Learning to Select and Order Vacation Photographs

Fereshteh Sadeghi[1,2] J. Rafael Tena[2] Ali Farhadi[1] Leonid Sigal[2]
[1]University of Washington [2]Disney Research
{fsadeghi,ali}@cs.washington.edu {rafael.tena,lsigal}@disneyresearch.com

Abstract

We propose the problem of automated photo album creation from an unordered image collection. The problem is difficult as it involves a number of complex perceptual tasks that facilitate selection and ordering of photos to create a compelling visual narrative. To help solve this problem, we collect (and will make available) a new benchmark dataset based on Flickr images. Flickr Album Dataset and provides a variety of annotations useful for the task, including manually created albums of various lengths. We analyze the problem and provide experimental evidence, through user studies, that both selection and ordering of photos within an album is important for human observers. To capture and learn rules of album composition, we propose a discriminative structured model capable of encoding simple preferences for contextual layout of the scene (e.g., spatial layout of faces, global scene context, and presence/absence of attributes) and ordering between photos (e.g., exclusion principles or correlations). The parameters of the model are learned using a structured SVM framework. Once learned, the model allows automatic composition of photo albums from unordered and untagged collections of images. We quantitatively evaluate the results obtained using our model against manually created albums and baselines on a dataset of 63 personal photo collections from 5 different topics.

1. Introduction

With abundance of devices that are able to capture images (including cell phones, electronic readers, and conventional cameras), the need for automated ways to sort and meaningfully present these images to users is becoming ever so important. We take inspiration from conventional photography, where often large photo collections are presented in the form of sparse important key moments in a particular viewing order. Viewing order may, in general, be different from the temporal order in which original photos were taken and depends on the *story* that the photographer is trying to tell. Typical key examples are traditional photo albums, wedding albums and photo slideshows. In all these instances a sub-set of photos is chosen, by a professional or the users themselves, and are arranged in a particular viewing order (e.g, see Figure 1). The process is, however, manual and often laborious for large image collections.

Figure 1. **Problem formulation:** The goal of this work is to select and order a small sub-set of photos, from a larger image collection, to form a visual story.

To this end, we look at the problem of automated album creation, where given a large set of unordered and untagged vacation photographs the method is tasked with selecting and ordering a smaller sub-set (e.g., 5 or 10) of these photos that make a compelling visual narrative. In addressing this problem, one needs to model (i) the image quality of individual photographs, (ii) the level of semantic understanding of the photo's content, and (iii) the preferences that people have for selection and ordering of photographs, based on (i) and (ii), for creation of albums.

We are motivated by recent work in image set summarization. However, our problem is sufficiently different. In particular, most summarization techniques rely heavily on the time stamps [7, 30, 10, 16, 25], geo tags [25] and even social tags [13] within Facebook-like on-line communities [16]; others rely on implicit photographer biases [24] for view selection. In [1], feature matching and epipolar geometry between pairs of images is used to provide partial time order of a subset of images. In contrast, we focus on visual content in a single image set where identities of people, exact geo location and temporal order of images is unknown. In addition, we address the narrative/storytelling aspects of the task, largely unaddressed by prior literature.

We take inspiration from advice given in photography that advocates the use of storytelling as one of the top elements in photo album design. While chronological order-

ing and geo-tag information can enhance event-based storytelling, other storytelling strategies exist and are readily used by photographers, *e.g.*, "*don't be afraid to break from [chronological ordering], by grouping photos that make sense together for impact*"[1]. Past albuming frameworks, including [7, 30, 13, 5], largely rely on temporally chronological ordering (given by time-stamps) for presentation. Our approach is more general, and can build storylines in a data-driven fashion by leveraging preferences of single, or multiple, user(s) and visual content. In our preliminary experiments, we found that people manually trying to make a short story-driven album in 40% of the cases prefer to place two consecutive photos out of chronological order. This implies that for an album of 5 photos, only about 13% of all created albums will maintain exact chronological ordering and 87% will have one or more photos out of chronological order.

Contributions: Our human study experiments show that the structure and ordering of photos in an album highly affect the human preferences. Based on these results, we formulate the problem of album creation within a framework of discriminative Structured SVM. Our model contains two sets of terms that encode learned rules for album composition. *Unary* terms, independently, model the preference for certain photographs to be used in pre-defined positions within the album. *Pairwise* terms implicitly model (1) exclusion preferences, which ensure that alike photographs are not selected more than once within an album without reason, (2) local ordering preferences that help to convey the narrative through local album consistency, and (3) long-term consistency as, for example, often exists between the first and the last frame. Finally, we propose a novel dataset, Flickr Album Dataset, that consists of nearly 9K annotated images, spanning 5 topics and 63 photo collections.

2. Previous Work

Image set summarization: In early work, *digital tapestry* [22] and *picture collages* [28], small collections of photos are summarized by a single large output image that combines salient and spatially compatible blocks from the input image set. Both [22] and [28] assume that the images are given and only address the problem of spatial layout. An extension, [21], in addition, addresses the issue of image selection by choosing a sub-set of photographs based on entropy of textures and presence/absence of faces.

Many approaches in image summarization make assumptions about criterions necessary to select the summary images. Typically those include: *image quality* – measure of image interestingness and visual attractiveness, *diversity* – a measure of non-redundancy, and *coverage* – measure of how complete the summary is with respect to the original image set [25]. Optimization of these criterions often takes the form of unsupervised clustering where clusters are formed based on metadata (e.g, GPS or text tags [25]), time stamps [5, 10, 16, 25] and/or appearance dissimilarity [10]; representative exemplars are then chosen to summarize each cluster. For example, in [10] face size and positioning are used to select a representative photo for each cluster; [4] use centrality measure for a graph with edges weighted by the near-duplicate similarity and [26] uses combination of visual and time stamp data. In [13] a social image value is proposed, based on the relationship of people in an image and intended audience, assuming social relationships between portrayed subjects are known and subjects can be identified. In contrast to other methods, that often use heuristic measures for clustering, in [26], authors propose to use human-centered measures by building a social game – Epitome.

Notable distinctions are the works of [16] and [7]. Similar to our overall goal, in [16] authors propose a system for building stories from personal image collections on social media. The approach is interesting because it attempts to encode rules of dramaturgy and cinematography for storytelling. However, the method relies on a greedy procedure that tries to match certain marginal statistics obtained from Facebook-like social media profiles (e.g, desired ratio of faces to non-faces, desired distribution over tagged actors/characters, *etc.*). In [7], time ordering across Flickr photo streams is used, instead, as a method for learning large scale storyline graphs; however, the method relies on temporally consistent diversity clustering, making it more appropriate for summarization.

All mentioned image summarization approaches have three key limitations: (1) they assume knowledge of metadata, (2) they often produce summaries that are order independent or purely temporal (based on time stamps), and (3) the rules for selection and ordering are often hand coded. In contrast, our goal is to learn a human-preference-centric album creation model. As we show in Sec. 3.1 semantic (not necessarily temporal) ordering of photos is critical to predict how an actual user would choose to select and order photos in an album. Coverage and diversity also take a different form. Finally, we assume that our images come without any metadata, or social context data [13, 16].

Scene understanding: The task of semantic scene understanding has a long history in computer vision and we omit a detailed review of literature due to space limitations. Classically, scene understanding relied on global holistic features (*e.g.*, GIST [18], filter banks, bag of words models [9]). More recently, however, richer representations that rely on spatial layout of objects [11, 12], interactions between objects and/or attributes to encode the scene have been successfully proposed. We build on these richer scene descriptions by utilizing both holistic and ObjectBank [12] features as representation. We also take advantage of pre-trained scene and attribute classifiers from [29, 19].

[1] http://digital-photography-school.com/5-top-tips-for-designing-good-photo-book-layouts

Topic	Number of Collections	Number of Images	Average Images per Collection
Disney	16	3,194	200
Beach	25	3,014	121
London	5	557	111
Paris	9	1,205	134
Washington	8	692	87
Total:	63	8,662	137

Table 1. **Flickr Album Dataset:** Statistics of the collected dataset.

Importance and memorability in images: Our method is implicitly related to recent studies that look at what people find important [2] or memorable [6] in images. Unlike [2, 6] that take a perceptual approach toward understanding these effects, we treat this information as latently present in our annotated albums and try to learn the model that implicitly encodes this information. In addition, we deal with collections of photographs, and so the question of importance must be cast jointly on an album.

Iconic images: Our work is indirectly related to iconic images, explored by Raguram *et al*. [20] and Berg and Berg [3]. Iconic images are representative images of a specified object category. In this light, the creation of an album, as we defined it, can be thought of as a simultaneous estimation of a set of categories, their ordering within a visual narrative, and an iconic representation (an image) for each.

3. Notation, Data Collection and Analysis

To give a detailed introduction, we first discuss the data and experiments that explore the problem domain of album creation. We then focus on model formulation that allows us to create albums from photo collections automatically.

We collected a dataset from Flickr that contains a total of $S = 63$ image collections on 5 topics related to vacation photographs: 1) trip to a Disney theme park (DS), 2) beach vacation (BC), and trips to 3) London (LN), 4) Paris (PR), and 5) Washington DC (WS). Collections contain between 44 and 1353 photographs, i.e, $44 \leq N_f \leq 1353$, where N_f is the number of images/photos in the collection $f \in [1, S]$. The statistics of the collected dataset are listed in Table 1. Images in different collections are of varying quality, ranging from amateur to semi-professional.

More formally, we can represent an image collection $\mathbf{I} = \{I_1, I_2, ..., I_{N_f}\}$ by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_f}\}$ where $\mathbf{x}_i$ corresponds to features computed based on image/photograph I_i. Each album can then be encoded as $\mathcal{A} = \{\mathbf{x}_{y_1}, \mathbf{x}_{y_2}, ..., \mathbf{x}_{y_M}\}$, where $y_i \in [1, N_f]$ are the selected indices for the images in the collection $\mathbf{I}$. Note that $\mathbf{y} = \{y_1, y_2, ...y_M\}$ can be interpreted as a structured output that contains ordered set of indices.

Annotation for training: Each photo collection was annotated by $A = 4$ independent annotators. The annotations came in two forms: (1) album annotations and (2) shot annotation. For *album annotation*, we asked annotators to create photo albums of length $M = 5$ photographs. Our positive annotated training data then takes the form of pairs $\{(\mathbf{X}_i, \mathbf{y}_{i,j})\}_{i \in [1,S], j \in [1,A]}$. Note that we only have annotations of the albums and do not have an explicit negative set; instead we rely on our learning procedure to mine for negative examples (bad albums). Annotators were not given any explicit instructions beyond the fact that they had to select and order M photographs from an album in a way that "*tells a story*". For *shot annotation*, one annotator was tasked with grouping images into sets of near duplicates; only images that are next to one another, temporally, can be grouped into a single shot. Recall that at test time we do not assume any knowledge about the image sets, their content and any of the annotations.

3.1. Data Analysis

The task of album creation is a high level cognitive task that involves aesthetics and storytelling. It is unclear how well people themselves are capable of performing such a task, or whether they are consistent in it. Figure 2 shows the albums created by the annotators for 3 different collections (one from 3 out of the total of 5 topics). It is clear that certain regularities exist and there is consistency in the selection and ordering of photographs. This is emerging naturally, since instructions to annotators did not specify anything about the content or ordering of images.

Perceptual verification experiments: We conducted a 4-way forced choice perceptual experiment, to ensure that observations we made about our annotations are perceptually quantifiable and to explore which aspects of the task are important. The four conditions tested were: (1) random selection of 5 images from the collection; (2) uniform selection of 5 images sorted in temporal order; (3) human annotated album of 5 images; and (4) human annotated album of 5 images, but where the order of the images is perturbed. Note, that since it's a 4-way choice, the chance of selecting any of the conditions at random is 25%. We asked participants on Amazon Mechanical Turk to carry out 3 trials (randomizing selections where appropriate) for 76 photo collections and 4 annotations each. The total of $3 \times 76 \times 4 \times 5 \approx 4500$ HITs are summarized in Table 2.

We can draw a number of conclusions, looking at Table 2. First, we find that annotated albums are substantially better (preferred in 32.8% of cases) than other conditions. Second, we can clearly see that ordering plays an important role, since performance drops significantly (from 32.8% to 28.0% on average; and in case of BC from 34.9% to 26.6%) when the order is perturbed. These results reveal that structured ordering of photos has a high impact in human preferences. Therefore, we propose a first, to our knowledge, *structured* album selection method which can simultaneously rank (select) and order the photos.

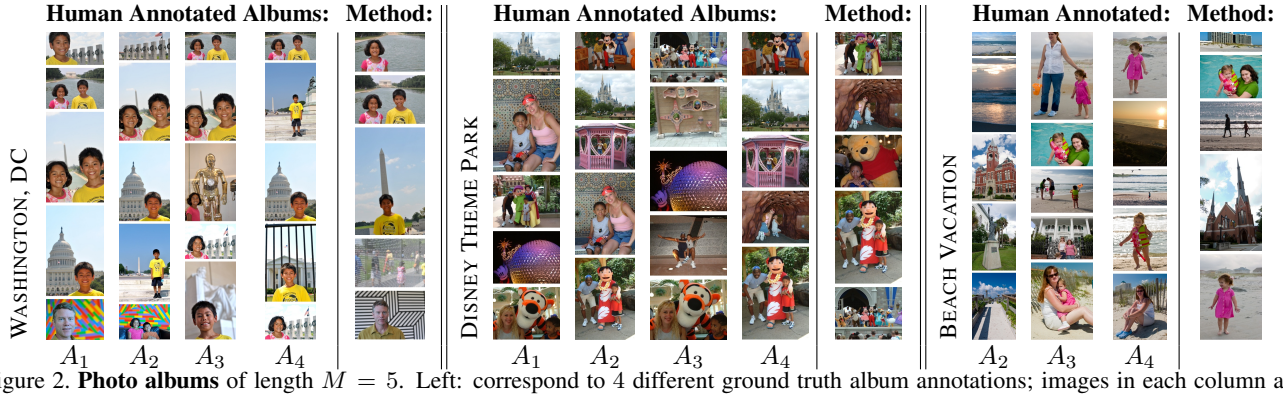

Figure 2. **Photo albums** of length $M = 5$. Left: correspond to 4 different ground truth album annotations; images in each column are shown in the selected order (top to bottom). Right: album automatically generated by our method. Note that while the automatically generated album is not identical to those on the left, it does use the same images and has a similar visual flavor.

Image Select	Image Order	Condition is chosen over all others (in %) DS	BC	LN	PR	WS	Avg
Rand	Rand	21.3	19.6	15.2	24.8	15.2	19.6
Uniform	Time	19.7	18.8	16.5	13.6	27.4	19.7
Annot.	Annot.	**30.2**	**34.9**	**36.3**	**31.6**	**32.5**	**32.8**
Annot.	Rand	28.8	26.6	32.0	30.1	24.8	28.0

Table 2. **Perceptual experiments with ground truth annotations:** Consistently better performance in 4-th row compared to the 2-nd suggest that selection of images is important; however, significant degradation of performance occurs when order of selected images is perturbed (2-nd vs. 3-rd row) suggesting that ordering is nearly as important. See text for more details.

4. Modeling Photo Albums

Given a new set of photos $\mathbf{I}$, our goal is to select a sub-set of these photos $\mathcal{A}$ (encoded by the index vector $\mathbf{y}$) which are arranged in the form of an album. We model the album using the graphical model in Fig. 3, where each unshaded node represents a frame in the album and relations between the frames are encoded by the edges. Parameters along the edges allow the model to capture semantic relationships of frame appearances and pairwise relationships such as correlations and exclusion. During the training we are interested in learning weights $\boldsymbol{w}$ for a selection function (that will also implicitly order the images), encoded by the graphical model, $f : \mathbf{X}(\mathbf{I}) \rightarrow \mathbf{y}$ that returns indices $\mathbf{y}$ of photos in the set $\mathbf{I}$ which maximize the score function $F_{\boldsymbol{w}}$ that operates on the feature representation of the frames $\mathbf{X}$. The structured output $\mathbf{y} = \{y_1, y_2, ..., y_M\}$ defines an album. The selection function is of the form:

$$\mathbf{y}^* = f(\mathbf{X}) = \arg\max_{\mathbf{y}\in\mathcal{Y}} \; F_{\boldsymbol{w}}(\mathbf{X}, \mathbf{y}) \tag{1}$$

The score function $F_{\boldsymbol{w}}$ factors into the sum of local and pairwise potentials which measure the quality of the album $\mathbf{y}$ based on the photos (and their order) in that album as a function of the weight vector $\boldsymbol{w} = [\boldsymbol{w}_i, \boldsymbol{w}_{ij}]$:

$$F_{\boldsymbol{w}}(\mathbf{X}, \mathbf{y}) = \boldsymbol{w}_i^\top \Phi_i(\mathbf{X}, \mathbf{y}) + \boldsymbol{w}_{ij}^\top \Psi_{ij}(\mathbf{X}, \mathbf{y}). \tag{2}$$

Unary potential: The unary potential $\boldsymbol{w}_i^\top \Phi(\mathbf{X}, \mathbf{y})$ measures how well a particular image $\mathbf{x}_{y_i}$ fits the corresponding position $i \in [1, M]$ in the album, independent of other images in the album. To keep the dimensionality of parameters low, and make it independent of the album size (a desirable property), we assume there is no independent preference between middle frames of the album. We do, however, independently model preferences for the first and last frames as those tend to be semantic and different between each other and all other (middle) frames in the album. This results in three different terms that contribute to our unary potential:

$$\boldsymbol{w}_i^\top \Phi(\mathbf{X}, \mathbf{y}) = \boldsymbol{w}_{\mathrm{f}}^\top \mathbf{x}_{y_1} + \boldsymbol{w}_{\mathrm{l}}^\top \mathbf{x}_{y_M} + \sum_{k=2}^{M-1} \boldsymbol{w}_{\mathrm{m}}^\top \mathbf{x}_{y_k} \tag{3}$$

where $\mathbf{x}_{y_1}$ describes the appearance of the photo which is selected as the *first* frame; $\mathbf{x}_{y_k}$, $\mathbf{x}_{y_M}$ the appearance of the photos which are selected as the *middle* and *last* frames.

Pairwise potential: The pairwise potential $\boldsymbol{w}_{ij}^\top \Psi_{ij}(\mathbf{X}, \mathbf{y})$ models different pairwise contextual relationships between album frames. In general, we are after learning the inclusion or exclusion principles – given a selection for album frame i what frame should (or should not) appear in place j (terms 2–4 in Eq. 4) and long term correlations – given a selection for the first frame what is the likely last frame (term 1 in Eq. 4) To this end we formulate pairwise potentials by combining four different terms:

$$\begin{aligned} \boldsymbol{w}_{ij}^\top \Psi_{ij}(\mathbf{X}, \mathbf{y}) \;=\; & \boldsymbol{w}_{\mathrm{fl}}^\top \psi(\mathbf{x}_{y_1}, \mathbf{x}_{y_M}) + \sum_{k=2}^{M-1} \boldsymbol{w}_{\mathrm{fm}}^\top \psi(\mathbf{x}_{y_1}, \mathbf{x}_{y_k}) \\ + & \sum_{k=2}^{M-1} \sum_{j=2}^{M-1} \boldsymbol{w}_{\mathrm{mm}}^\top \psi(\mathbf{x}_{y_k}, \mathbf{x}_{y_j}) \mathbb{1}(k \neq j) \\ + & \sum_{k=2}^{M-1} \boldsymbol{w}_{\mathrm{lm}}^\top \psi(\mathbf{x}_{y_k}, \mathbf{x}_{y_M})), \end{aligned} \tag{4}$$

where $\psi(\mathbf{x}_i, \mathbf{x}_j)$ are the pairwise feature described in Sec. 6.

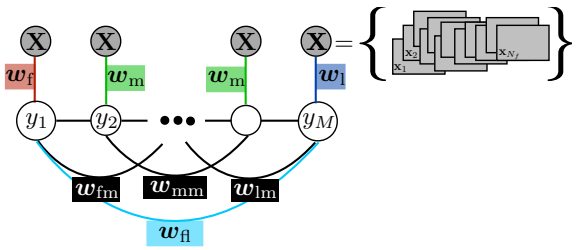

Figure 3. **Model:** Illustrated is the discriminative structured model proposed for album selection. See Section 4 for details.

5. Learning and Inference

For learning the weight vector $\boldsymbol{w}$, we can cast the problem of album selection as a structured learning task, which can be solved using structured SVM. The choice of Structured SVM was motivated by their ability to learn discriminatively and mine for hard negative samples (since our negative set is exponentially large). The key difference from the standard formulation, however, is that for each image collection, $\mathbf{I}$, instead of only having one true album, we have A different labelings for possible albums from A annotators. This aspect can be accounted for using the following, slightly modified, structural SVM formulation:

$$\min_{\omega,\xi\geq 0} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{S}\sum_{j=1}^{A}\xi_{i,j} \tag{5}$$

$$\text{s.t } F_{\boldsymbol{w}}(\mathbf{X}_i, \mathbf{y}_{i,j}) - F_{\boldsymbol{w}}(\mathbf{X}_i, \mathbf{y}_i^*) \geq \Delta(\mathbf{y}_{i,j}, \mathbf{y}_i^*) - \xi_{i,j},$$
$$\mathbf{y}^* \in \mathcal{Y} \setminus \mathbf{y}_{i,j}, \quad \forall i, j.$$

In the above formulation, S refers to the number of image collections in the training set and A refers to the number of different ground truth albums that we have for each collection. The structured output $\mathbf{y}$ contains the index of the photos which are selected in the album. Also, $\mathbf{y}_{i,j}$ refers to the jth ground truth album of the ith collection and $\mathbf{y}^*$ refers to the highest scoring album selected by the inference rule in Eq. (1); Δ is the loss function discussed in next section.

Expansion of the positive set: In practice, since the number of positive annotations is relatively small, in comparison to the number of images in the photo collection, we augment the positive set of album annotations by expansion using the *shot annotations* discussed in Section 3. For every selected image by the annotator we treat all images within the same shot as (near duplicates) as being equally likely as appearing in that place within an album. Given a hypothetical scenario where each shot is 2 frames, for an album of 5 frames we get 2^5 positive samples in the expanded positive set from each original positive album annotation.

Inference: Since our score function $F_{\boldsymbol{w}}$ includes pairwise terms and our graph is fully connected, we can not find the exact maxima. We approximate the solution to the inference problem using the TRW-S method [8].

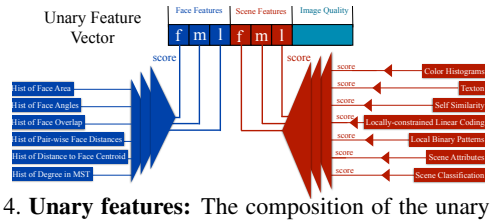

Figure 4. **Unary features:** The composition of the unary feature vector is illustrated on top; classifiers are denoted with triangles.

Loss: Unlike standard SVMs which use simple 0/1 loss functions, we incorporate a more complex loss function which enables us to penalize output albums based on how they deviate from our annotated concept of albums. In our case we can have multiple *correct* solutions as we are given multiple (A in total) annotations for each album. In addition, these annotations are not exhaustive in that typically other images exist in the photo collection that can be substituted for the ones selected by annotators without affecting the quality or story of the resulting album. This is due to the fact that there is redundancy in photos present in the collection (which is a common phenomenon); multiple pictures of identical content and nearly identical visual quality can be in the collection. We handle this to an extent, by considering images that come from the same respective *shots* as those in annotations to be equally good alternatives. To handle all these issues, we propose the following loss:

$$\Delta(\mathbf{y}_{i,j}, \mathbf{y}^*) = 1 - \max_j \Delta_{sim}(\mathbf{y}_{i,j}, \mathbf{y}^*), \tag{6}$$

where $\Delta_{sim}(\mathbf{y}_{i,j}, \mathbf{y}^*) \in [0, 1]$ measures the *semantic similarity* between $\mathbf{y}_{i,j}$ and $\mathbf{y}^*$ and max accounts for multiple annotations. We further decompose this similarity to one on the individually selected frames: $\Delta_{sim}(\mathbf{y}_{i,j}, \mathbf{y}^*) = \sum_{i=1}^{M} \omega_i \Delta_{sim}(y_i, y_i^*)$, where ω_i is the relative importance of the frame such that $\sum_i \omega_i = 1$ (in practice we let $\omega_i = \frac{1}{M}$). Intuitively, we want $\Delta_{sim}(y_i, y_i^*)$ to act like a 0/1 loss where $\Delta_{sim}(y_i, y_i^*) = 1$ if I_{y_i} and $I_{y_i^*}$ are *semantically similar* and $\Delta_{sim}(y_i, y_i^*) = 0$ otherwise. To this end we let $\Delta_{sim}(y_i, y_i^*) = 1$ if the predicted image and the ground truth annotation come from the same *shot* (*i.e.*, are near duplicates) and $\Delta_{sim}(y_i, y_i^*) = 0$ if they do not.

6. Features

To construct our unary and pairwise features that encode appearance and relationship of album frames, we employ a variety of image descriptors that fall into three categories: (1) *face features* – encode presence/absence of faces and their spatial layout in a frame, (2) *global scene features* – encode overall scene texture and color, and (3) *image quality* – encode the overall esthetic quality of the image. The illustration of our unary terms is given in Fig. 4.

Global scene features: For global scene features, we use a number of common feature representations: 1) color histograms, 2) HoG, 3) self similarity (SSIM) [23], 4) local bi-

nary patterns (LBP) [17], 5) texton histograms [15], 6) SUN scene attributes [29, 19], and 7) scene classifiers [29]. For scene attributes and classifiers, we make use of responses of pre-trained scenes and attribute classifiers from [29].

Based on each of these features we build independent 1-vs-All SVM (using LIBSVM) classifiers to classify an image as a first, middle, or last frame. We use homogeneous kernel map [27] of order 3 with χ^2 kernels for SSIM and textons and combination of intersection and χ^2 kernels for LBP and color histograms. This results in a 7-dimensional (one dimension for each feature type) response vector, $\nu \in \mathbb{R}^7$, for each classification task. This gives 21 scores which can be interpreted as confidence of a given frame being the first, the last or the middle frame based a given feature type. These scores make up the final scene feature $\mathbf{x}^{(glob)} \in \mathbb{R}^7$.

Face features: We first detect all faces in each image using the Fraunhofer[2] face detector. Based on these detections we compute spatial layout features that contain certain aspects of *proxemics* [31]. The face detector returns a bounding box containing the face and eyes. Based on this information we compute: 1) *histogram of face area* – fraction of the normalized image area occupied by a face detection; 2) *histogram of face overlap* – we compute the overlap of two faces A and B using the following two measures $\frac{A \cap B}{A \cup B}$ and $\frac{A \cap B}{A}$; 3) *histogram of pair-wise face distances* – pairwise distances between centroids of all detected faces; 4) *histogram of distances to face centroid* – distance of each face detection from the centroids of all the detections in an image, 5) *histogram of face orientations* – we approximate orientation of the face by the angle of the line connecting the two eyes to the horizontal, and 6) *histogram of the node degree in the minimum spanning tree graph*[3]. This results in a concatenated feature vector of dimension 102. Similar to above, we train a 1-vs-All SVM (using LIBSVM) with an intersection kernel to classify photos into the first, middle, and last frame, resulting in score vector $\mathbf{x}^{(face)} \in \mathbb{R}^3$. Our face features are symbolically illustrated in Figure 6.

Image quality: For image quality we pre-train a standard classifier based on generic low-level visual features as suggested in [14] using the PhotoNet[4] dataset. The PhotoNet dataset comes with ground truth quality assessment labels. Having this classifier trained on PhotoNet, we apply it to our images producing a quality score: $\mathbf{x}^{(qual)} \in \mathbb{R}$.

[2]http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst/detektion.html

[3]We use Euclidean distance, in an image, between any two face detections to construct a fully connected face graph.The weight along each edge corresponds to the distance between detections. We then compute the Minimum Spanning Tree (MST) of this graph. The MST shows the proximity of faces and we use the degree of each vertex as a measure for the arrangement of the people in the image. As the final measure, we compute the normalized degree of each node by dividing the degree of each node by the maximum possible degree of a node in a n-node graph $(n-1)$.

[4]http://ritendra.weebly.com/aesthetics-datasets.html

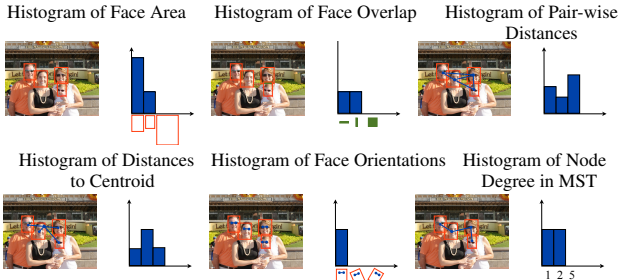

Figure 5. **Illustration of Face Features:** See text for description.

Unary and pairwise features: For unary features we simply concatenate all feature types to form: $\mathbf{x}_i = [\mathbf{x}_i^{(face)}, \mathbf{x}_i^{(glob)}, \mathbf{x}_i^{(qual)}]$. The pairwise feature, $\psi(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^7$, encodes dissimilarity of individual feature channels as a feature-wise dot product of global scene features $\mathbf{x}_i^{(glob)}$.

7. Experiments

Data: We introduce a Flickr Album Dataset, specifically for the purposes of automated photo album creation. The dataset consists of 63 photo sets collected from Flickr. Complete discussion of the dataset is given in Section 3.

Experimental setup: We train models, one for each topic, in a leave-one-out fashion. This results in 16-fold cross validation setup for Disney theme park trips, 25-fold for beach vacations; 5-, 9- and 8-fold cross validation for sets depicting vacations in London, Paris and Washington DC respectively. Note that we are able to train effectively even with few sets thanks to i) expansion of the positive set (that we discuss in Section 5), ii) low-dimensionality of our features, and iii) the large margin structured SVM formulation, that is particularly well suited for learning from little data.

Baselines: We compare performance to six different clustering baselines (obtained using K-means) as outlined in Table 3. These are consistent with literature on image set summarization. Note that while some baselines do make use of time stamp information, our method does not.

These baselines are stronger than they may appear at the first glance. For example, T-K10-T-Q over segments the image set using 10 clusters, by temporally segmenting the set into bursts of shots – this often results in clustering of images into events. The largest 5 clusters are then chosen for the summary, under assumption that more important events will contain larger number of photos. Image quality is used to pick the most representative image from each cluster, ensuring that the overall image quality is maximized. The ordering among, effectively, representative images from important events, is done based on timestamps (which is the predominant strategy for event-based storytelling). I-K10-T-Q is similar but uses visual similarity, instead of temporal burst frequency, for initial clustering, ensuring each cluster corresponds to visually coherent set of images.

Baseline Name	K	Cluster Based On	Image Selection from Cluster	Image Ordering
I-K05-S-R	5	VSim	Random	Size
I-K05-S-Q	5	VSim	Image Qual.	Size
I-K10-T-R	10	VSim	Random	Time
I-K10-T-Q	10	VSim	Image Qual.	Time
T-K10-T-R	10	Time	Random	Time
T-K10-T-Q	10	Time	Image Qual.	Time

Table 3. **Baselines:** Table outlines the 6 baselines we compare against. K is the number of clusters used by K-means. Note, when K is 10 the collection is clustered into 10 clusters, from which the 5 biggest ones are taken to form an album. Clustering is either done on visual similarity (VSim), based on global scene features, or time stamps (Time); note, our method makes no use of time stamps. Once clusters are selected one image is drawn from each cluster, either at random (Random) or based on image quality measures discussed previously (Image Qual). The final set of 5 images is ordered either by size of the clusters they came from (larger cluster first) or using time stamps in temporal order (Time).

Frame selection: We first test our ability to choose frames for particular placement in the album. We consider choosing frames for first, middle and last place in the album. We define accuracy as follows: for a predicted album we check if the selected first frame matches any frame within a shot (near duplicate set) containing the annotated album's first frame, same for the last frame and all middle frames. Results can be seen in Fig. 6 (left). On average our method outperforms the next best baseline (I-K10-T-R) by 9.9% in accuracy, which is actually 35% improvement. Our results are considerably better than the baselines in all but the London category; we believe this is due to the limited number of training collections (only 4) for London.

Feature importance: The contribution of individual features to the selection of the *first* frame in an album is given in Fig. 6 (right); we show remaining plots for *last* and *middle* placement in the supplemental material. All features seem important, but of the largest importance are the face features we propose as those allow to model "social" context of the photo. Note, chance performance is low because few images are selected to be first (or last) in an album.

Automated album creation (quantitative performance): In Fig. 6 (left), we omit the ordering among the different types of frames within albums. In Fig. 6 (middle) we show an alternate measure of performance that takes the ordering into account. This measure is similar to the loss function we use to learn our model. Given a predicted album we first compute the accuracy of this album with respect to each of the 4 annotations. We define accuracy as the fraction of the frames within a predicted album that appear within *shots* (we define a *shot* as a set of near duplicate frames) of the corresponding frame in the annotated album. For example, 100% accuracy means that all 5 frames in the predicted album were from the same shots as frames in the annotated album; 80% may mean that first frame is mismatch (or any one of the other frames), etc. Since we have 4 annotation we consider accuracy as the max across the 4 accuracies computed with respect to each annotation. This measure, unlike the one we use for frame selection experiment above, takes exact ordering of selected images into account. We further compare our model to the various baselines considered with respect to this measure. The performance can be seen in Fig. 6 (middle). Our model achieves accuracy of 11.03% on average and the closest baseline (I-K10-T-R) – 7.46% on average; our method has nearly 47% higher accuracy.

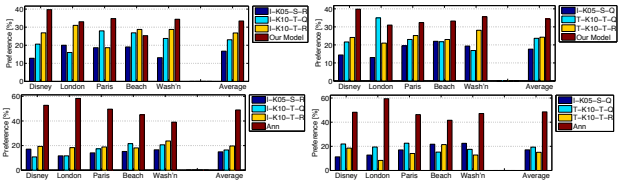

Figure 7. **Forced 4-way choice experiment:** Percentage of trials where a given method was selected over 3 competitors; see text.

Automated album creation (perceptual experiments): To objectively compare the automatically created albums to the various baselines and ground truth we run 4-way forced choice perceptual experiments on Amazon Mechanical Turk (for our method we ask 20 subjects to label their preference; for ground truth album annotations 10, since there are more trials – 4 annotations for each album). The tests are randomized, in terms of presentation order, and hence ensure unbiased preference estimates without post-filtering of MTurk results (though observed relative improvements may be lower due to uniform noise). Since we can only compare 4-choices at one time we run the experiment in 4 parts: I-K05-S-R vs. I-K10-T-Q vs. I-K10-T-R vs. Our Model (or Ground Truth Annotations – Ann) and I-K05-S-Q vs. T-K10-T-Q vs. T-K10-T-R vs. Our Model (or Ann). Our model *consistently* outperforms the competition by a large margin, except for London as discussed above, and is somewhat lower than ground truth (which can be seen by comparing Fig. 7 (top row) to (bottom row)). On average we achieve improvement of 24.7% over the next closest baseline, in terms of performance, of I-K10-T-R.

Creating longer albums: Our model is specifically designed to be agnostic to album length (due to the reuse of terms for all middle frames). This fact allows us to generate longer albums (by running inference on augmented graph with more intermediate nodes, while reusing the learned weights for all the terms) using the same model that was trained to produce 5-frame albums. The results are shown in Fig. 8 where we generate 10-frame albums. Visually the results are appealing, and show similar structure to the shorter albums, despite the fact that the model was not trained to produce an album of size 10; showing generality of method.

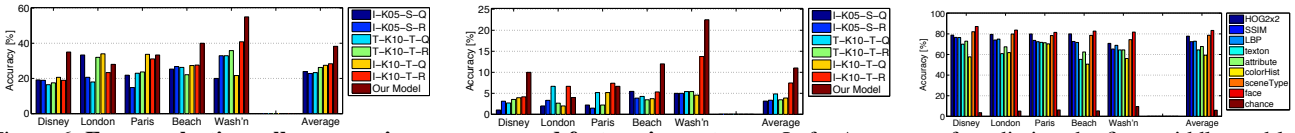

Figure 6. **Frame selection, album creation accuracy and feature importance:** Left: Accuracy of predicting the first, middle, and last frames using our method vs. the baselines. Center: Accuracy of automatic album construction using our method vs. the baselines. Right: Feature contribution towards final classifier for predicting which image goes first in the album.

Figure 8. **Longer albums:** We use the learned 5-frame model to generate albums of length 10. In general, we can generate albums of arbitrary length without re-training the model.

8. Discussion

We explore automatic creation of photo albums, that go beyond chronological ordering, to predict how an actual user may choose to select and order photos to tell a compelling visual story of a vacation trip. In doing so, we rely purely on visual information for features and exemplar album annotations to drive the discriminative learning procedure. To the best of our knowledge, our framework is the first attempt at this complex visual task. While here we learn models from multiple annotations, in essence trying to capture population average of what a *good* album is, we can easily learn albums from exemplars provided by individuals (without any alterations to our framework). This way a user, by creating a few example albums, can teach the system how to form albums of similar semantic structure.

References

[1] T. Basha, Y. Moses, and S. Avidan. Photo sequencing. In *ECCV*. 2012.

[2] A. C. Berg, T. L. Berg, H. D. III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, 2012.

[3] T. L. Berg and A. C. Berg. Finding iconic images. In *2nd Internet Vision Workshop at CVPR*, 2009.

[4] W.-T. Chu and C.-H. Lin. Automatic summarization of travel photos using near-duplication detection and feature filtering. In *ACM MM*, 2009.

[5] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM MM*, 1(3):269–288, 2005.

[6] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, 2011.

[7] G. Kim and E. P. Xing. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *CVPR*, 2013.

[8] V. Kolmogorov and M. J. Wainwright. On the optimality of tree-reweighted max-product message-passing. In *UAI*, 2005.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[10] J. Li, J. H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Pacific Rim Conf. on Multimedia*, pages 1536–1540, 2003.

[11] L. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *ICCV*, 2007.

[12] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[13] A. C. Loui, M. D. Wood, A. Scalise, and J. Birkelund. Multidimensional image value assessment and rating for automated albuming and retrieval. In *ICIP*, 2008.

[14] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[16] P. Obrador, R. de Oliveira, and N. Oliver. Supporting personal photo storytelling for social albums. In *ACM MM*, pages 561–570, 2010.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[19] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[20] R. Raguram and S. Lazebnik. Computing iconic summaries for general visual concepts. In *1st Internet Vision Workshop at CVPR*, 2008.

[21] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake. Autocollage. *ACM Trans. Graph.*, 25(3):847–852, 2006.

[22] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. In *CVPR*, 2005.

[23] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[24] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *ICCV*, 2007.

[25] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *ACM ICMR*, 2011.

[26] P. Vajda, I. Ivanov, J.-S. Lee, and T. Ebrahimi. Epitomize your photos. *Int. J. Comput. Games Technol.*, 2011.

[27] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *TPAMI*, 34(3):480–492, 2012.

[28] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum. Picture collage. In *CVPR*, 2006.

[29] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[30] J. Yang, J. Luo, J. Yu, and T. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Trans. MM*, 2012.

[31] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012.