This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

Probabilistic Approach to People-Centric Photo Selection and Sequencing

Vassilios Vonikakis, *Member, IEEE,* Ramanathan Subramanian, *Senior Member, IEEE,* Jonas Arnfred, *Member, IEEE,* and Stefan Winkler, *Senior Member, IEEE*

***Abstract*—We present a crowdsourcing (CS) study to examine how specific attributes *probabilistically* affect the selection and sequencing of images from personal photo collections. 13 image attributes are explored, including 7 people-centric properties. We first propose a novel dataset shaping technique based on Mixed Integer Linear Programming (MILP) to identify a subset of photos in which the attributes of interest are uniformly distributed and minimally correlated. Shaping enables the synthesis of compact, balanced and representative datasets for CS, and facilitates effective learning of the selection likelihood of an image as well as its relative position in a sequence, given its attributes. We further present an ILP-based slideshow creation framework to select and arrange (a subset of) *appealing* images from a personal photo library. Quantitative and qualitative evaluations confirm that our method outperforms regression-based and greedy approaches for photo selection and sequencing, generating slideshows similar in quality to those created by humans.**

***Index Terms*—Crowdsourcing, Slideshow Creation, Personal Photo Libraries, Image Appeal, Mixed Integer Linear Programming.**

I. Introduction

THE ease with which photos can be captured and stored today has resulted in the proliferation of *personal* photo-libraries. Interacting with these libraries can be tedious, especially if one is interested in images of a specific person or group. As such, there has been extensive work on assisting users to interact with large personal photo-collections. Browsing, summarization and organization techniques have been developed based on four main approaches: **Event-based**, where photo-collections are grouped and analyzed based on specific events, acts or scenes [1]–[4] (e.g. *"my latest vacation trip"*); **Location-based** via GPS data [5] or image content [6] (*"photos I took at the Eiffel tower"*); **Time-based** using EXIF timestamps [7], [8] (*"photos I took last month"*); **Attribute-based** using lower-level attributes such as global color and texture [8], [9] (*"all bluish images"*), or high-level semantics [10] (*"images taken on a rainy day"*).

While the aforementioned approaches and their combinations offer diverse ways of interacting with photo albums, they do not account for a critical characteristic of personal photo-collections: the majority of images depict *people* performing various activities such as traveling, sports, enjoying family moments, gatherings with friends, etc. Moreover, there exists substantial evidence that photos with people are inherently different from other types of images, with respect to:

- **Image memorability**: The presence of humans is a key indicator of how memorable an image is [11].
- **Image appeal**: The presence of people in a photo, and the facial expressions they exhibit, are found to be the second and fifth major factors influencing image appeal, among 38 attributes [12]. Also, images with faces were found to attract more *likes* in social networks [13].
- **Attention/saliency**: Humans direct their gaze towards faces [14], [15], while low-level saliency rules break down when people are present in a scene [16].
- **Emotions**: Facial expressions in photos are found to emotionally impact viewers [17].

The above observations suggest that human presence in images impacts visual perception and viewer experience. Consequently, contrary to other types of images, people-centric photo collections are more challenging to browse/organize/summarize, since instances of a specific person (or group) may span different events, locations and time-periods. Therefore, traditional event-, location-, time- or attribute-based approaches are not ideal for *People-Centric Browsing* (PCB), where these attributes become less relevant. Typical PCB use-cases include (but are not limited to) (a) wanting to see photos of someone close, (b) compilation of a slideshow/collage for recalling life memories during an anniversary/birthday, and (c) visually introducing a person to others through pictures of him/her. Fig. 1 depicts the typical sequence of steps for synthesizing a people-centric photo collection/slideshow.

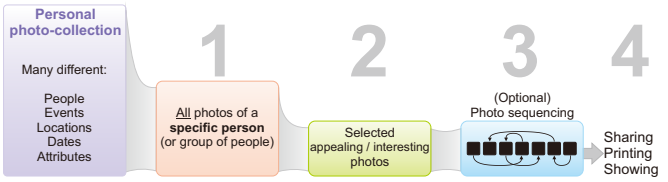

Fig. 1. **People-centric slideshow creation:** Flow diagram of steps (numbered 1–4) for generating a people-centric photo collection/slideshow.

Because manual selection of person-specific images from a large photo-collection (stage 1 in Fig. 1) is onerous, a number of computational approaches have recently been proposed. These include *social context* and *co-occurrence* [18]–[20], *clothing* [21], [22], *timestamps/geo-tagging* [23], *clustering* [24] or *interactive tagging* [25]. Additionally, commercial

Vassilios Vonikakis and Stefan Winkler are with the Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore. (Email: bbonik@adsc.com.sg, stefan.winkler@adsc.com.sg)

Ramanathan Subramanian is with the International Institute of Information Technology, Hyderabad (India). (Email: s.ramanathan@iiit.ac.in)

Jonas Arnfred is with Amazon, UK. (Email: jonas@ifany.org)

applications such as Apple iPhoto or Google Picasa offer semi-automatic ways for face tagging. Although person/group identification in photos is a critical first step for PCB, creation of a pleasing photo slideshow involves efficient *selection* and *sequencing* (stages 2 and 3 in Fig. 1) of *appealing* photos for optimal user experience. Automating the selection and sequencing processes entails many research questions such as:

- How many images should be included in a slideshow?
- Should the person of interest be alone or with others?
- How do emotions exhibited by people affect selection?
- What are the appeal requirements of the included photos?
- Do selection and sequencing characteristics vary among user groups (*e.g.*, male/female)?

Our previous works [26], [27] have affirmed that selecting images for PCB is not an arbitrary process. Underlying factors such as facial expressions influence the selection of images and their relative positions in a slideshow. This work extends these findings, while attempting to answer the aforementioned questions. More specifically, our objective is to build an automatic system for assisting users in stages 2 and 3 of Fig. 1 by generating *people-centric* slideshows, utilizing the preferences of the *average user* learnt via crowdsourcing (CS). These collections/slideshows may directly be used for presentation/sharing or serve as a baseline for customizations. Overall, our approach combines ideas from human-centered computing, image appeal and CS.

Similar to [11], [26], [28], [29], and in contrast with typical rating studies, we use CS for an *exploratory* task; workers perform photo selection and sequencing without any explicit directives, while we attempt to discover and model their preferences in order to automate the process. However, none of the prior studies have paid particular attention to image attribute distributions or the correlations among them. The danger here is threefold: (i) Certain attribute values may be under/over-represented in the photo-collection, constraining user preferences; (ii) Strong correlations among attributes can confound interpretation of user preferences; (iii) Success of CS studies hinges on presenting workers with small chunks of representative data, so as to maximize their engagement and task performance.

To this end, we introduce a new *dataset shaping* technique based on Mixed Integer Linear Programming (MILP). The proposed optimization leverages on the (possible) redundancies in a large dataset to generate a more compact version with a *specified target distribution* across each dimension, while simultaneously minimizing linear correlations between dimensions. It improves over [30], which does not minimize cross-dimensional correlations. Our proposed technique makes large-scale exploratory CS studies feasible and valid, since workers are provided with small-yet-representative chunks from the original dataset. Fewer items make workers' task more manageable and improve the quality of CS results.

Another salient aspect of this work is the adoption of a probabilistic approach for determining the selection likelihood of an image as well as its relative placement in the slideshow sequence. Prior studies that analyzed the impact of low-level attributes on image selection – including ours – [26]–[29] used regression models to provide *generic* insights on how image attributes influence image appeal. In contrast, the proposed probabilistic framework provides *detailed* insights on how an image attribute can influence photo selection across the whole range of values. Finally, while the significance of affect/emotions in images with people is well-known [1], [28], [31], it has hardly been examined in depth by studies on image appeal. This work expressly examines the impact of emotional dimensions, namely *valence* (indicating pleasant/unpleasantness) and *intensity*, on photo selection and sequencing. Overall, our work makes the following contributions:

1) We present a large-scale CS study for discovering user preferences regarding the compilation of people-centric photo summaries or slideshows. The study provides insights on how the variations in 13 image attributes, including 7 people-centric traits, affect the selection likelihood and arrangement of photos. In particular, it is the first work to expressly study the influence of affective dimensions on image appeal.
2) We propose a novel MILP-based dataset shaping technique, allowing the enforcement of specific target distributions across dimensions, while minimizing correlations among them. The technique can be used to compile compact-yet-representative subsets, which is useful for exploratory CS. A Matlab implementation of the technique is available for download (Section VIII).
3) We propose a novel ILP-based technique for automatically generating people-centric slideshows, employing probabilistic knowledge learned from CS regarding photo selection and arrangement.

The rest of the paper is organized as follows. Section II reviews the literature to motivate this work and highlight its novelty. Section III provides a detailed description of the dataset shaping method used for the CS study and an ILP-based technique for automatic people-centric slideshow creation. Section IV and V respectively discuss the protocol adopted for the CS study and related findings. Quantitative and qualitative evaluation is presented in Section VI, while conclusions are provided in Section VII. Finally, details about the available Matlab implementations of the proposed techniques can be found in Section VIII.

II. Related Work

Broadly, our work lies at the intersection of image appeal and CS, sharing common elements with previous works from both domains as elaborated below.

CS has been the driving force in many image-related studies lately. Generally, it is used either to directly annotate datasets or rate/rank images/videos according to specific attributes. For example, a large dataset of landscape images is annotated with semantic tags using CS to train multiple regressors in [10]. Some CS works employ *gamification*, where the main task is structured as part of a game. Notable examples are *Epitome* [32] where workers summarize photo albums, and *Memory* [11], where the performance of workers in a memory game is used to study image memorability. In all the

above cases, CS data is filtered by modeling workers' biases and variance using known labels [33], or following specific practices to improve data quality [34].

In other cases, CS is used for *discovering* user preferences in subjective tasks, and for learning about the 'average' user. Our work falls in this category, and is similar to [28] where user preferences for location-based photo summarization are learned using RankSVM. Filtering the CS results for such exploratory analysis is challenging, since no 'ground truth' really exists. In this respect, we adopt the approach of [35] by embedding *microtasks* in the CS experiment, which have been shown to improve workers' engagement, and provide a reliable way of benchmarking their performance (see Section IV).

Our work also examines aesthetics and image appeal, for which a significant body of work exists. For example, ACQUINE [36] is a system that rates user-uploaded photos for aesthetic quality, based on real-time prediction using SVMs. Compositional photographic rules are learned in [37], enabling optimal cropping of panoramic images. Other approaches include multi-scale decomposition, perceptual understanding and tuning via psychophysical studies [38], [39]. Most recent works on image appeal [40], [41] compare CS with lab-based results to discover which practices increase reproducibility. Nevertheless, they use generic images, which do not specifically focus on people. Furthermore, no particular attention is given to the prior distribution of image attributes in the examined datasets.

Savakis *et al.* [12] show that *image appeal* is more complicated than mere image quality and aesthetics, and discover a number of attributes contributing to it, human presence being among the key ones. Based on these findings, other works have studied the impact of human-related attributes on image appeal and aesthetics. For example, Khan and Vogel [42] analyze facial composition rules in portrait images, while Obrador and Moroney [43] show the importance of sharpness and mean luminance difference between faces and background. Perhaps the most extensive work concerning aesthetics for PCB is [31], with which our work shares similarities. Its authors consider pose, social and facial expression-related attributes (mouth openness and prolongation, eye openness and texture) and present methods for automatic selection and cropping of images from personal photo collections [44].

In this work, we adopt face composition attributes from [42], and follow a more detailed approach towards understanding the impact of facial emotion-related attributes on image appeal. More specifically, instead of classifying heuristic affective traits [44] or the 7 prototypical Ekman emotions [45], we follow the *dimensional* approach [46] where continuous values for the VA and IN affective attributes are estimated via regressors trained on the annotated Radboud dataset [47].

III. A Probabilistic Approach to Image Selection and Sequencing

Our main objective is to discover and understand criteria according to which users *select* and *arrange* photos of people, from personal photo-collections. More specifically, we seek to learn how photo attributes affect the likelihood that an image will be (a) selected (among many others) and (b) placed at a specific location in a sequence. This requires modeling of the relationships among image attributes and selection/placement criteria for automated slideshow creation.

From a probabilistic viewpoint, this requires estimating the underlying conditional probability $P_j(sel=1|a_m,t)$; the probability that image j will be selected (among many others) for the t^{th} temporal segment of a sequence, given that its m^{th} visual attribute is a_m. Since it is reasonable to assume that the prior for the m^{th} attribute is independent of the temporal distribution, *i.e.*, $P(a_m) \perp\!\!\!\perp P(t)$, we have:

$$P_j(sel=1|a_m,t) = \frac{P(a_m,t|sel=1)\,P(sel=1)}{P(a_m)\,P(t)} \qquad (1)$$

$P(a_m,t|sel=1)$ expresses the probability that – for the t^{th} temporal segment in a sequence – the m^{th} visual attribute computed over all *selected* images equals a_m. This probability can be estimated from (a sufficient number of) previously observed image selections, performed by either a single user or a group (crowd). In the former case it will reflect individual preferences, whereas in the latter it will reflect the preferences of the 'average' user. We estimate $P(a_m,t|sel=1)$ from a large-scale CS study as described in Section IV. However, the same approach could also be applied to learn individual user preferences for a personalized approach to image selection and sequencing. Marginalizing Eq. (1) over time, $P_j(sel=1|a_m)$ denotes the chance of image j being selected if its m^{th} visual attribute is a_m.

A. Dataset Shaping

In our exploratory CS study, workers need to perform selection and arrangement of photos by examining the *entire* dataset. This procedure involves a number of challenges which need to be surmounted to render the CS study valid. Firstly, the CS scenario necessitates narrowing down the number of images that workers have to interact with to a *manageable* amount. This is due to the fact that large sets are difficult to browse, and would not allow workers to exhibit the same level of engagement throughout the task. At the same time, the narrowed-down subset should have enough diversity so as to include a variety of attributes, both *low-level* (*e.g.*, sharpness or colorfulness) and *high-level* (*e.g.*, portrayed emotions).

Secondly, datasets should be *balanced* to minimize any selection biases. This implies that each image attribute should have an (approximately) uniform prior, and any deviation from a uniform prior in the submitted results would be reflective of workers' preferences, representing a potential selection criterion. Consequently, as long as the priors for all image attributes are uniform ($P(a_m) \sim U$), $P(sel=1|a_m,t)$ can be estimated from $P(a_m,t|sel=1)$, since by definition $P(t) \sim U$ and $P(sel=1)$ can be obtained from photo selection statistics. If this requirement is not met however (*e.g.*, randomly picking the initial set of photos), many image attributes may not be fully represented across their possible range of values; *e.g.*, low and medium colorfulness images may be included, but not high colorfulness. Fig. 2(b) depicts this situation for a toy

2-dimensional example (here dimensions can be viewed as image attributes). Although the distribution along the y axis is uniform, not all values are equally represented along x.

Thirdly, all attributes should be as uncorrelated as possible. Failing to do so could potentially cause problems both in interpretation and prediction. For example, if two attributes are strongly correlated, an increased selection of images corresponding to either attribute would not clearly reflect user preferences as either attribute may be responsible for the observed selections. Additionally, it could also result in increased multicollinearity while building regression models. Fig. 2(a) depicts an instance with strong linear correlation between two dimensions. Although the distribution is approximately uniform in both dimensions, a strong linear correlation would cause the aforementioned problems.

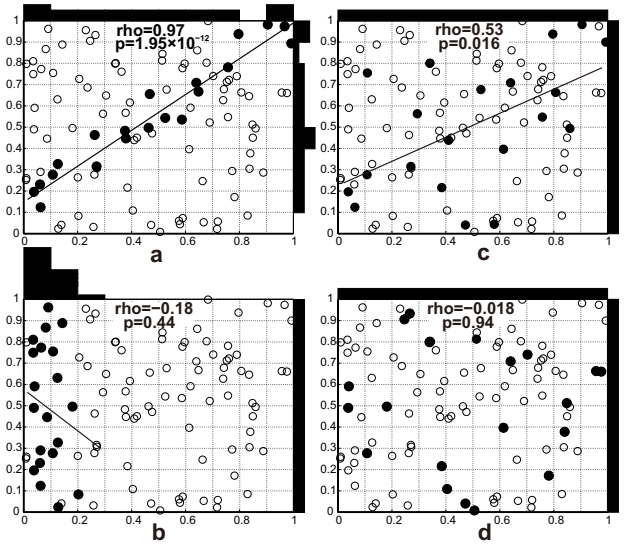

Fig. 2. **Toy 2D example for dataset shaping:** Different ways of selecting 20 out of 100 *identical* data points. Filled circles represent selected data, while distribution along x and y is depicted via a 10-bin histogram along the edges of each plot. (a) Strong positive correlation between dimensions. (b) Underrepresented x axis. (c) Enforcing a uniform distribution along both dimensions using Eq. (4). (d) Enforcing uniform priors while minimizing cross-correlations using Eq. (9).

All the above specifications can be summarized as follows:

1) The **dataset size** should be sufficiently large in order to include enough variance, and at the same time manageable in size so that the workers can pay attention to all images.
2) All attributes should preferably have a **uniform distribution** so that they are equally represented in the CS dataset, thereby minimizing any inherent bias.
3) **Correlations between attributes** should be minimized.

The first two specifications (dataset size and distribution) are addressed in Section III-A1, while the third is discussed in Section III-A2. These techniques are general in nature, and could be used in other areas apart from CS. For example, our dataset shaping technique may be used in machine learning to create balanced training subsets from larger unbalanced datasets, or to evaluate performance of an algorithm across datasets with different distributions. As such, our technique can be seen as a complement to dimensionality reduction: instead of reducing feature dimensions while maintaining the number of observations, we reduce the number of observations while imposing distributional constraints across various dimensions. Details about the available Matlab implementation of the proposed technique can be found in Section VIII.

1) Optimizing Dataset Size and Distribution: Let $S = \{\mathbf{q}_i \mid \mathbf{q}_i \in \mathbb{R}^M, \mathbf{q}_i \sim D_S^M\}_{i=1}^K$ be an initial set of K observations of a random variable $\mathbf{q}$ with priors D_S across M dimensions forming matrix $\mathbf{Q} = [q_{ij}]_{K\times M}$. Then, assuming that there is sufficient *redundancy* across all M dimensions, the objective is to select a subset of observations $\hat{s} \subset S$ with $\hat{s} = \{\hat{\mathbf{q}}_i \mid \hat{\mathbf{q}}_i \in S, \hat{\mathbf{q}}_i \sim D_{\hat{s}}^M\}_{i=1}^N$ and $N \ll K$, where $\hat{\mathbf{Q}} = [\hat{q}_{ij}]_{N\times M}$ denotes the reduced data matrix (or data subset). Enforcing $D_{\hat{s}}^M = D_S^M$ ensures that the subset $\hat{s}$ has the same distribution as S, and specifically enforcing $D_{\hat{s}}^M = U$ will result in a *balancing* effect. Evidently, varied target distributions may be synthesized via $D_{\hat{s}}$ depending on the problem.

Let matrix $\mathbf{D} \in \mathbb{R}^{H\times M}$ represent the target distribution $D_{\hat{s}}^M$ such that each of its columns $\mathbf{D}_{*j}$ contains the probability mass function (PMF) of $D_{\hat{s}}^M$ across the j^{th} dimension, quantized into H intervals. Let $B = \{\mathbf{B}^m\}_{m=1}^M$ denote a set of M binary matrices, with $\mathbf{B}^m \in \mathbb{Z}_2^{H\times K}$, such that each binary element b_{ij}^m denotes whether or not the j^{th} item of S belongs to the i^{th} interval of the target PMF, for dimension m. Finally, we introduce a binary vector $\mathbf{x} \in \mathbb{Z}_2^K$ such that each element x_i is a *decision variable* determining whether the i^{th} item of S belongs to subset $\hat{s}$. The problem can then be formulated as:

$$\min_{\mathbf{x}} \sum_{m=1}^{M} \|\mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m}\|_1 \ \text{ s.t.} \ \|\mathbf{x}\|_1 = N \tag{2}$$

which implies: *select those N elements from S that minimize the L1 distance from the target PMF and thus approximate* $D_{\hat{s}}^M$. The above minimization can be solved by introducing auxiliary vectors $Z = \{\mathbf{z}_i\}_{i=1}^M$ with $\mathbf{z}_i \in \mathbb{R}_+^H$ so that:

$$\left.\begin{aligned}\mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m} &\le \mathbf{z}_m\\ \mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m} &\ge -\mathbf{z}_m\end{aligned}\right\} \Rightarrow \left.\begin{aligned}\mathbf{B}^m\mathbf{x} - \mathbf{z}_m &\le N\mathbf{D}_{*m}\\ -\mathbf{B}^m\mathbf{x} - \mathbf{z}_m &\le -N\mathbf{D}_{*m}\end{aligned}\right\} \tag{3}$$

$\forall$ dimensions m and minimizing over Z. The final optimization can be expressed as a MILP:

$$\text{Minimize } \mathbf{c}^\top\tilde{\mathbf{x}} \text{ s.t. } \mathbf{A}\tilde{\mathbf{x}} \le \mathbf{b} \tag{4}$$

with $\mathbf{c} = \begin{bmatrix}\mathbf{0}_K^\top & \mathbf{1}_{HM}^\top\end{bmatrix}^\top$, $\tilde{\mathbf{x}} = \begin{bmatrix}\mathbf{x}^\top & \mathbf{z}_1^\top \cdots \mathbf{z}_M^\top\end{bmatrix}^\top$ and

$$\mathbf{A} = \begin{bmatrix} \mathbf{1}_K^\top & \mathbf{0}_{HM}^\top \\ -\mathbf{1}_K^\top & \mathbf{0}_{HM}^\top \\ \hline \mathbf{B}^1 & \\ \vdots & -\mathbf{I}_{HM} \\ \mathbf{B}^M & \\ \hline -\mathbf{B}^1 & \\ \vdots & -\mathbf{I}_{HM} \\ -\mathbf{B}^M & \end{bmatrix}, \mathbf{b} = \begin{bmatrix} N \\ -N \\ \hline N\mathbf{D}_{*1} \\ \vdots \\ N\mathbf{D}_{*M} \\ \hline -N\mathbf{D}_{*1} \\ \vdots \\ -N\mathbf{D}_{*M} \end{bmatrix}$$

where $(.)^\top$ denotes transpose, $\mathbf{A} \in \mathbb{Z}^{(2+2HM)\times(K+HM)}$, $\mathbf{b} \in \mathbb{R}^{(2+2HM)}$ and $\mathbf{c} \in \mathbb{Z}_2^{K+HM}$, while $\tilde{\mathbf{x}}$ is also of size $K+HM$ and contains both integer and real optimization variables. The first two rows of $\mathbf{A}$ and $\mathbf{b}$ address the equality constraint of the integer variables $\|\mathbf{x}\|_1 = N$, expressed as two inequality

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

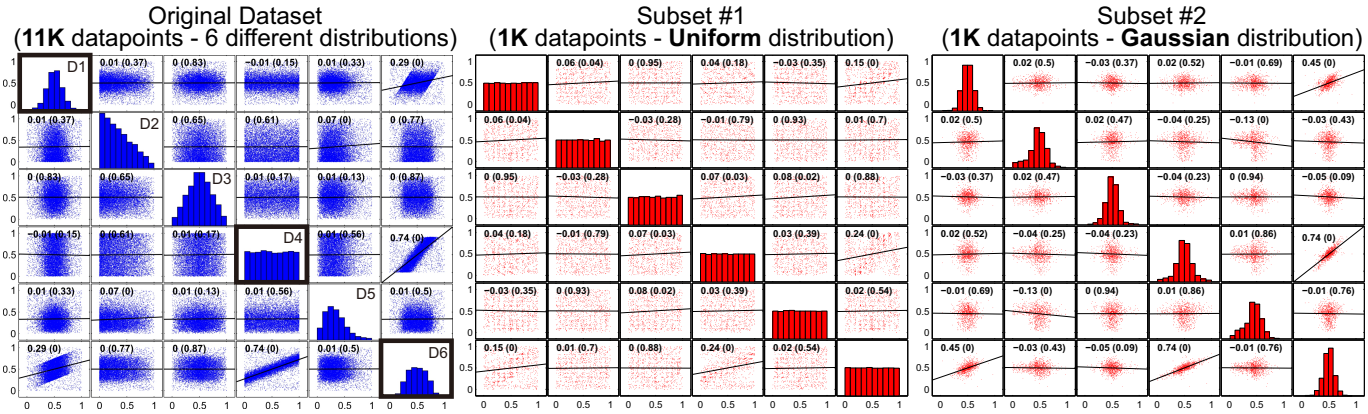

Fig. 3. **Data shaping illustration:** (best viewed under zoom) Covariance scatter plots for a 6-dimensional dataset with 11,000 data points. Distribution for each dimension is given by a histogram, while Pearson correlation (ρ) between dimensions and corresponding p-value (in parentheses) are mentioned for each scatter plot. Dimension 6 (D6) is a linear combination of D1 and D4. Two subsets of 1,000 datapoints are generated with our data shaping technique, so as to have Uniform and Gaussian distributions. Minimal correlations between sampled data points are enforced via Eq. (9).

constraints $\sum_{i=1}^{K} x_i \leq N$ and $-\sum_{i=1}^{K} x_i \leq -N$. The lower two sections address the constraints for the real auxiliary variables from the upper and lower parts of Eq. (3).

MILP problems are NP-hard combinatorial problems. However, modern branch and bound algorithms can solve many real world problems quickly and reliably [48]. Such algorithms solve the LP relaxation problem to obtain fractional solutions and create two sub-branches by adding new constraints [49]. As an indication, Matlab's *intlinprog* solver takes approximately one second for solving Eq. (4) for 3500 integer variables and 130 constraints on a typical quad-core PC with 8GB RAM.

2) Minimizing Cross-dimensional Correlations: We now need to minimize cross-dimensional correlations in the selected subset $\hat{s}$. Since correlation is a scaled version of covariance, one can minimize the latter instead. This is done by diagonalizing the covariance matrix of $\hat{\mathbf{Q}}$ using the following approach:

$$\min \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} \left| \mathrm{cov}\left(\hat{\mathbf{Q}}_{*m}, \hat{\mathbf{Q}}_{*n}\right) \right| \tag{5}$$

which essentially means: *minimize the sum of absolute covariances of all possible* $\binom{M}{2}$ *combinations from* M *dimensions of the selected subset* $\hat{s}$. The covariance between two dimensions of the data matrix $\hat{\mathbf{Q}}$ can be expressed relative to the elements of matrix $\mathbf{Q}$ using the decision variables of vector $\mathbf{x}$ as follows:

$$\begin{aligned} \mathrm{cov}\left(\hat{\mathbf{Q}}_{*m}, \hat{\mathbf{Q}}_{*n}\right) &= \sum_{j=1}^{N} \left(\hat{q}_{jm} - \overline{\hat{q}_{*m}}\right)\left(\hat{q}_{jn} - \overline{\hat{q}_{*n}}\right) = \\ &= \sum_{j=1}^{K} x_j \left(q_{jm} - \overline{\hat{q}_{*m}}\right)\left(q_{jn} - \overline{\hat{q}_{*n}}\right) \end{aligned} \tag{6}$$

where $\overline{\hat{q}_{*n}}$ is the mean of the n^{th} column of $\hat{\mathbf{Q}}$. The absolute value of the covariance is bounded as a result of the triangle inequality:

$$\begin{aligned} 0 \leq \left| \sum_{j=1}^{K} x_j \left(q_{jm} - \overline{\hat{q}_{*m}}\right)\left(q_{jn} - \overline{\hat{q}_{*n}}\right) \right| \leq \\ \leq \sum_{j=1}^{K} x_j \left|q_{jm} - \overline{\hat{q}_{*m}}\right| \left|q_{jn} - \overline{\hat{q}_{*n}}\right| \end{aligned} \tag{7}$$

This means that instead of minimizing Eq. (5), we can minimize the upper bound of each covariance as Eq. (7) indicates. This is expressed as follows.

$$\begin{aligned} &\min \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} \sum_{j=1}^{K} x_j \left|q_{jm} - \overline{\hat{q}_{*m}}\right| \left|q_{jn} - \overline{\hat{q}_{*n}}\right| \equiv \\ &\equiv \min \sum_{j=1}^{K} x_j \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} \left|q_{jm} - \overline{\hat{q}_{*m}}\right| \left|q_{jn} - \overline{\hat{q}_{*n}}\right| \equiv \min \mathbf{v}^{\top}\mathbf{x} \end{aligned} \tag{8}$$

where $\mathbf{v} = [v_1, \cdots, v_K]^{\top}$, $v_i = \sum_{m=1}^{M-1}\sum_{n=m+1}^{M} \left|q_{im} - \overline{\hat{q}_{*m}}\right| \left|q_{in} - \overline{\hat{q}_{*n}}\right|$. This can be combined with the objective of Section III-A1 in the same MILP by substituting the zero vector $\mathbf{0}_K$ of $\mathbf{c}$ in Eq. (4) with $\mathbf{v}$:

$$\mathbf{c} = \left[\lambda \mathbf{v}^{\top} \quad \mathbf{1}_{HM}^{\top}\right]^{\top} \tag{9}$$

where λ is a scalar that controls the relative weighting of the two objectives. When $\lambda = 0$, Eq. (9) is transformed to Eq. (4), and the cross-dimensional correlation objective is not applied. Higher values of λ will increase its contribution by introducing a penalty weight to each observation as defined by $\mathbf{v}$. Observations contributing more to increased correlation between dimensions are penalized, and thus have lower probability of being selected. However, calculating $\mathbf{v}$ requires prior knowledge of the mean value $\overline{\hat{q}_{*n}}$ in each of the dimensions of the final subset $\hat{s}$. This is achieved by the first objective, which enforces the target distribution $D_{\hat{s}}^{M}$ on the final subset $\hat{s}$. Consequently, the value of λ should be selected so as to balance the impact of the two objectives. In this work, we set $\lambda = 0.5$.

Figs. 2(c) and 2(d) depict the output of Eq. (4) and Eq. (9) respectively, for a simple 2D example in which $D_{\hat{s}}^{M} = U(0,1)$. Although the objective function of Eq. (4) achieves the target uniform distribution, it does not minimize the correlation between the two dimensions ($\rho = 0.53$ and $p = 0.016$). On the other hand, Eq. (9) achieves both objectives: the target distribution is approximated with minimum linear correlation between the two dimensions ($\rho = -0.018$).

Fig. 3 depicts a more thorough demonstration of the proposed shaping technique, on a larger ($N = 11{,}000$) 6-dimensional dataset with various distributions. Dimension 6 (D6) is a linear combination of D1 and D4. As a result, there are significant correlations ($\rho_{6,4} = 0.74$, $\rho_{6,1} = 0.29$) between

those. Our technique is able to not only enforce the target distribution in each dimension, but also reduce the resulting correlations between dimensions. For example, in Subset #1 $\rho_{6,4} = 0.24$ and $\rho_{6,1} = 0.15$.

B. Image Selection and Sequencing

On estimating $P(sel = 1|a_m, t)$ either from a CS study or from a single user, one can use it to select *unseen* images and arrange them into a sequence. Following a notation similar to Section III-A, let $S = \{\mathbf{q}_i \mid \mathbf{q}_i \in \mathbb{Z}^M\}_{i=1}^{K}$ be the initial set of people-centric photos characterized by M visual attributes. Each attribute value is assumed to be quantized into H levels (bins). Let matrix $\mathbf{B} = [b_{ij}]_{K\times M}$ contain quantized attribute values of K images in set S, such that each $b_{ij} \in \mathbb{Z}$ denotes the bin in which the j^{th} attribute of the i^{th} image belongs. The objective is to select a subset of *appealing* photos $\hat{s} \subset S$, with $\hat{s} = \{\hat{\mathbf{q}}_i \mid \hat{\mathbf{q}}_i \in S\}_{i=1}^{N}$, $N < K$, based on $P(sel = 1|a_m, t)$ and (optionally) arrange them into a sequence. The sequence is assumed to be partitioned into T temporal segments ($T \leq N$). These segments represent *stages* of the slideshow, and should not be confused with actual image positions. With $T = 3$ for example, N photos would be distributed equally to the *beginning*, *middle* and *end* of a slideshow. We assume that $P(sel = 1|a_m, t)$ is in the form of a set of M matrices (one for each visual attribute) $\mathcal{P} = \{\mathbf{P}^m\}_{m=1}^{M}$, with $\mathbf{P}^m = \left[p_{ij}^m\right]_{H\times T}$; each $p_{ij}^m \in \mathbb{R}$ denotes the probability that an image may be selected for the j^{th} sequence segment given that its m^{th} visual attribute is within the i^{th} bin. The final probability of an image j being selected for the i^{th} segment of an N-image sequence is given by the product of the selection probabilities of all attributes a_m. These probabilities are stored in matrix $\mathbf{E} = [e_{ij}]_{N\times K}$, where each $e_{ij} \in [0, 1]$ denotes the probability that image j will be selected for the i^{th} slideshow position.

$$e_{ij} = \prod_{m=1}^{M} P_j(sel = 1|a_{jm} = k, t = d(i)) = \prod_{m=1}^{M} p_{kd(i)}^m \qquad (10)$$

where $d(x) : N \to T$ is a quantization function assigning the x^{th} out of N images to one of T temporal segments.

If image sequencing is not required, the marginal probabilities over time can be used, eliminating the temporal part of Eq. (10). In this case, e_{ij} becomes e_j and provides an indication of how *appealing* image j is, based on its visual attributes and $\mathcal{P}$, as discussed in Section V-C. When both image selection and sequencing are required, e_{ij} indicates how *appropriate* image j is for position i, according to the learnt probabilities $\mathcal{P}$. Clearly, selecting the most appropriate image for each position will result in a more appealing slideshow. Consequently, the sum of all selection probabilities for the slideshow images, $U_{\hat{s}} = \sum_{i=1}^{N} e_{i\hat{s}(i)}$, where $\hat{s}(i)$ is the image index in slideshow $\hat{s}$ at position i, can serve as an *appeal* measure. Maximizing $U_{\hat{s}}$ ensures that slideshow $\hat{s}$ is the most appealing slideshow among all possibilities, based on the learnt selection probabilities and initial image pool S. A straightforward approach is to use a *greedy* algorithm, in which we always choose the image with the highest available selection probability e for each position. However, this approach does not guarantee the maximization of $U_{\hat{s}}$, because selecting images with the highest probability e in the beginning may result in worse selections for later positions. Alternatively, a compromise (in terms of e) in some positions may result in a sequence with higher overall $U_{\hat{s}}$. Since this is a combinatorial problem, we introduce a method based on ILP that ensures maximization of $U_{\hat{s}}$.

A binary vector $\mathbf{x} \in \mathbb{Z}_2^{NK}$ is introduced, each element x_i of which is a decision variable determining whether or not the elements of $\mathbf{E}$ will be selected. The problem can then be formulated as the following ILP:

$$\max\ \mathbf{c}^\top\mathbf{x} \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b} \equiv \min\ -\mathbf{c}^\top\mathbf{x} \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b} \qquad (11)$$

where $\mathbf{c}$ is a vectorized form of $\mathbf{E}$, while $\mathbf{A} \in \mathbb{Z}_2^{(2+2N+K)\times NK}$ and $\mathbf{b} \in \mathbb{Z}^{(2+2N+K)}$ represent the following constraints:

1) Total number of selected images should strictly be N.
2) There should be strictly one image selection per position.
3) Each image can be used up exactly once.

The exact form of $\mathbf{A}$ and $\mathbf{b}$ is the following:

$$\mathbf{A} = \begin{bmatrix} \mathbf{1}_{NK}^\top \\ -\mathbf{1}_{NK}^\top \\ \hline vec(\mathbf{O}_{1*})^\top \\ -vec(\mathbf{O}_{1*})^\top \\ \vdots \\ vec(\mathbf{O}_{N*})^\top \\ -vec(\mathbf{O}_{N*})^\top \\ \hline vec(\mathbf{O}_{*1})^\top \\ \vdots \\ vec(\mathbf{O}_{*K})^\top \end{bmatrix}, \mathbf{b} = \begin{bmatrix} N \\ -N \\ \hline 1 \\ -1 \\ \vdots \\ 1 \\ -1 \\ \hline 1 \\ \vdots \\ 1 \end{bmatrix}$$

where $\mathbf{O}_{i*}$ is a $N \times K$ matrix with all 0 elements except row i which is 1. Similarly, $\mathbf{O}_{*j}$ is a $N \times K$ matrix with all 0 elements except column j which is 1. The first two rows of $\mathbf{A}$ and $\mathbf{b}$ address the equality constraint that the total number of selected images should be N. The middle section addresses the equality constraint that there should be only one image per position. Finally, the last section addresses the inequality constraint that each image can be used only once. A Matlab implementation of this ILP is available in the supplementary material (Section VIII).

IV. Crowdsourcing

A. Main Experiment

The CS experiment was set up on Amazon Mechanical Turk (AMT). 465 English-conversant workers were recruited in total. All workers were paid 50¢, and an additional 10¢were given as bonus for exceptional performance (see Section IV-C). The experiment was designed so as to replicate the actual workflow shown in Fig. 1. All workers had to go through 2 basic steps: starting from an initial balanced set of people-centric photos, they had to make image selections (step 1), and arrange their selections in a slideshow sequence (step 2). Upon task acceptance, workers were taken to a web page which presented visual instructions pertaining to the experiment. To ensure motivation, we explicitly mentioned that exceptional workers would be paid a bonus for their effort.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

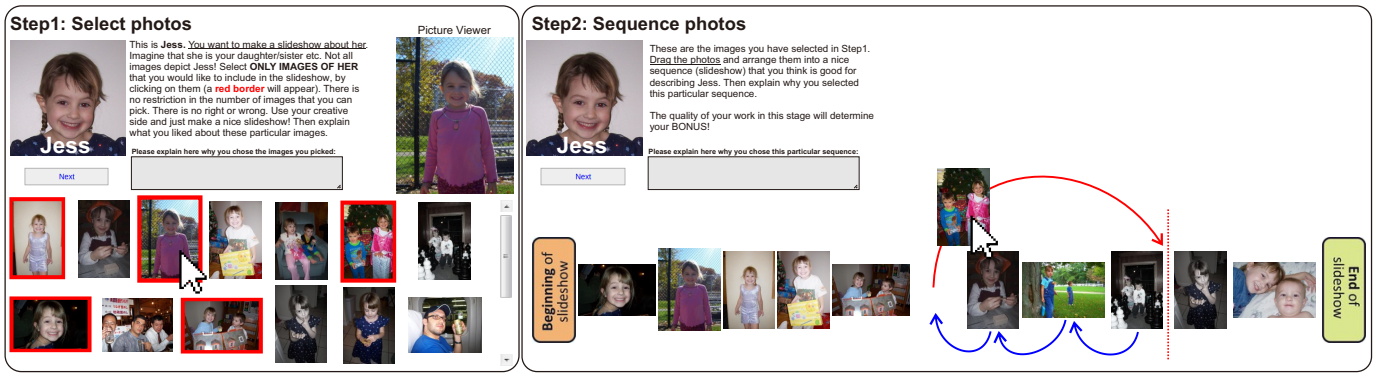

Fig. 4. Snapshots of the user interface employed in our CS study for *selection* (left) and *sequencing* (right).

TABLE I
VISUAL ATTRIBUTES USED IN THE CS EXPERIMENT.

	Attribute	Indicates	Implementation
People-centric	Face count	Group or individual	Viola-Jones frontal face detector along with skin detection to eliminate false-positives [50]
	Face size	Type of shot: close-up, full-body, long distance	Ratio of face bounding box size to image size
	Face exposure	Face exposure	Mean luminance within bounding box
	Face composition	Aesthetic impression	Minimum distance from the face center to the 5 power-points in the image (4 for the *rule of thirds* and 1 for the center of the image) [42]
	Face yaw	Frontal / profile	Yaw of the head pose as estimated by the IntraFace library [51]
	Face valence	Pleasant / unpleasant facial expression [-1,1]	Regressor trained on Radboud images with VA annotations [47] with geometric features extracted on 49 facial points detected by IntraFace [51]
	Face intensity	Neutral / apex [0,1]	Regressor trained on Radboud images with IN annotations [47] with the same geometric features as for VA
Image-centric	Capturing Period	Primacy/recency, age of person	EXIF timestamps in conjunction with character's age (only for *male*, *female*, *couple* albums)
	Scene type	Indoorness/outdoorness	Ranking function based on *Relative Attributes* [52], using gist and color histograms
	Sharpness	Level of fine details	Computed similar to [53]
	Exposure	Overall exposure	Mean value of the luminance component
	Contrast	Overall contrast	Coefficient of variation (σ/μ) for the luminance channel
	Colorfulness	Color vividness	Computed similar to [54]

To begin with, workers had to answer questions related to demographics and their familiarity with photo collections. Workers were then *randomly* assigned to one of five different albums (see Section IV-B) and were presented with a graphical user interface (GUI). Special attention was given to make the experiment *fun*, *creative* and *relaxing*, rather than a typical annotation task. More specifically, a desktop-quality GUI was designed, sharing elements with other widely used software like Windows Movie Maker. It allowed simple and intuitive image selection and browsing, as well as interaction on a film-like thumbnail timeline. No restrictions were imposed on the workers, regarding the number of selected images or the total time spent, allowing them to freely express their creativity.

During image selection, the GUI featured 4 regions as shown in Fig. 4(left): a character *introduction* area, a *thumbnail browsing* area, a *picture viewer* and a *text box*. The introduction area included the name(s), portrait(s) and brief description(s) of the album character(s). This was done to make the context relevant to workers and the task more personal. The thumbnail browsing area depicted thumbnails of the album images in random order. The picture viewer displayed a magnification of any thumbnail over which the mouse pointer hovered. This way, workers could effortlessly browse images at coarse and fine levels. Clicking on a thumbnail would activate (deactivate) a red border around it, indicating that this particular image is selected (deselected). Finally, workers had to type comments justifying their selections in the textbox to ensure accountability. The sequencing task shown in Fig. 4(right) was designed similar to video editing software, and selected photos were displayed as thumbnails. The worker could freely 'drag and drop' thumbnails to any position to create the sequence of their choice. Here again, they had to justify their sequencing via a textbox. All user-specific metadata such as number of clicks, time to task-completion, and length of provided comments, were recorded. The experiment was conducted thrice corresponding to three initial set sizes (40/60/80 photos).

B. Albums and Image Attributes

Five large people-centric photo albums were compiled as the initial sources for the CS study. Each albums depicted a different central character (theme), namely, *adult male* (316 photos), *adult female* (381 photos), *couple* (281 photos), *girl* (278 photos) and *baby* (177 photos). The first three were selected from personal photo-collections, since no datasets depicting the same individual(s) over an extended period of time and spanning multiple events are available. The *girl* and *baby* albums were taken from the Gallagher dataset [21], which also includes family moments exemplifying a personal photo-collection. The approach described in Section III-A was then applied on these source datasets to acquire balanced subsets with approximately uniform distributions. Three different subsets (of size 40, 60 and 80 images) were created to explore the impact of dataset size on image selections, resulting in 15 different balanced subsets (3 sizes $\times$ 5 themes). 13 image attributes were analyzed (7 of which are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

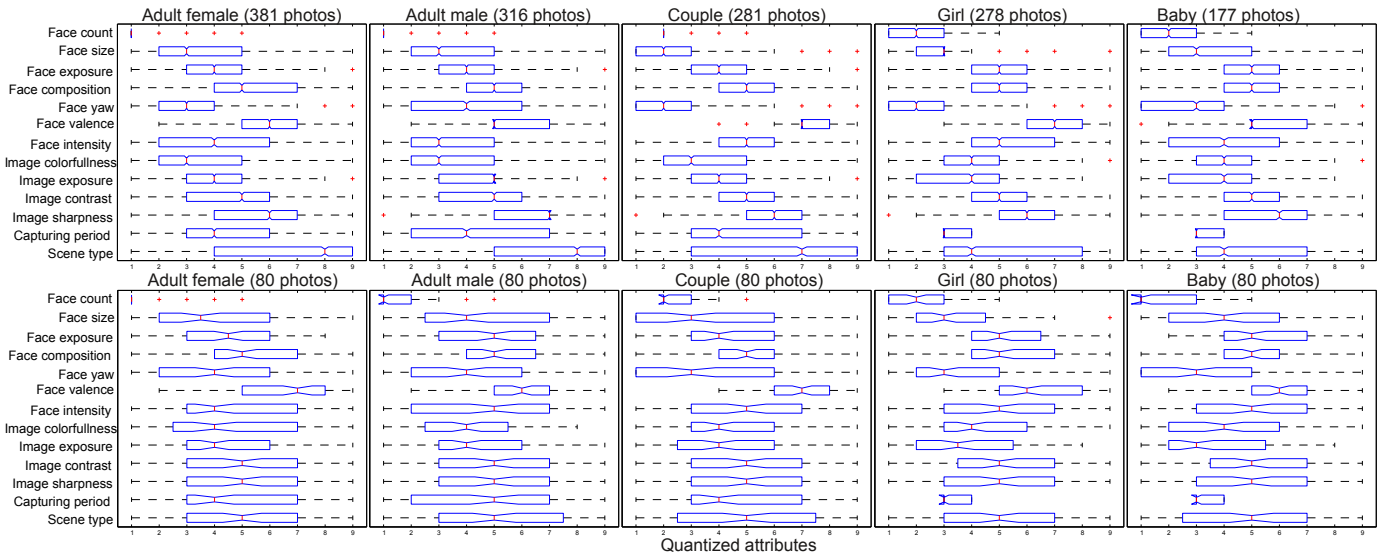

Fig. 5. **Effects of balancing:** (Top) Box and whisker plots for the quantized attributes of the original five character albums. (Bottom) Box and whisker plots for the 80-image subsets, after enforcing a *uniform* target distribution.

related to faces) to determine their influence on selection and sequencing.

Table I presents a complete list of considered attributes and their extraction methodology. Some attributes (sharpness, colorfulness, contrast and exposure) were selected based on prior works [12], [55]. Others such as face size and face count were selected to describe close-up and group photos, which have been found to influence image appeal [12]. Attributes like scene type, recency and facial affect (valence and intensity), which have not been studied previously, were also considered. All the above attributes were *min-max* normalized to $[0,1]$, and quantized to 9 discrete levels ($H = 9$) to generate the balanced subsets. For valence (VA), which spans the $[-1,1]$ interval, values of -1 and 1 were respectively assigned to the 1^{st} and 9^{th} quantization levels. There were no images with high negative valence (VA≈-1) in any album.

Upon completion of the selection and sequencing steps by crowdworkers, $P(a_m,t|sel=1)$ was estimated for 13 image attributes and 5 temporal segments ($T = 5$) among the selected images to understand user preferences. To account for small deviations from a perfectly uniform attribute prior, $P(a_m,t|sel=1)$ was normalized by $P(a_m)$ in order to get $P(sel=1|a_m,t)$. The resulting $P(sel=1|a_m,t)$ was smoothed via a two-dimensional Gaussian filter (3×3 window, $\sigma = 1$) [56].

Fig. 5 presents the results of data shaping on our CS study. The top half of Fig. 5 presents the statistical characteristics of the original 5 character albums specified above. Values for most attributes span the entire range, with the exception of face count and size, valence and capturing period (especially for the girl and baby characters). The bottom half presents the statistics of the balanced 80-image subsets used in the CS study. As mentioned previously, so long as the sufficient redundancy assumption is met, the resulting distribution is closer to uniform, with a median value close to the center of the scale and equal quartiles. Also, larger datasets have a higher chance of satisfying the redundancy assumption.

TABLE II
METADATA USED TO EVALUATE WORKER RELIABILITY, IN DESCENDING ORDER OF IMPORTANCE.

$\mathbf{a}_j$	Description	Sign of w	Indicates a worker...
a_1	Ratio of selected distractor photos over selected album photos	w_1: –	paid attention to the task (engagement)
a_2	Comments regarding selection (number of characters)	w_2: +	had a specific reason for his/her selections
a_3	Length of comments – sequencing (number of characters)	w_3: +	had a specific reason for his/her sequencing
a_4	Number of selected images	w_4: +	was engaged in the task
a_5	Time spent selecting images (s)	w_5: +	was not rushing
a_6	Time spent arranging images (s)	w_6: +	was not rushing
a_7	Time spent on instructions (s)	w_7: +	was not rushing
a_8	Time spent on questions (s)	w_8: +	was not rushing
a_9	Number of clicks in sequencing	w_9: +	was diligent in this task
a_{10}	Not owning any cameras (binary)	w_{10}: –	had no photo-experience
a_{11}	Number of wrong answers (binary)	w_{11}: –	paid attention to questions
a_{12}	Contradictory answers (binary)	w_{12}: –	paid attention to questions

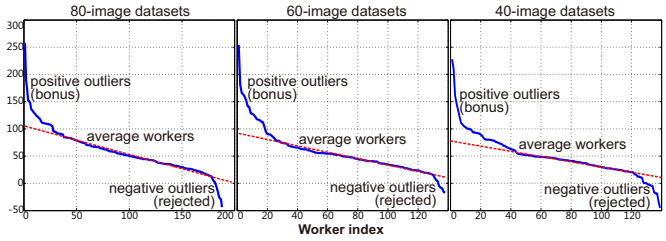

Fig. 6. **Worker evaluation:** Filtering of workers (x-axis) based on reliability score R (y-axis). On fitting a linear model (dotted red line) to the blue curve, positive outliers were paid a bonus, whereas negative outliers were rejected.

C. Evaluation of Workers

Evaluation of workers' reliability is usually challenging in CS, and more so with an *exploratory* study like ours, where workers perform a subjective task with no 'ground truth'. Consequently, techniques that attempt to model the bias and variance of workers based on known data [33] are ineffective. Therefore, we embedded *microtasks* [35] in the main experiment, which are known to improve workers' engagement and serve as indirect indicators of their reliability. More specifically, we introduced a *facial recognition* microtask

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

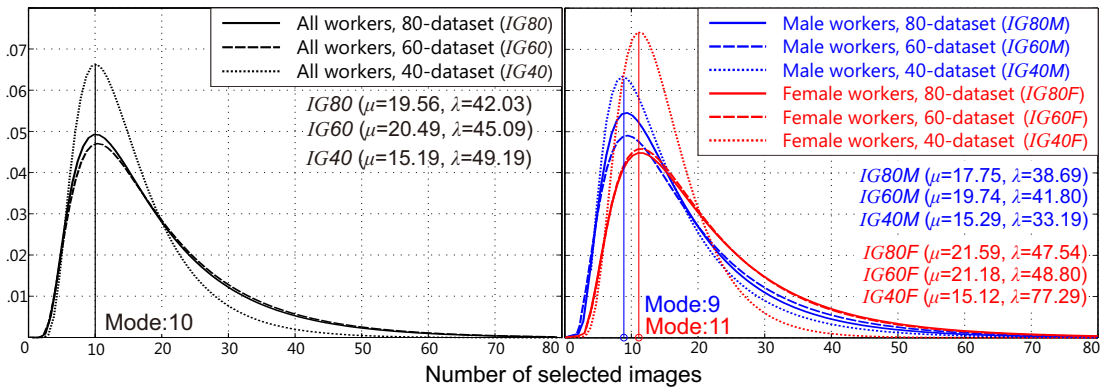

Fig. 7. IG PDF fitting (and their corresponding parameters μ and λ) for all datasets and workers (best-viewed in color).

in the main experiment: among the images that were presented to workers, some *distractor* images (comprising unknown characters) were presented (see bottom row of left part of Fig. 4). The ratio of distractor images to the number of album images was 0.375, *i.e.*, 30, 23 and 15 distractor images were shown along with the 80, 60 and 40 album subsets. Apart from the fact that the central character was missing, distractor photos were similar in all aspects to the album images. Therefore, workers had to carefully select photos containing the main character in the album. This approach allows for a *quantifiable measurement* of the worker's engagement. Results including distractor images were strong indicators of a worker not paying the necessary attention to the task.

Based on this and other important cues, the reliability R of worker j was estimated as $R_j = [\mathbf{s} \odot \mathbf{a}_j]^\top \mathbf{w}$, where $\odot$ denotes element-wise vector multiplication. $\mathbf{a}_j$ is a vector denoting worker metadata, $\mathbf{s}$ is a vector of scaling factors to account for different dimensions of $\mathbf{a}_j$, and $\mathbf{w}$ is a weight vector denoting the type of influence ($+$ or $-$) and degree of importance to each element in $\mathbf{a}_j$. Table II presents 12 types of metadata employed for estimating worker reliability. It should be noted that the actual values of the weight vector $\mathbf{w}$ are not important; only the relative weighting between them. In our case, weights were selected so as to satisfy the following relationships: $|w_1| \gg |w_2| = |w_3| > |w_4| > |w_5| = |w_6| > |w_7| = |w_8| > |w_9| > |w_{10}| = |w_{11}| = |w_{12}|$. Overall, the presence of distractor images in the final result was heavily penalized. Detailed comments, longer task completion times and more GUI clicks were rewarded, whereas lack of experience with photos and inaccuracies in responses were only mildly penalized.

The reliability score R conveys how a worker performed *relative* to others. Fig. 6 depicts the ranking of workers for the 80, 60 and 40 image datasets based on R. Three distinctive regions are evident: workers that exhibit (i) exceptional performance, (ii) average performance (the majority falls in this category), and (iii) poor performance. To identify these regions, we fit a linear model to the sorted reliability scores. All scores falling within a range of ε of the linear model were considered to be average workers. Any positive outliers were considered exceptional workers and were rewarded with a bonus of 10¢. Any negative outliers were considered poor workers, whose results were rejected. This approach has 2 main advantages – first, it provides a solution to the difficult problem of worker evaluation for a subjective task with no ground-truth. Second, all decisions are based on *relative* performance of workers, rather than 'hard' decisions based on absolute performance criteria. For example, if a worker accidentally selected one distractor image but otherwise exhibited positive performance, his/her results could still be considered for inclusion. Based on this evaluation scheme, 49 out of 465 workers ($\approx 10.5\%$) who participated in the CS study were rejected.

V. Crowdsourcing Results

This section discusses the main observations from our crowdsourcing study in terms of (a) image selection, and (b) the type of image attributes that are preferred by users (as shown in Fig. 8). On average, workers were found to use about 10 photos per people-centric slideshow, irrespective of the number of available images. Distance from the camera, face composition, facial affect (VA and IN), and image colorfulness considerably influence image selections. Significant differences between male and female workers were observed with respect to their preference for facial valence; males showed more propensity to selecting photos with negative facial emotions as compared to females.

A. Number of Selected Images

A series of probability density functions (PDFs) were fitted to the filtered CS photo selection data using Maximum Likelihood Estimation (MLE). Out of 20 different PDFs, the one that exhibited the best fit for all dataset sizes was the Inverse Gaussian (IG). Fig. 7 displays fitted IG PDFs across all CS datasets and workers, and also for male and female workers. Four conclusions can be drawn from the best-fit distributions.

- Distributions are heavily skewed, and there is considerable variability in the total number of selected photos among workers, highlighting the subjectivity of the task.
- The most frequent number of selections (mode) for a slideshow is around 10 images, irrespective of the dataset size.
- Selection variance appears to increase with dataset size up to a certain degree; while 60- and 80-image datasets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

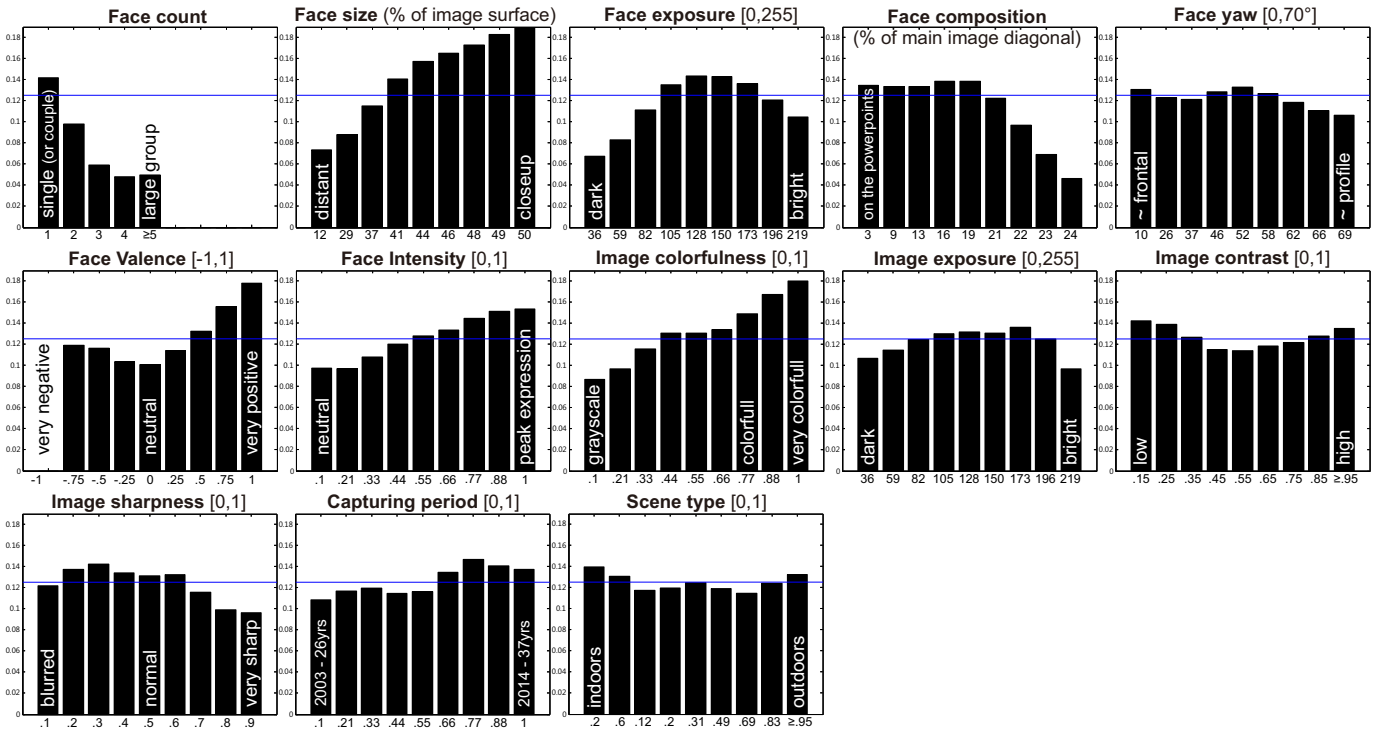

Fig. 8. **Crowdsourcing results:** Marginal selection likelihood (over time) plots from the CS study considering all five album themes. Blue line indicates random baseline, assuming a selection of 10 (most frequent choice) out of 80 images.

have very similar PDFs, variance for the 40-image dataset is smaller.

- A difference is observable between male and female workers regarding the number of selected photos. On average, females tend to select 11 photos, and males 9 photos.

B. Type of Images

In contrast to prior studies that present *general* observations like "*People prefer colorful images*", we are interested in specific trends such as "*Colorfulness values above x increase selection probability by y%*". Fig. 8 depicts marginal likelihoods (over time) for the considered 13 image attributes based on the CS study. Here we assume that $P(sel=1) = 10/80 = 0.125$, since 10 was found to be the mode for selections on the 80-image dataset. This measure can serve as a baseline, indicating the probability of random selection (blue line in Fig. 8). The depicted results refer to a *generic* photo-album case, where we combine the results for all the five themes used for the CS study.

Face count: Presence of a solitary (target) person (or two people for the *couple* album) increases image selection probability over the random baseline. At the same time, presence of more people in the photo decreases $P(sel=1|a_m)$ by more than 50% compared to the baseline.

Face size: Workers generally prefer larger faces (portaits) of the target, resulting in a clear monotonically increasing trend for face size. Interestingly, a bounding-box to image-size ratio of ≈40% seems to be a threshold for image selection. Images with face sizes above this threshold have up to 50% higher chance of selection, whereas $P(sel=1|a_m)$ for photos with small faces reduces by up to 40%.

Face exposure: High or low face exposure negatively affects image selection probability. Faces with average luminance levels in the interval $[100,200]$ increase $P(sel=1|a_m)$ by up to ≈12% over the baseline. Deviating from this interval decreases $P(sel=1|a_m)$ by up to 50%. Another interesting observation is that users are more tolerant to overexposed faces compared to underexposed ones.

Face composition negatively affects $P(sel=1|a_m)$. There is an interval where distance from the image power points does not affect selection probability. This interval spans from 0% (exactly on the power points) to ≈20% away from the power points (relative to the image diagonal). Above this threshold, $P(sel=1|a_m)$ decreases monotonically by up to 64%.

Face yaw impacts selection only weakly based on our CS results. There seems to be a slight preference for frontal and profile $\approx 45^\circ - 50^\circ$, while other head orientations decrease $P(sel=1|a_m)$ up to 16% with respect to the baseline.

Face valence (VA) is a *critical* attribute influencing image selection. Starting from *neutral*, which results in decreased $P(sel=1|a_m)$ by ≈20% relative to random, there is a strong monotonic increase in $P(sel=1|a_m)$ for positive VA. High positive VA images have a 44% higher chance of being selected. Interestingly, this preference extends to high negative VA as well, although a more gentle slope is observed in this case. Overall, users clearly prefer emotional faces over neutral ones.

Face intensity of expression (IN) positively affects selection probability. Intense facial emotions (above ≈0.5) increase $P(sel=1|a_m)$ by up to 24%, while mild/neutral emotions decrease $P(sel=1|a_m)$ by up to 20%. Overall, users prefer intense facial emotions as compared to neutral.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

Fig. 9. **ILP-based estimates of selection probability** $P(sel = 1|a_m)$**:** Top-five and bottom-five photos with the highest and lowest selection probabilities for 3 sub-albums of the Gallagher dataset [21]. Top row: 'baby' album, middle row: 'girl' album, bottom row: 'boy' album.

Image colorfulness: Colorfulness also critically influences image selection probability, and there is a roughly monotonic increase in $P(sel = 1|a_m)$ as colorfulness increases. Very colorful images have a 44% higher chance of selection, while grayscale photos have 32% lower chances compared to the baseline.

Image exposure: In a luminance range of about $[80, 200]$, image exposure does not seem to affect $P(sel = 1|a_m)$. Outside of this interval, $P(sel = 1|a_m)$ drops by up to 24% for overexposure and 15% for underexposure. Contrary to face exposure, users seem to be more forgiving of global underexposure. This could be due to the fact that photos taken at night (or darker settings) are not uncommon in personal photo-collections.

Image contrast: CS results for image contrast are not straight-forward to interpret. A slight 'V' shape pattern is observed, revealing a preference for images with high or low contrast. The latter is counter-intuitive. A possible explanation could be that low-contrast photos exhibit a 'washed-out' effect and might have been considered as artistic, similar to the filters used in applications like Instagram.

Image sharpness: Similar to contrast, image sharpness results are not easy to interpret. A general tendency is that low to mid-level sharpness is preferred over very sharp images. This is also counter-intuitive. A possible explanation is that very sharp images may look unnatural. On the other hand, the size of the images in the picture viewer of our CS interface may not have been large enough for workers to notice image blur.

Capturing period was only used for the adult albums, and incorporated the aging of album characters over 11 years (from their mid 20s to mid 30s), as well as camera evolution over a decade. Based on the CS results, it seems that there is some preference for more recent images. This could be due to the better image quality that newer cameras offer, rather than having to do with character age. It should be noted that older images were captured by early low-resolution digital cameras ($<$ 1MP), images around 2006 were captured by 5MP point-and-shoot cameras and the latest ones by 16MP DSLR cameras.

Scene type: This attribute also does not seem to influence selections. No large deviations from the baseline are observed. There is a slight increase in preference for fully indoor and outdoor scenes, rather than semi indoor-outdoor.

C. Predicting Image Appeal Based on Crowd Preferences

Apart from providing detailed insights regarding user preferences, the image selection trends of Fig. 8 can be used to estimate *appeal* of new people-centric photos, irrespective of their temporal placement (higher appeal should result in higher chances of selection). This is depicted in Fig. 9, where images from 3 sub-albums of the Gallagher dataset [21] are ranked according to their predicted image appeal. Upon computing the 13 visual attributes for new images, using the marginal likelihoods of Fig. 8 on the marginalized (over time) version of Eq. (10), provides an estimation of their selection probability. These estimations can then be used to rank images and automatically select the most appealing ones from the current set. A Matlab script implementing this approach is available in Section VIII.

Some profound qualitative observations can be derived from the results of Fig. 9. Mid-to-close range shots of the album character seem to be favored over full-body shots. Colorful images are clearly preferred to dull ones, while brighter and more visible faces are favored. Most of the photos with high predicted selection probability depict intense happy faces. Conversely, photos that are dark and containing groups of people, longer distance from the camera and neutral facial expressions seem not to be favored.

D. Male/Female Comparison

A separate analysis was performed to discover differences (if any) between male and female worker preferences in the CS study. Overall, no large deviation was noted between the two groups. Fig. 10 presents the two attributes (VA and face exposure) for which large differences were noted. As evident from Fig. 10, chances of males selecting faces with negative VA can be up to 24% higher compared to females. Conversely, females favor more positive VA although not pronouncedly

TABLE III
PERFORMANCE COMPARISON FOR THE **IMAGE SELECTION** TASK (KENDALL'S τ_b AND CORRESPONDING p-VALUES IN PARENTHESIS).

Methods		Model tested on: Baby	Girl	Adult male	Adult female	Couple	Mean
Image attributes (6)	Regression [57]	0.0929 (2.38E-01)	0.1189 (1.30E-01)	0.0739 (3.50E-01)	0.2446 (1.73E-03)	0.1205 (1.23E-01)	0.1302 (1.69E-01)
	Proposed	0.142 (7.08E-02)	0.1013 (1.98E-01)	0.09956 (2.08E-01)	0.2082 (7.66E-0.3)	0.1127 (1.50E-01)	0.1327 (1.27E-01)
Image + people attributes (13)	Regression	0.6179 (3.36E-15)	**0.5331 (1.04E-11)**	0.4564 (6.96E-09)	0.4109 (1.39E-07)	0.3692 (2.22E-06)	0.4775 (4.74E-07)
	Proposed	**0.6370 (4.70E-16)**	0.5210 (3.09E-11)	**0.4571 (6.62E-09)**	**0.4733 (1.31E-09)**	**0.3822 (9.65E-07)**	**0.4940 (1.95E-07)**

TABLE IV
PERFORMANCE COMPARISON FOR THE **IMAGE SEQUENCING** TASK (MEAN KENDALL'S τ_b AND STANDARD DEVIATION IN PARENTHESIS).

Methods		Model tested on: Baby	Girl	Adult male	Adult female	Couple	Mean
Image attributes (6)	Regression	0.0866 (0.2175)	-0.0097 (0.1937)	0.0130 (0.3009)	-0.0015 (0.2422)	-0.0553 (0.2832)	0.0066 (0.2475)
	Greedy	0.0135 (0.2519)	0.0006 (0.2974)	-0.0500 (0.1942)	-0.0152 (0.2101)	-0.0341 (0.2272)	-0.0171 (0.2362)
	Proposed	0.1079 (0.2675)	0.0358 (0.2325)	0.0154 (0.2325)	0.0032 (0.2525)	-0.0217 (0.2477)	0.0282 (0.2466)
Image + people attributes (13)	Regression	0.0094 (0.2691)	-0.0265 (0.2389)	0.0125 (0.2837)	-0.0349 (0.2429)	-0.0996 (0.2656)	-0.0278 (0.2601)
	Greedy	0.0294 (0.2589)	0.0039 (0.2765)	0.0514 (0.2102)	0.0737 (0.2182)	0.0228 (0.2933)	0.0362 (0.2514)
	Proposed	**0.3602 (0.1927)**	**0.3046 (0.1549)**	**0.3230 (0.1797)**	**0.3813 (0.1890)**	**0.3345 (0.1620)**	**0.3407 (0.1757)**

with respect to males. Again it should be noted that highly negative VA images were not included in the study, and thus our inferences are limited to the observed range. Exposure of faces is another attribute for which differences were noted– females tend to tolerate underexposed/overexposed faces more. Conversely, males prefer the $[80, 150]$ luminance interval.

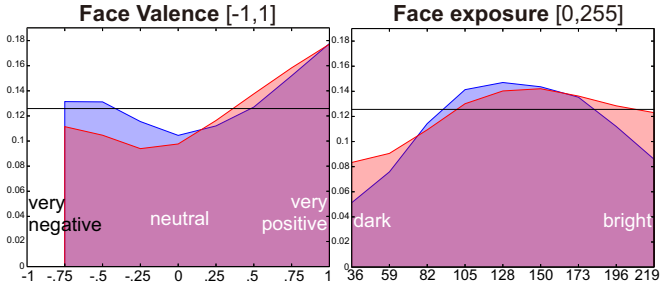

Fig. 10. Comparison of marginal likelihood (over time) area plots for male (blue) and female (red) workers. Black line denotes random baseline.

VI. EXPERIMENTAL RESULTS

We now compare the proposed method quantitatively against a regression-based approach [57] as well as a greedy baseline. In addition, we perform a qualitative evaluation via a psychophysical user study.

A. Quantitative Analysis

1) Image Selection: Our work shares similarities with [57], which predicts image *attractiveness* with a regression model using text and image-based attributes. However, [57] focuses mainly on *generic* images and not people-centric photos, which are our focus. Specifically, no people-centric attributes are considered in [57], but only low-level descriptors. We compare our proposed approach with a regression implementation similar to [57] for two scenarios: (i) using all our 13 attributes, and (ii) using only the 6 image-based attributes (*i.e.*, ignoring the people-centric ones). Among the image-based attributes used in [57], 4 are also considered in our work, while we approximate color saturation and naturalness with colorfulness and indoor/outdoorness. This comparison can reveal the advantages of (i) employing people-centric features for predicting image appeal, and (ii) employing a probabilistic framework which enables detailed analysis as compared to general trends discoverable via regression.

We adopt the evaluation protocol of [57], where the descending evaluator rank order (crowdworkers in our case) is used as ground truth (GT). The generated ranks are compared to GT via Kendalls τ_b rank correlation coefficient (adjusted for ties). High τ_b indicates greater agreement with GT. We use an album-based cross validation approach. From the 5 albums (400 images) used in CS, 4 albums are used for training while the hold-out album (80 photos) is used for testing. Once $P(sel = 1|a_m)$ or regression weights are learnt from training, appeal rankings are estimated on the test set.

Table III shows the obtained τ_b and corresponding p-values. Clearly, superior appeal estimation is achieved when people-centric attributes are used by both the proposed and regression methods. This highlights that photos of people are different in nature, requiring people-centric features for predicting their appeal. Our probabilistic approach exhibits better performance than regression for all albums except 'girl'. Interestingly, an album-based *decreasing* performance trend is observed for both methods: 'baby' $>$ 'kid' $>$ 'adult' $>$ 'couple'. This could be due to a corresponding increase in the complexity of semantic content of these albums, with 'couple' exhibiting perhaps the most complex of all. In such cases, even more high-level attributes may be required, such as gaze, body pose, or interactions.

2) Image sequencing: We employed a similar protocol as above for evaluating image sequencing. Specifically, (workers') selected images are given as input, and temporal sequencing is predicted via 3 approaches: (i) the proposed ILP

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2699859, IEEE Transactions on Multimedia

approach with $T = 5$ segments (Section III-B), (ii) a greedy algorithm for the maximization of $U_{\hat{s}}$ (see Section III-B), and (iii) a regression approach where image features are used to predict a photo's slideshow position. Crowdworkers' photo sequences are used as GT, and Kendalls τ_b is used for evaluation. Also, an album-based cross validation is employed.

Table IV depicts the mean τ_b (over all GT sequences of the tested album) and corresponding standard deviations for the three approaches. An immediate observation is that τ_b values for image sequencing are considerably lower compared to selection, with less discrepancy on the performance of different albums. This indicates that photo arrangement is more challenging to model than selection, as subjective variability may be higher for this task.

Overall, the proposed ILP approach achieves the best results by far, while the greedy algorithm performs marginally better than linear regression. Similar to image selection, the use of people-centric attributes leads to a considerable performance improvement. Interestingly, this difference is mostly evident for the ILP approach, as compared to others. Nevertheless, this observation reinforces the fact that images depicting people are inherently different compared to generic photos.

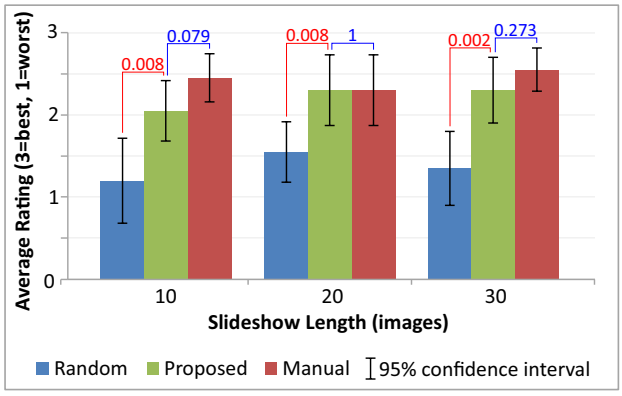

Fig. 11. **User study results:** User liking ratings with 95% CIs for 3 different slideshow lengths. t-test p-values are shown in red ($p < 0.05$) or blue ($p > 0.05$) colors.

B. Psychophysical User Study

To further examine how our approach effectively learns visual preferences, we conducted a user study. To this end, we compiled a dataset of 715 photos directly taken from a personal photo-collection (see Fig. 12). The photos were taken by either mobile phones or cameras, and were depicting the person alone or with others in various settings. These photos were *not* part of the dataset used in the CS study, and were shown to users *exactly* as they had been downloaded from the capturing devices. Consequently, image quality varied significantly, with technically flawless photos interspersed with images laden with artifacts such as motion blur or underexposure. Valid faces were detected in 524 of the 715 photos and these were the ones employed in the user study.

A photo expert who had considerable experience with slideshow creation was asked to compile slideshows comprising 10, 20 and 30 images from the 524 images. These slideshows are representative of *manual* performance. Incorporating the selection probabilities $P(sel = 1|a_m, t)$ from the CS study, and applying the proposed slideshow creation method (Section III-B), 10, 20 and 30-image *automated* (*i.e.*, proposed) slideshows were created. 15 users (13 male, 2 female) participated in the study. Furthermore, 3 *random* slideshows comprising 10, 20 and 30 photos were created for each user. All in all, each user was presented with 9 slideshows (manual/automated/random with 10/20/30 photos) in random order, and was asked to rank them as per their liking on a scale from 1 (worst) to 3 (best). Finally, participants were asked to briefly justify their rankings, and specify the attributes that influenced them.

Fig. 11 presents average scores for the manual, proposed and random versions for the three different slideshow lengths, along with the 95% confidence intervals (CIs). Student's t-tests were carried out for all pairwise combinations within the same slideshow length to examine whether the observed variations in ratings are statistically significant ($\alpha = 0.05$). Two main conclusions can be drawn:

- No statistically significant differences were noted in liking ratings for the manual and proposed slideshows for all lengths. For the 20-image slideshow, both approaches received identical scores. Thus, our slideshow creation approach performs *similar* to a human expert.
- There were significant differences between liking ratings provided for the slideshows created with the proposed approach vs. random versions for all lengths. This confirms that our method clearly outperforms random photo selection and sequencing.

Fig. 12 depicts the 20-image slideshows used in our study. 'Random' denotes a randomly compiled slideshow, which was generated anew for each user. Qualitative analysis of these slideshows, over all 3 different lengths (10, 20 and 30 photos), gave the following observations about the behavior of the proposed approach. First, close-up images are clearly preferred over images where the central character is far from the camera. Smiling photos comprise the majority of selected images. Colorful images are preferred over dull ones. Globally dark images (*e.g.*, taken at night) and photos with underexposed faces are avoided. Finally, high positive valence images are clearly preferred at the beginning of the slideshow, and the first/final image tends to include the central character alone.

VII. Conclusions

We presented an extensive crowdsourcing study to evaluate how specific visual photo attributes can impact image selection and sequencing for people-centric slideshow creation. 13 image attributes are explored, of which 7 are people-centric. A novel dataset shaping technique based on Mixed Integer Linear Programming (MILP) was proposed to arrive at compact but attribute-wise balanced datasets for the CS study, while minimizing correlations between attributes. Based on the CS results, image selection probabilities and relative image positions in a slideshow are estimated based on the different attribute values. Furthermore, a novel ILP-based slideshow creation method was introduced. Our technique is not necessarily limited to CS – image selections and

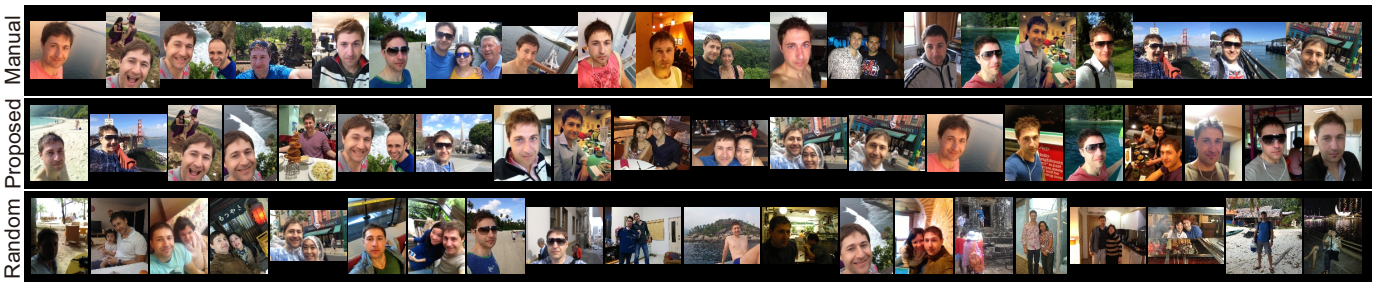

Fig. 12. **ILP-based slideshow creation:** 20-image slideshow included in the survey. Top row: Slideshow compiled *manually*; middle row: *proposed* method; bottom row: *random* selection and sequencing.

arrangements learned from a single user will result in *personalized* slideshows.

Quantitative evaluation confirms that our approach outperforms a regression-based method, and that people-centric attributes focused on faces and emotions play a critical role in estimating selection and sequencing preferences. Furthermore, a user study demonstrates that our method creates much better slideshows than random selection and sequencing, and achieves performance similar to a human expert. Consequently, future work will focus on including additional people-centric attributes such as *pose*, *activities*, *social interactions*, and *face attractiveness*. Our methodology can be coupled with near-duplicate detection [58], and combined with techniques such as [4], [59]–[61] to generate compact, aesthetic, and narrative photo streams.

VIII. Supplementary Material

Matlab implementations of the proposed methods are available at: https://sites.google.com/site/vonikakis/software-code/appealing_slideshows

Acknowledgment

This study is supported by the Human-Centered Cyber-physical Systems research grant from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[1] P. Obrador, R. de Oliveira, and N. Oliver, "Supporting personal photo storytelling for social albums," in *Proc. ACM Multimedia Conf.*, 2010.

[2] J. Yang, J. Luo, J. Yu, and T. S. Huang, "Photo stream alignment and summarization for collaborative photo collection and sharing," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1642–1651, 2012.

[3] M. L. Kherfi and D. Ziou, "Image collection organization and its application to indexing, browsing, summarization, and semantic retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 893–900, 2007.

[4] Y. Wu, X. Shen, T. Mei, X. Tian, N. Yu, and Y. Rui, "Monet: A system for reliving your memories by theme-based photo storytelling," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2206–2216, 2016.

[5] J. Yuan, J. Luo, and Y. Wu, "Mining compositional features from GPS and visual cues for event recognition in photo collections," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 705–716, 2010.

[6] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008.

[7] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 1, no. 3, pp. 269–288, 2005.

[8] D.-S. Ryu, W.-K. Chung, and H.-G. Cho, "A hierarchical photo visualization system emphasizing temporal and color-based coherences," *Multimedia Tools and Applications*, vol. 61, no. 3, pp. 523–550, 2012.

[9] G. Strong and M. Gong, "Similarity-based image organization and browsing using multi-resolution self-organizing map," *Image and Vision Computing*, vol. 29, no. 11, pp. 774–786, 2011.

[10] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Trans. Graphics*, vol. 33, no. 4, pp. 149:1–149:11, 2014.

[11] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469–1482, July 2014.

[12] A. E. Savakis, S. P. Etz, and A. C. P. Loui, "Evaluation of image appeal in consumer photography," in *Proc. SPIE*, vol. 3959, 2000, pp. 111–120.

[13] S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *Proc. Conf. Human Factors in Computing Systems*, 2014, pp. 965–974.

[14] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, p. 5, 2014.

[15] R. Subramanian, V. Yanulevskaya, and N. Sebe, "Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements," in *Proc. ACM Multimedia Conf.*, 2011, pp. 33–42.

[16] E. Birmingham, W. F. Bischof, and A. Kingstone, "Saliency does not account for fixations to eyes within social scenes," *Vision Research*, vol. 49, no. 24, pp. 2992–3000, 2009.

[17] B. Wild, M. Erb, and M. Bartels, "Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences," *Psychiatry Research*, vol. 102, no. 2, pp. 109–124, 2001.

[18] A. C. Gallagher and T. Chen, "Using group prior to identify people in consumer images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[19] J. Y. Choi, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 14–28, 2011.

[20] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: Recognizing people and social relationships," in *Proc. European Conf. Computer Vision (ECCV)*, 2010.

[21] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008.

[22] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] D. Lin, A. Kapoor, G. Hua, and S. Baker, "Joint people, event, and location recognition in personal photo collections using cross-domain context," in *Proc. European Conf. Computer Vision (ECCV)*, 2010, pp. 243–256.

[24] C. Zhu, F. Wen, and J. Sun, "A rank-order distance based clustering algorithm for face tagging," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 481–488.

[25] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag: Adding prior constraints into active learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1058–1065.

[26] V. Vonikakis, R. Subramanian, and S. Winkler, "How do users make a people-centric slideshow?" in *Proc. ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2013.

[27] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler, "Modeling image appeal based on crowd preferences for automated person-centric

collage creation," in *Proc. ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2014.
[28] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1231–1243, 2013.
[29] T. C. Walber, A. Scherp, and S. Staab, "Smart photo selection: Interpret gaze as personal interest," in *Proc. Conf. Human Factors in Computing Systems*, 2014, pp. 2065–2074.
[30] V. Vonikakis, R. Subramanian, and S. Winkler, "Shaping datasets: Optimal data selection for specific target distributions across dimensions," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, 2016.
[31] C. Li, A. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, 2010, pp. 3221–3224.
[32] I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi, "Epitome: A social game for photo album summarization," in *Proc. ACM Int'l Workshop on Connected Multimedia*, 2010.
[33] Q. Liu, A. T. Ihler, and M. Steyvers, "Scoring workers in crowdsourcing: How many control questions are enough?" in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 1914–1922.
[34] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.
[35] O. Alonso, C. Marshall, and M. Najork, "Crowdsourcing a subjective labeling task: A human-centered framework to ensure reliable results," Microsoft Research, Tech. Rep. MSR-TR-2014-91, 2014.
[36] R. Datta and J. Z. Wang, "Acquine: Aesthetic quality inference engine – real-time automatic rating of photo aesthetics," in *Proc. ACM Conf. Multimedia Information Retrieval*, 2010, pp. 421–424.
[37] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian, "Learning to photograph: A compositional perspective," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1138–1151, 2013.
[38] S. O. Gilani, R. Subramanian, H. Hua, S. Winkler, and S.-C. Yen, "Impact of image appeal on visual attention during photo triaging," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, 2013, pp. 231–235.
[39] T. O. Aydin, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Visualization and Computer Graphics*, vol. 21, no. 1, pp. 31–42, 2015.
[40] M. A. Saad, P. McKnight, J. Quartuccio, D. Nicholas, R. Jaladi, and P. Corriveau, "Online subjective testing for consumer-photo quality evaluation," *Journal of Electronic Imaging*, vol. 25, no. 4, 2016.
[41] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, 2016.
[42] S. S. Khan and D. Vogel, "Evaluating visual aesthetics in photographic portraiture," in *Proc. Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, 2012.
[43] P. Obrador and N. Moroney, "Low level features for image appeal measurement," in *Proc. SPIE*, vol. 7242, 2009.
[44] C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: A photo quality assessment and photo selection system," in *Proc. ACM Multimedia Conf.*, 2010, pp. 827–830.
[45] K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, and D. Feng, "Learning realistic facial expressions from web images," *Pattern Recognition*, vol. 46, no. 8, pp. 2144–2155, 2013.
[46] L. Zhang, D. Tjondronegoro, and V. Chandran, "Representation of facial expression categories in continuous arousalvalence space: Feature and correlation," *Image and Vision Computing*, vol. 32, no. 12, pp. 1067–1079, 2014.
[47] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the Radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.
[48] A. Atamturk and M. Savelsbergh, "Integer-programming software systems," *Annals of Operations Research*, vol. 140, no. 1, pp. 67–124, 2005.
[49] J. Clausen, "Parallel branch and bound principles and personal experiences," in *Parallel Computing in Optimization*, ser. Applied Optimization, A. Migdalas, P. Pardalos, and S. Storøy, Eds., 1997, vol. 7, pp. 239–267.
[50] P. Viola and M. J. Jones, "Robust real-time face detection," *Int'l Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[51] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *Proc. Face and Gesture Recognition Conf.*, 2015.
[52] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2011, pp. 503–510.
[53] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur," *IEEE Trans. Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.
[54] D. Hasler and S. E. Susstrunk, "Measuring colorfulness in natural images," in *Proc. SPIE*, vol. 5007, 2003, pp. 87–95.
[55] P. Obrador, "Region based image appeal metric for consumer photos," in *Proc. IEEE Workshop on Multimedia Signal Processing*, 2008, pp. 696–701.
[56] P. Jacob and P. E. Oliveira, "Relative smoothing of discrete distributions with sparse observations," *Journal of Statistical Computation and Simulation*, vol. 81, no. 1, pp. 109–121, 2011.
[57] J. San Pedro and S. Siersdorfer, "Ranking and classifying attractiveness of photos in folksonomies," in *Proc. Int'l World Wide Web Conf.*, 2009, pp. 771–780.
[58] V. Vonikakis, A. Jinda-Apiraksa, and S. Winkler, "Photocluster: A multi-clustering technique for near-duplicate detection in personal photo collections," in *Proc. Int'l Conf. Computer Vision Theory and Applications (VISAPP)*, vol. 2, 2014, pp. 153–161.
[59] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz, "Exploring photobios," in *Proc. SIGGRAPH*, 2011, pp. 61:1–61:10.
[60] Y. Liu, J. Fu, T. Mei, and C. W. Chen, "Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks." AAAI Conference on Artificial Intelligence, February 2017, pp. 1445–1452.
[61] V. Vonikakis and S. Winkler, "Emotion-based sequence of family photos," in *Proc. ACM Multimedia Conf.*, 2012, pp. 1371–1372.

Vassilios Vonikakis received the degree in Electrical and Computer Engineering as well as the PhD degree, from the Democritus University of Thrace, Greece. He currently works as a Research Scientist at the University of Illinois' Advanced Digital Sciences Center (ADSC) in Singapore. His research interests include Computer Vision, Affective Computing and Image Enhancement.

Ramanathan Subramanian received his Ph.D. degree in Electrical and Computer engineering from the National University of Singapore in 2008. He is currently an Associate Professor at the International Institute of Information Technology, Hyderabad (India). Prior to his current appointment, he was a Research Scientist at UIUC's Advanced Digital Sciences Center, Singapore. His research focuses on Human-centered and Human-assisted computing. In particular, he is interested in developing applications which employ implicit human behavioral signals such as gaze patterns and brain-based signatures for media and user analytics. He is a Senior Member of IEEE and a member of ACM.

Jonas Arnfred is a Software Development Engineer at Amazon, UK. Before that, he worked at UIUC's Advanced Digital Sciences Center (ADSC) in Singapore. He graduated with a MSc./BSc. degree (2013) from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Stefan Winkler is Distinguished Scientist and Director of the Video & Analytics Program at the University of Illinois' Advanced Digital Sciences Center (ADSC) in Singapore. Prior to that, he co-founded a start-up, worked for a Silicon Valley company, and held faculty positions at the National University of Singapore and the University of Lausanne, Switzerland. He has published over 100 papers and the book "Digital Video Quality" (Wiley). He is an Associate Editor of the IEEE Transactions on Image Processing.