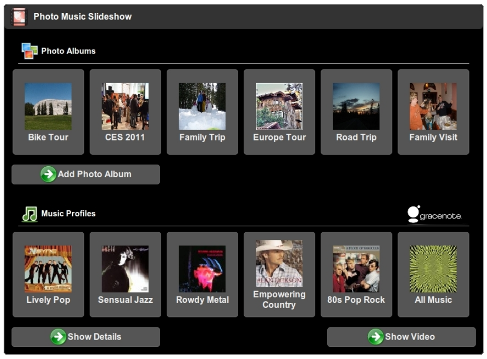

Fig. 6. Web based demo user interface that shows photo albums and predefined music profiles.

and realized throughout this process consisted of: using patterns of photo effects instead of single effects according to current music arousal, preferring chorus music segments over other segment types, minimum play-time of a song segment should not be shorter than 15 seconds, “Ken Burns” effects applied to every photo increased the appealing clearly, and stretching photo event valence and arousal values across the complete album to the full range of available music increases the diversity of music and the user satisfaction.

In addition to the quality-in-use evaluation, a survey similar to [4] was considered, but first results made it obvious that the presentations were more appealing than randomly chosen songs. To provide a more detailed analysis of the algorithm, we are developing a user test which measures the perceived difference in arousal and valence values between paired photo and music segments.

In the test, users are presented with photos, sans any music or transitions, and asked to describe how the current photo event’s overall emotional impression differs from the previous photo-event. The same questions are asked, sans any photos, concerning the selected song segments. This information is compared with the changes in arousal and valence estimated by the algorithm. Additionally, we ask them whether the selected music and visual effects are appropriate for the images.

Multiple pre-tests suggested several strengths and weakness of the algorithm. The approximation of the image’s valence was more accurate than the image’s arousal, especially for photo albums focused on people, and the arousal scores for music where more accurate than the valence scores for music. As well, the questionnaire for the music-image-effect combination indicated that in some situations, counter-selection of emotional parameters for music is preferred by the users, e.g. “thanks to the dynamic music the photos from this event are not so boring any more.”

5. CONCLUSIONS

A new approach for consumer slideshow presentations has been developed, which preserves the chronological ordering of photos as they were taken while enabling content-aware music selections. This new technology can power a wide range of application scenarios in service-oriented environments as well as in consumer software.

The iterative system development and evaluation methods led to several improvements of the technology and a substantially increased user satisfaction. User feedback also identified additional unimplemented features that will improve the overall system, such as photo scene recognition to enable even closer semantic mapping between music and photo events.

6. REFERENCES

[1] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, “Time as essence for photo browsing through personal digital libraries,” in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM Press New York, NY, USA, 2002.

[2] J.C. Chen, W.T. Chu, J.H. Kuo, C.Y. Weng, and J.L. Wu, “Tiling slideshow,” in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM Press New York, NY, USA, 2006, pp. 25–34.

[3] P. Dunker, C. Dittmar, A. Begau, S. Nowak, and M. Gruhne, “Semantic high-level features for automated cross-modal slideshow generation,” in *Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing*, 2009.

[4] Chin-Han Chen, Ming-Fang Weng, Shyh-Kang Jeng, and Yung-Yu Chuang, “Emotion-based music visualization using photos,” in *Advances in Multimedia Modeling*, vol. 4903 of *Lecture Notes in Computer Science*, pp. 358–368. Springer Berlin / Heidelberg, 2008.

[5] “Gracenote,” www.gracenote.com (April 22, 2011).

[6] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. 452–455.

[7] P. Dunker, S. Nowak, A. Begau, and C. Lanz, “Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach,” in *1st ACM international conference on Multimedia information retrieval*. ACM, 2008.

[8] D.P.W. Ellis, “Beat tracking by dynamic programming,” *Journal of New Music Research*, 2007.

[9] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” *Image Processing, IEEE Transactions on*, vol. 11, no. 4, pp. 467–476, 2002.

[10] Marc Wick, “Geonames,” www.geonames.org (April 22, 2011).

[11] AC Loui, A. Savakis, E.K. Co, and NY Rochester, “Automated event clustering and quality screening of consumer pictures for digital albuming,” *Multimedia, IEEE Transactions on*, vol. 5, no. 3, pp. 390–402, 2003.

[12] J. Ayars, D. Bulterman, A. Cohen, K. Day, E. Hodge, P. Hoschka, E. Hyche, M. Jourdan, M. Kim, K. Kubota, et al., “Synchronized multimedia integration language (SMIL 2.0),” *World Wide Web Consortium Recommendation, Aug*, 2001.

[13] “Stupeflix,” www.stupeflix.com (April 22, 2011).

[14] Jakob Nielsen, “Iterative user interface design,” *IEEE Computer*, 1993.