Synthesizing Illumination Mosaics from Internet Photo-Collections

Dinghuang Ji, Enrique Dunn, Jan-Michael Frahm
Department of Computer Science, The University of North Carolina at Chapel Hill
{jdh,dunn,jmf}@cs.unc.edu

Abstract

We propose a framework for the automatic creation of time-lapse mosaics of a given scene. We achieve this by leveraging the illumination variations captured in Internet photo-collections. In order to depict and characterize the illumination spectrum of a scene, our method relies on building discrete representations of the image appearance space through connectivity graphs defined over a pairwise image distance function. The smooth appearance transitions are found as the shortest path in the similarity graph among images, and robust image alignment is achieved by leveraging scene semantics, multi-view geometry, and image warping techniques. The attained results present an insightful and compact visualization of the scene illuminations captured in crowd-sourced imagery.

1. Introduction

Internet photo-collections can provide a vast sample of the space of possible viewpoint and appearance configurations available for a given scene. This work addresses the organization and characterization of this image space by exploring the link between time-lapse photography and crowd-sourced imagery. Time-lapse photography strives to depict the evolution of a given scene as observed under varying image capture conditions. While the aggregation of a sequence of images into a video may be the most straightforward visualization for time-lapse photography, the integration of multiple images in the form of a mosaic provides a descriptive 2D representation of the observed scene's temporal variability. We denote these time-lapse mosaics as *illumination mosaics* and show an example in Fig. 1.

The problem of mosaic construction can be abstracted as a three-stage process of image registration, alignment, and aggregation. However, the representation of the appearance dynamics introduces the qualitative challenge of producing an aggregate mosaic that is both coherent with the original scene content and descriptive of the fine-scale appearance variations across time. The associated technical challenges addressed in this work are 1) identify within an unorganized image set an image sequence depicting the desired content appearance transition and 2) construct an illumination mosaic that accurately depicts the observed appearance variability while mitigating scene artifacts due to changes in scene content and capture parameters.

Figure 1. Example time-lapse image of the Coliseum, the top image is automatically generated by our method, and the bottom is manually made by a photographer (courtesy of Richard Silver).

We address these challenges by exploring the spectrum of capture variability available in Internet photo-collections and propose a novel framework to obtain illumination mosaics. We briefly summarize the functionality of our processing pipeline. The input data to our framework are a reference image depicting the desired image composition to be used to generate the illumination mosaic and a crowd-sourced image collection of the scene of interest. We initially use semantically-aware global image features characterizing an imaged scene's composition and ambient illumination properties in order to determine the scope of the variation to be represented in the mosaic. Then, a limited

connected graph is built based on image similarities, from which we find a smooth path between two nodes, defining an ordered set of images to be used for mosaicing. Our subsequent image alignment and stitching leverages 2D warping, segmentation, and color mappings to achieve smooth image transition while mitigating scene aberrations. We demonstrate our method on several landmark datasets, and show both qualitative and quantitative results.

2. Related Work

A possible way to automate appearance-based mosaic generation is to transfer the color of images taken at different times of the day into a single image. Along these lines, [25] and [24] propose to match color statistics between images, which could be used in style transfer. Akers *et al*. [4] introduce a method to create illustrations from a set of images of an object taken from the same point of view under variable lighting conditions. Chia *et al*. [9] were the first to leverage the rich image content on the Internet to color a grayscale image. However, this method can not be applied to time-lapse images which contain dramatic appearance change. Shih *et al*. [30] propose an automatic "time hallucination" method to synthesize a plausible image at a different time of day. Laffont *et al*. [17] further define 40 transient attributes to characterize a scene's appearance change, and transfer these attributes to new images. While these color transfer methods could generate illumination mosaics, they rely on large datasets of time-lapse videos and we empirically found them to look artificial.

There exists a large body of research on modeling the temporal order of images. Seitz *et al*. [21] introduce an approach for synthesizing time-lapse videos of landmarks from online photo collections, which aims to visualize long-term temporal change of dynamic elements in the scenes. While our method aims to visualize the appearance change of scenes from night to day. Wang *et al*. [34] propose low-dimensional manifolds to model the gradual appearance change of materials. In order to find smooth transitions between images of faces, Shlizerman *et al*. [15] build a graph with faces as nodes and similarities as edges, and solve for shortest paths on this graph. For natural scenes like the appearance of the sky, Tao *et al*. [33] analyze semantic attributes of sky images, train classifiers to categorize them, and find smooth sequences of appearance change. To find intermediate images in the sequence, they build an image graph and connect images with nearest neighbors (in terms of color distance). Instead of the sky, we focus on generating the temporal change of more general scenes and adopt local color transfer techniques to better portray the color transition. Schindler *et al*. [26] propose a constraint-satisfaction method for determining the temporal ordering of images based on visibility reasoning of reconstructed 3D points. They further present a framework [27] for estimating temporal variables in structure from motion problems and obtaining the temporal order of images. Their methods work for images taken over decades of time. Palermo *et al*. [23] extract features that are temporally discriminative and show outstanding results in temporal classification of historical images. Kim *et al*. [16] propose a non-parametric approach for modeling and analysis of the topical evolution for Internet images with time stamps. Jacobs *et al*. [14] created a large dataset of over 500 static web-cameras around the world and propose a method to analyze consistent temporal variations in these scenes. Our proposed method mines unorganized crowd-sourced data to identify a suitable visual datum to construct illumination mosaics.

Given images taken in short time periods, Basha *et al*. [6] recover the temporal order by extracting features from dynamic elements in a scene, and comparing their relative positions with static feature points. They further relax the strong assumption that two images must be captured by the same static camera by utilizing the temporal information from successive images captured by the same moving camera [7]. Several methods address the problem of non-rigid shape and motion recovery from a set of still images when temporal order is not used directly. Avidan and Shashua [5] recover the temporal order of a 3D moving point by assuming a fixed shape of the trajectory. In this paper, we aim to find image sequences from night to day by modeling image relationships with color features. Our method does not rely on any motion cues or priors, but instead builds temporal sequences exclusively from appearance transitions.

There has been tremendous progress in modeling unordered Internet image collections [10, 1, 12, 28]. The work of Snavely *et al*. [31, 32] enabled the spatially smooth traversal from Internet images of landmark scenes. Lee *et al*. [18] propose a system to "rephotograph" historical photographs. Xu *et al*. [35] use collections of images to infer the motion cycle of animals. Hays *et al*. [11] propose an image completion algorithm which fills in empty areas by finding similar image regions in a large dataset. With a different goal, we aim to visualize the temporal change of scenes by leveraging appearance transfer techniques.

To create an illumination mosaic we compose the information from multiple images into a single photo, which has been discussed in [37, 29, 2]. Besides these previous work, Agarwala *et al*. [3] adopt graphcut and gradient domain fusion to choose good seams between images and reduce visible artifacts in a composite image. To stitch a set of images, Levin *et al*. [19] introduce several formal cost functions for the evaluation of the quality of stitching. Zhang *et al*. [38] propose a hybrid alignment model that combines homography and content-preserving warping to provide flexibility for handling parallax. However, this method is not designed to align image sequences and did not show results to align images with very different illuminations.

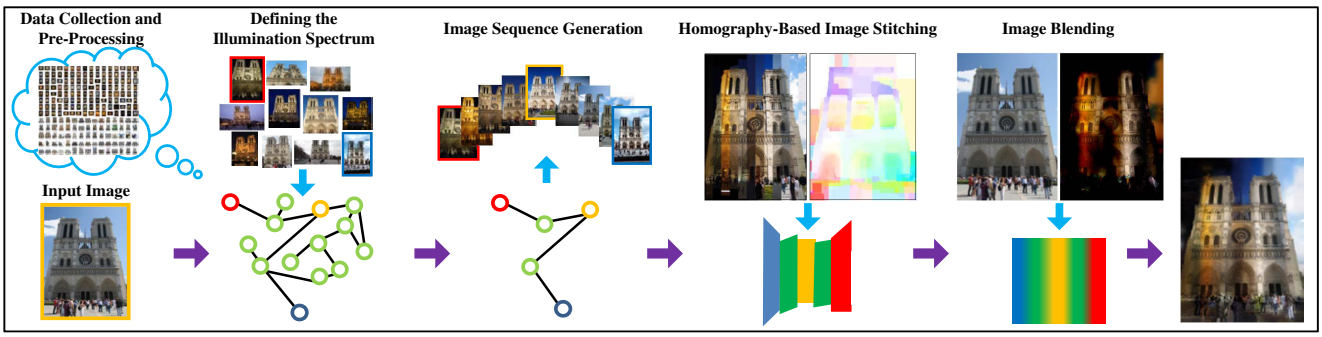

Figure 2. Framework of our method. Given an input image I, our method determines an appearance neighborhood $\mathcal{N}_{GIST}(I)$ within a photo collection. We identify two extremum elements of $I^- \in \mathcal{N}_{GIST}(I)$ and $I^+\mathcal{N}_{GIST}(I)$ to determine a path within an appearance similarity graph, which corresponds to image sequence used for mosaic integration. We perform robust homography-based region warping to aggregate a mosaic. Finally, we transfer color from the mosaic into our reference image.

3. Illumination Mosaic Generation

In order to depict the illumination spectrum of a scene, our method relies on building discrete representations of the image appearance space through connectivity graphs defined over a pairwise image distance function. To generate illumination mosaics, we want to select an image sequence which 1) shares similar spatial composition, 2) features a smooth color transition between the images, and 3) conveys a large variety of scene appearances. We now detail our proposed framework for identifying the appearance variability in a photo collection, and subsequently using it to build illumination mosaics. Fig. 2 shows an overview of our pipeline.

3.1. Data Collection and Pre-Processing

To obtain the image data for different landmarks, we first perform a keyword-based query to the Flickr photo sharing website. In order to remove unrelated images, we employ the iconic selection pipeline proposed in [10]. We perform GIST-based([22]) image clustering and discard images that cannot establish a pairwise epipolar geometry to the cluster center. We perform K-means clustering enforcing an approximate average cluster size of 50 images. Given that all non-discarded images can be registered to the cluster center, it is possible to estimate a local 3D model of the scene. However, for efficiency purposes, we do not perform full dataset geometric verification, but instead rely on pairwise image registration to determine 2D image alignments.

3.2. Defining the Illumination Spectrum

The composition of our illumination mosaics requires us to specify both the desired spatial composition of the image output and the range of appearance variability to be depicted. We take as input (from the user) a reference image I that will define the spatial layout/composition of our output illumination mosaic and will be used to define subsequent image alignment and warping operations. Next, we identify, within our registered image set, elements that define the scope of our displayed appearance variation. We select a local appearance neighborhood to the reference image, which is comprised of the nearest K=300 images in terms of the Euclidean distance of their corresponding GIST descriptors. That is, we compute the GIST descriptor for the input reference, and by leveraging the pre-computed GIST descriptors for our registered dataset, we determine an image set $\mathcal{N}_{GIST}(I)$ of its K nearest neighbors. The motivation for initially focusing on a reduced local neighborhood is to ensure spatial content similarity among images, which will facilitate subsequent image alignment and warping.

In order to exploit the diversity of image capture characteristics found in a crowd-sourced photo collection we need to identify image measurements that are discriminative w.r.t. the variations we want to portray in our mosaics. We focus on a specific type of global appearance variations: the transitions between dark and bright ambiance. To enable this characterization we leverage image statistics of disjoint semantic elements within a scene to define an aggregate scene descriptor. More specifically, we perform foreground and sky segmentation on the input image and compute histogram statistics for each of the disjoint image segments.

Sky Segmentation. Empirically we found that using the sky detector proposed in [13] to extract the sky region provides unreliable results for images captured at night. For each image we estimate an homography-based warp to its nearest GIST-neighbor. We then compute local NCC for the two images, where local patches with NCC values larger than 0.5 will be deemed to belong to foreground buildings, and patches with NCC values less than 0.2 are labeled as background. The intuition is that static structure will have consistent NCC even in different illuminations while sky regions and transient objects will not. Graphcut is adopted to generate a more complete segmentation for the building and sky (shown in Fig. 3).

Quantifying Image Intensity. For the pixels contained in the sky segment we compute a 100-bin intensity histogram $\mathcal{H}_b$ of the blue color channel. We compute the in-

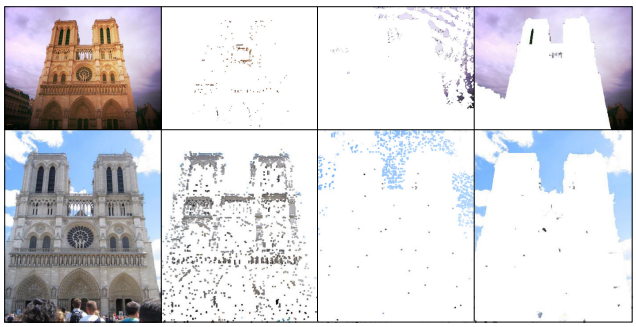

Figure 3. Sky/building segmentation. (a) Original images, (b) Foreground mask, (c) Background mask, (d) Sky segmentation.

tensity values (i.e. histograms bins) corresponding to the top 5 frequencies and select their median as our intensity measure for that image, given that image histogram will usually have multiple peaks. We choose images I^+ and I^- having the highest and lowest intensity values within $\mathcal{N}_{GIST}(I)$ as the two respective extremes of our illumination spectrum.

3.3. Image Sequence Generation

The goal of this step is to find an image sequence that depicts the gradual variation between the previously selected pair of images, I^- and I^+, which define the scope of our output illumination spectrum. We build this path by determining and concatenating an image sequence $I^- \rightarrow I$ and an image sequence $I \rightarrow I^+$, where all the aforementioned images are elements of our registered camera set. Henceforth, we will consider the $I^- \rightarrow I$ transition, but it is to be understood that the same steps apply to the second half of the image transition sequence.

Aggregated Image Appearance Descriptor. We combine a global image GIST feature descriptor to capture the image composition, a color histogram to represent the sky color, and a histogram of the dark channel prior image to choose photos that contain well-illuminated images. We restrict our color histogram to sky regions to account for landmarks which may be arbitrarily illuminated at night. We use all three color channels to enable more fine-grained discrimination of ambient illumination among subsequent images. These three features are normalized and concatenated to form a global image feature representation.

Image Similarity Graph. Based on our global image descriptor we define a discrete representation of our appearance space based on image pairwise similarity. We incrementally build a graph where each image is treated as a node, similar to [34], we use both k-rule and ϵ-rule to construct a neighborhood graph. The edge weights connecting two nodes are computed by L^2 distance of image features. To find a balance between path descriptiveness and compactness, we iteratively augment the local image neighborhoods around both I^- and I^+ until we attain a single connected component from which to attain a minimum-length path between the nodes corresponding to I^- and I^+. Moreover, at each iteration k(which starts from 1), each image in the registered camera set is only connected to its k nearest neighbors. Outliers in the graph are reduced using the ϵ-rule, which removes edge connections that have weights (i.e. descriptor distance) more than $\epsilon = 1.3d_p$, where d_p is the average edge distance in the graph. Once a k-connected graph is defined at each iteration, we search for a connecting path between I^-, I and I, I^+ by using Dijkstra's method.

3.4. Homography-Based Image Stitching

Our scene warping is a two-stage process that leverages pairwise homography transfers between elements of our image sequence. First, we compute a homography warping $\mathbf{H}_j$ between every image I_j in the generated sequence and the input image I, which transfers the local surface appearance characteristics under a local planarity approximation, i.e. $I_j^{'} = \mathbf{H}_j(I_j)$. Second, we apply dense SIFT Flow [20] warping to the homography-warped image to compensate for fine-scale scene parallax not modeled by the local planarity assumption, i.e. $I_j^{''} = \mathbf{S}(\mathbf{H}_j(I_j))$.

Robust Homography Chains. If the homography matrix H_{ba} aligns I_b to I_a, according to the chain rule, the homography matrix that aligns a third image I_c through I_b to I_a is $H_{ca} = H_{ba} \cdot H_{cb}$. Likewise, if we have N images and want to register the n_{th} image to the first one, the homography matrix could be written as $H_{1,N} = \Pi_{i=1}^{N} H_{i,i+1}$.

However, in our experiments we found computing feature-based homographies directly between neighboring images is unreliable, especially for images captured at night. Since we only extract color features from the sky, the colors on the building facades between neighboring images can be very different (i.e. in Fig. 4). While simplifying image alignment to a homography model provides a more inclusive geometry fitting framework (i.e. less constraints) we observed that reliably building an homography chain across the entire input sequence was still elusive. As mitigation we explored the use of bridge images to attain pairwise homography estimates through transitivity Fig. 5(c).

We measure the confidence for our homography estimation based on the output of the pairwise RANSAC estimation process. We measure the number of inlier matches $m_{i,j}$ between images I_i and I_j and the image area $a_{i,j}$ of the convex hull of the attained inlier set normalized by total image size. Note that $m_{i,j}$ is symmetric while $a_{i,j}$ is not. Using these values we define a pairwise homography confidence score between images I_i and I_j as

$$C_{i,j} = m_{i,j} \cdot (a_{i,j} + a_{j,i}) \tag{1}$$

and use it to search for an alternative intermediate *bridge* image between every adjacent image pair in the sequence. The motivation is to omit unreliable adjacent estimates

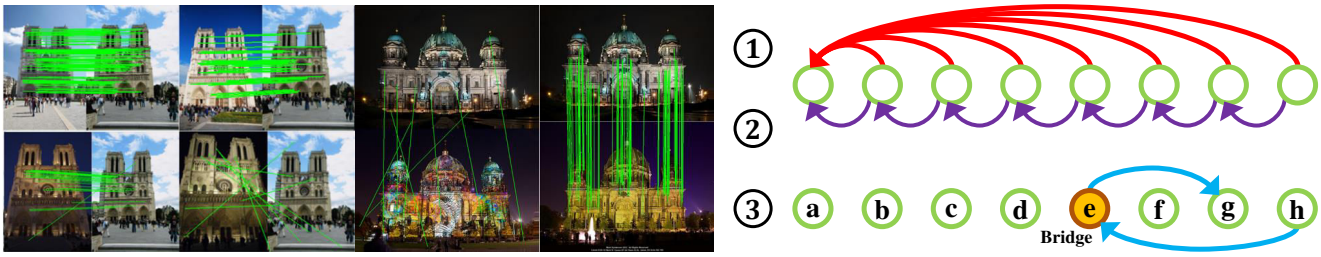

Figure 4. Motivation for robust homography chains. (left) The reliability of direct pairwise homography estimation of an entire image sequence to a single reference image is not uniform across the sequence. Moreover, neighboring images may exhibit drastic appearance variation (especially at night), hindering direct homography chains. Green lines depict RANSAC inlier matches. (right) Schematic representation of (1) direct pairwise estimation, (2) direct homography chains , and (3) our proposed bridge-based homography estimation.

through the transitivity of a third image. Given an image I_i, the bridge image I_k is selected as the non-adjacent image with highest confidence path to the adjacent image. The bridge I_k image will be used to join two successive images I_i and I_{i+1} whenever the following condition is satisfied

$$C_{i,i+1} < \max_{k \neq i}(r_{i,k} \cdot C_{i,k} + r_{i+1,k} \cdot C_{k,i+1})/2 \quad (2)$$

in which $r_{i,k}$ is the area ratio of image i and k, and this is used to regularize cases when image k has higher resolution than image i. Similarly, we use a confidence threshold to eliminate images in the sequence that do not attain reliable homography estimations, and reconnect the sequence through the same bridge image search process as before.

Stitching & Refinement. Upon establishing a robust local homography chain across the entire sequence $\{I_j\}$, we warp all the images into the reference image I. Next, we apply dense SIFT Flow warping [20] to the homography-warped images to compensate for fine-scale scene parallax not modeled by the local planarity assumption. Finally, we form a mosaic by sequentially aggregating equal-sized vertical stripes from each of the images in the sequence to form a single, combined image. It is constructed such that the first (leftmost) vertical stripe is obtained from the first image in the sequence, the second stripe from the second image, and so forth. In this manner, the mosaic depicts a single, recognizable view of a scene, but is composed of stripes taken from different images (see Fig. 5(b)). The length of the output sequence is data-dependent as it is a function of both the size and composition of the image set used to determine our illumination spectrum. However, replacing Dijkstra shortest path search in our implementation with Yen's k-shortest path algorithm [36] would enable the user to set sequence length a priori.

3.5. Image Blending

We note that the generated stitched mosaic M may have strong color and structural artifacts among adjacent mosaic segments, see Fig. 5(b). The reason for these artifacts include: 1) Inconsistent foreground objects, i.e. pedestrians, cars, or other transient objects. These transient objects cannot find correspondences in other images and will cause registration artifacts. 2) Uneven resolutions for different stripes. Our generated image sequence does not enforce a common resolution for all images. When warping low-resolution images to high-resolution images, up-sampling will introduce blur artifacts. 3) Artifacts caused by dense registration. Although SIFT Flow generally works well for aligning static structures, sometimes it fails in texture-less regions (such as windows and tower top). Also, if the appearance or structure of the foreground elements changes dramatically, dense registration may introduce artifacts.

Color Transfer. In order to keep the fine-grained details of the mosaic, while at the same time conveying a large range of scene appearance, we decide to transfer the color from the image mosaic M to the reference image I. Shih *et al*. [30] propose a locally linear model learned from time-lapse video, allowing them to synthesize new color data while retaining image details. Moreover, for the image pair (M,I) we want to estimate local transformations which characterize the color variations between two images. The locally linear model proposed by [30] is used to relate the color of pixels in M to the color of pixels in I. We denote the patch centered on pixel p_k in the match image by $\mathbf{v}_k(M)$, and $\mathbf{v}_k(I)$ is the corresponding patch in the target image. Both are represented as $3\times N$ matrices in RGB color space; using patches of $N = 5\times 5$ pixels. The local linear transform applied to patch k is represented by a 3×3 matrix $\mathbf{A}_k$, and is estimated with a least-squares minimization:

$$\arg\min_{\mathbf{A}_k} \|\mathbf{v}_k(I) - \mathbf{A}_k\mathbf{v}_k(M)\|_F^2 + \gamma\|\mathbf{A}_k - \mathbf{G}\|_F^2 \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The second term regularizes A_k with a global linear matrix G estimated on the entire image (using a small weight $\gamma = 0.008$ in all tests). We obtain the optimal transform A_k in closed form:

$$\mathbf{A}_k = \left(\mathbf{v}_k(I)\mathbf{v}_k(M)^T + \gamma\mathbf{G}\right)\left(\mathbf{v}_k(M)\mathbf{v}_k(M)^T + \gamma\mathbf{I}_3\right)^{-1} \quad (4)$$

Since the mosaic and reference image are already aligned, there is no need to compute a correspondence map

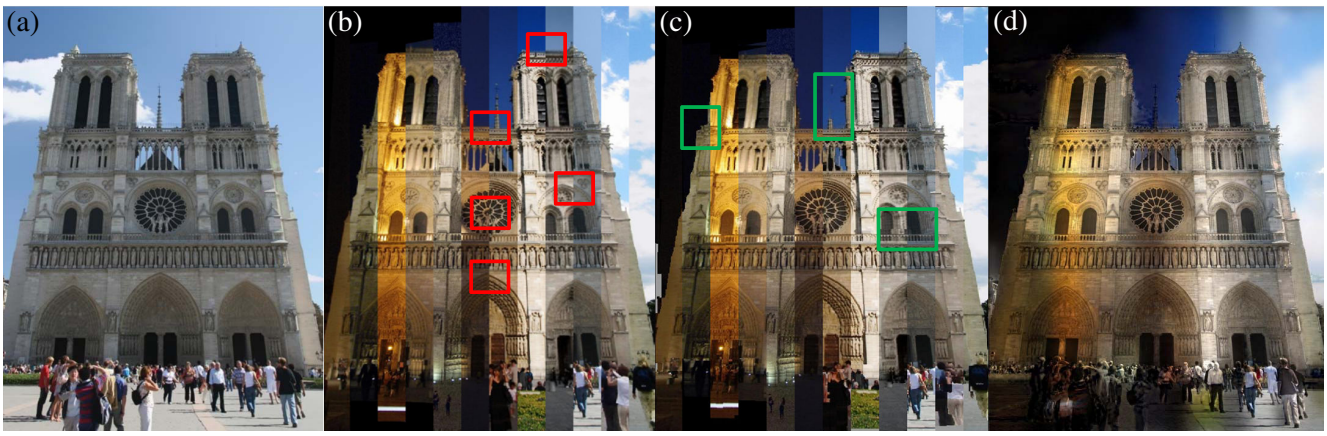

Figure 5. Mitigation of mosaicing artifacts. (a) Input reference image (b) Homography-based image stitching (red rectangles highlight alignment problems). (c) SIFT-flow dense registration refinement partially resolves alignment issues, at the expense of small-scale structure aberrations (highlighted green boxes) (d) Output image after transfering color from the mosaic to the reference image.

between them. We adopt the linear equation system proposed in [17] to solve the color transfer problem. Fig. 5(d) shows the color transfer results, compared to Fig. 5(b), and the artifacts highlighted in green are gone, and there is no detail loss from the reference image.

Local Stripe Reordering (optional). The image sequence is generated through global image appearance descriptors. However, there can be local appearance variations in the images, resulting in color inconsistencies among adjacent elements within the mosaic. Examples include clouds, partial foreground occlusions, or reduced overlap with the reference image. Addressing this contingency within the image sequencing step of our mosaic generation would entail an explosive growth of our image similarity graph, as each stripe needs to be connected to every other stripe in all other images within the appearance neighborhood. Accordingly, our approach is to resolve this issue through a post-processing step. We propose a method to locally reorder the stripes in the final mosaic to make the sky transitions look more natural by only reordering the contents of the sky regions. To this end, we leverage our existing sky segmentation and extract a sky-only intensity color histogram for each stripe. We sort the stripes by the median of their top 5 frequencies in the intensity histogram. We then transfer color from each image in the new sequence into the sky regions of the output mosaic. We repeat the process until the sequence convergences.

4. Experiments

Data Acquisition. We downloaded 10 online datasets from Flickr, and the statistics of our system's data associations are presented in Table 1. We categorize images with average intensity of their sky regions below 100 as night images (intensity value range from 0 to 255).

Homography Chain Evaluation. To evaluate the effectiveness of our bridge-based image stitching method, we design a metric to quantitatively compare alternative stitching methods. We first compute an edge map for the reference image and all warped images used to form the output

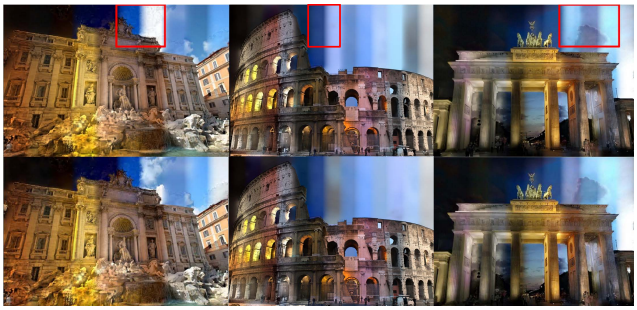

Figure 6. Sky reordering. Top: mosaics before reordering, red rectangles highlight the inconsistent stripes. Bottom: reordered mosaics, the sky appearance inconsistencies are mitigated.

Name	# Downloaded	# Clustered # (night / day)	Stripe Reordering
Notre Dame Cathedral	60291	3615 / 5260	No
Berliner Dom	51892	2197 / 3986	Yes
Brandenburg Gate	63796	2671 / 5198	Yes
Mount Rushmore	53612	583 / 2423	Yes
Coliseum, Rome	49220	910 / 1027	Yes
Trevi Fountain	94370	1612 / 3689	Yes
Manarola	54535	1023 / 4058	Yes
Potala Palace	25039	450 / 1996	Yes
Tiananmen Square	70384	658 / 3142	Yes
St. Peter's Cathedral	91060	2557 / 3297	Yes

Table 1. Composition of our downloaded image datasets. The number of clustered images corresponds to images that were able to register through geometric verification to their cluster center. In most cases (˜90%), stripe reordering is applied to generate smoother appearance transition (For Notre Dame dataset, stripe reordering didn't change its original sequence).

Dataset	Align to next	Bridge	SIFT Flow	Align to next + SIFT Flow	Bridge + SIFT Flow
Notre Dame	0.4179	0.4152	0.3509	0.4387	**0.6152**
Berliner Dom	0.3634	0.4539	0.3398	0.3812	**0.5529**
Trevi Fountain I	0.3967	0.4159	0.4123	0.6141	**0.6503**
Trevi Fountain II	0.4420	0.4292	0.3889	**0.6020**	0.5752
Forbidden City	0.3595	0.3969	0.3554	**0.4513**	0.4431
Mount Rushmore	0.4223	0.4563	0.2973	0.5257	**0.5708**
Brandenburg Gate	0.4095	0.5352	0.4130	0.4791	**0.5875**
Manarola	0.3415	0.4105	0.3306	0.4776	**0.5429**
Potala Palace	0.4251	0.5254	0.3875	0.5025	**0.5683**
Coliseum, Rome I	0.4085	0.4253	0.3169	0.6219	**0.6873**
Coliseum, Rome II	0.3468	0.4416	0.4152	0.5758	**0.7048**

Table 2. For each dataset, we create three sequences with different reference images and compute our predefined values. For Trevi Fountain I&II and Coliseum, Rome I&II, they differ in the viewing angle. Bold-font numbers highlight the best matching score, eight out of the ten datasets achieve the best results using our method. For the other two datasets, we are very close to the best scores.

mosaic, using Canny edge detection [8]. Using these edge-maps, we then compute the average per-pixel NCC values between each stripe in the reference image and its corresponding warped region in the mosaic using a 5×5 aggregation window. To focus on the inter-stripe alignment accuracy, we restrict our evaluation to edge pixels found in the boundaries between mosaic stripe elements. We compare our image stitching method (Bridge + SIFT Flow) with three methods: (1) Align image to neighbor, (2) Align with bridge, and (3) Align with SIFT Flow. From Table 2, we can see that most datasets benefit from bridge-based image stitching compared with the "Align to next" strategy. Moreover, many of the"Align to next" outputs suffer from incorrect homography estimates (due to highly different illumination conditions) which render severely distorted mosaics. Note that using robust homography chains in conjunction with dense SIFT Flow refinement provides enhanced accuracy when compared to either of them in isolation.

Color Transfer Results. We compare with three methods to create illumination mosaics: two previous works [25](a), [30](b), and our method without bridge homography connections(c). Method (a) adopts the same image sequence used in our method as input, and transfers color from all images to the reference image in the sequence using the approach proposed in [25]. Method (b) implements the method in [30] using the same reference image and the video datasets created by the original paper as input. We randomly select frames from all videos, extract their GIST and color features, compute the nearest neighbors w.r.t. the input image, and use that video as the input time-lapse source. We then manually select a temporal sequence from the video and transfer the color with the pipeline proposed in [30]. Method (c) also uses the same image sequence as input. We warp the sequence using only SIFT Flow, and transfer the color using the locally affine method proposed in [30]. The comparative results in Fig. 7 illustrate both the wide range of appearance variation achieved by our approach as well as the recovered fine-scale chromatic and scene structure details. Moreover, from Fig. 7 we can see that method (a) cannot generate a smooth color transition sequence. Method (b) can generate a smooth color transition, however a lack of drastic color change makes it surreal. Better results may be obtained if we enlarge the time-lapse video dataset and include more scenes. While method (c) generates reasonable color transitions overall, it suffers from severe local artifacts (i.e. the sky at night, blue regions on the building, etc.). Our method (d) can both keep the fine-grained details in the image and obtain smooth sky color transitions.

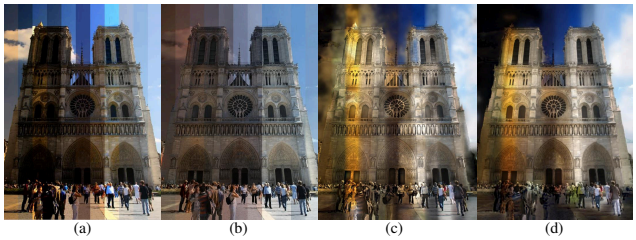

Figure 7. Comparative results for baseline color transfer methods. Column (d) is generated by our color transfer method, refer to the text for specification of baselines.

Qualitative Results. The generality and robustness of our approach is highlighted by applying our method to several image collections as shown in Fig. 8. Challenging appearance variations, such as drastic texture appearance changes (i.e. Berliner Dom), are addressed by leveraging the spatial composition similarity among images. Note that while our method relies on local homography-based structure transfer, deviations from non-planar scene structure (i.e. Mt. Rushmore) are mitigated by SIFT Flow refinement.

Quantitative Discussions. In the experiments, we observe a change in the color and smoothness in the color-transferred image by tuning the regularization factor γ. To make a convincing conclusion how γ influences the quality of the final images, we devise two metrics to quantitatively evaluate smoothness and color change. The first is a *smoothness ratio*, where we compute a sum of the image's horizontal gradients near the stripe boundaries and denote it as V_s. For the original mosaic M, this value is the largest, since no smoothing is applied. We then compute the smoothness ratio for every image as $V_s^{\gamma_i}/V_s^M$, where $V_s^{\gamma_i}$ is the smoothness of the γ_i-modified image, and V_s^M is the smoothness of the original mosaic. To describe a change in the color, we measure *color deviation* as the color histogram difference of the original mosaic and γ_i-modified image in Euclidean space. As we can see in Fig. 10, when the value of γ increases, the image is overall smoother, but it contains higher color deviation (i.e. notice the top left corner of the coliseum, where the red pattern fades away with increasing γ). In Fig. 11 we show the plots for the smoothness

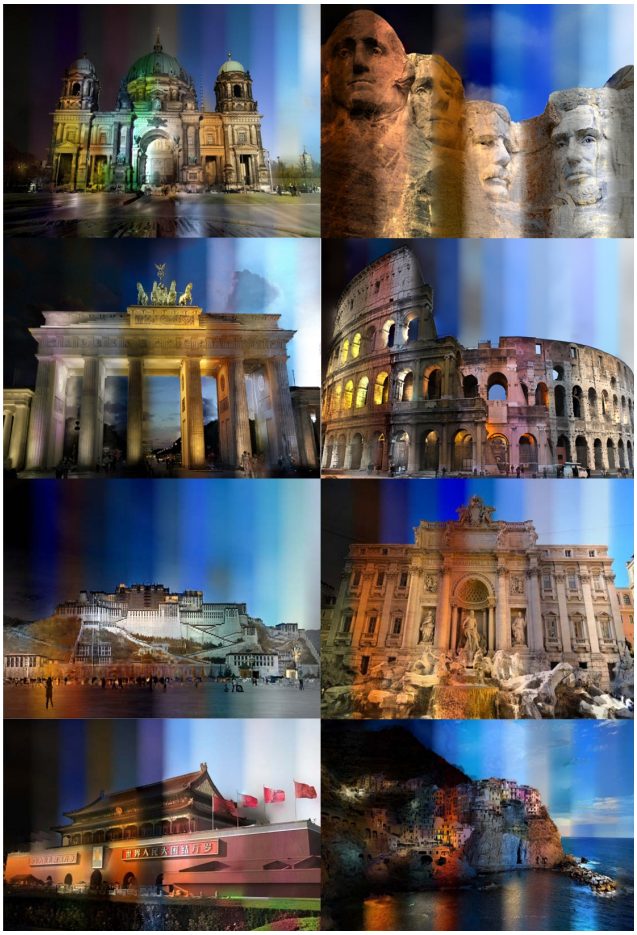

Figure 8. Illumination mosaics for eight downloaded datasets.

Figure 9. Failed cases for our method. Artifacts appear mainly on the domes and round facades which deviate from planar surfaces.

Figure 10. Color-transferred images with different γ, (left) $\gamma = 0.008$, and (right) $\gamma = 0.08$.

ratio and color deviation as γ increases. With increasing γ, the smoothness ratio keeps decreasing, i.e. the transition

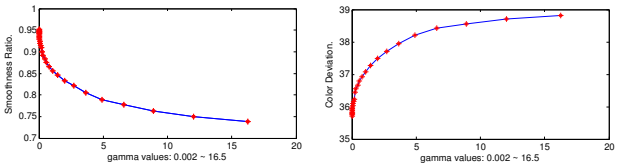

Figure 11. The effects of γ on the final mosaic: (left) smoothness ratio, and (right) color deviation.

is smoother, and the trend is to converge to a value that is equal to V_s^{Ref}/V_s^M, where V_s^{Ref} is the smoothness of the input image. The color deviation will also converge if γ goes to infinity, since the color transfer will be dominated by global linear matrix $\mathbf{G}$ (as shown in Eq. 4). One interesting thing to point out is when γ continues to decrease, the color-transferred image will contain increasingly many artifacts as without the regularization term, Eq. 4 the estimation will not be stable.

5. Conclusions

We propose a robust data-driven framework to automatically generate illumination mosaics of landmark scenes from Internet photo collections. Current limitations of our method are as follows: First, the length of the generated image sequence is uncontrolled, as it is mainly influenced by the size and internal distribution of the image dataset. Accordingly, results are sensitive to the image set homogeneity and redundancy w.r.t. the selected reference image, i.e. more densely sampled viewpoints will tend to generate better results. Since our method relies on a homography chain to align the images, it works reliably on scenes with mostly planar regions. For scenes with facades having multiple depths, if the misalignment can't be mitigated by SIFT Flow, artifacts will be produced (i.e. Fig. 9). Conversely, this characteristic may be leveraged to automatically discover viewpoints within the photo collection from which to generate the illumination mosaic. Second, since the color transfer method models the color transformation as an linear transformation, the scope of realistically reproducible appearance variation is intrinsically constrained. In the experiments, larger γ value might introduce some blurriness at the expense of a better transition.

Future work along these lines includes ascertaining a more global characterization of the appearance space by means of large-scale manifold learning techniques to increase the generality of the transitions modeled by our approach. Moreover, the generalization of our framework in order to model and represent the appearance variability depicted in video collections is yet another of our goals.

Acknowledgement This research is supported by NSF grants IIS-1349074 and CNS-1405847. We thank Jared Heinly for his help in proofreading and editing this paper.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, and S. Seitz. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] A. AGARWALA, M. AGRAWALA, M. COHEN, and D. SALESIN. Photographing long scenes with multiviewpoint panoramas. *Proceedings of SIGGRAPH*, page 853, 2006.

[3] A. Agarwala, A. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive Digital Photomontage. *Proceedings of SIGGRAPH*, 2004.

[4] D. Akers, F. Losasso, J. Klingner, M. Agrawala, J. Rick, and P. Hanrahan. Conveying Shape and Features with Image-Based Relighting. *IEEE Visualization*, 2003.

[5] S. Avidan and S. Shashua. Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. *PAMI*, 2010.

[6] T. Basha, Y. Moses, and S. Avidan. Photo Sequencing. *Proceedings of ECCV*, 2012.

[7] T. Basha, Y. Moses, and S. Avidan. Space-Time Tradeoffs in Photo Sequencing. *Proceedings of ICCV*, 2013.

[8] J. Canny. A computational approach to edge detection. i. *Pattern Anal. Mach. Intell.*, page 679, 1986.

[9] A. Chia, S. Zhuo, R. Kumar, Y. Tai, S. Cho, P. Tan, and S. Lin. Semantic Colorization with Internet Images. *Proceedings of ACM SIGGRAPH ASIA*, 30(6):156–156, 2011.

[10] J.-M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a Cloudless Day. *Proceedings of ECCV*, 2010.

[11] J. Hays and A. Efros. Scene Completion Using Millions of Photographs. *Proceedings of SIGGRAPH*, 2007.

[12] J. Heinly, J. Schnberger, E. Dunn, and J.-M. Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). *Proceedings of CVPR*, 2015.

[13] D. Hoiem, A. Efros, and M. Hebert. Geometric Context from a Single Image. *ICCV*, 2005.

[14] N. Jacobs, N. Roman, and R. Pless. Consistent Temporal Variations in Many Outdoor Scenes. *Proceedings of CVPR*, 2007.

[15] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. Seitz. Exploring Photobios. *ACM Transactions on Graphics*, 30, 2011.

[16] G. Kim, E. Xing, and A. Torralba. Modeling and Analysis of Dynamic Behaviors of Web Image Collections. *Proceedings of ECCV*, 2010.

[17] P. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(149), 2014.

[18] K. Lee, S. Luo, and B. Chen. Rephotography Using Image Collections. *Pacific Graphics*, 2000.

[19] A. Levin, A. Zomet, S. Peleg, and Y. Weiss. Seamless Image Stitching in the Gradient Domain. *Proceedings of ECCV*, 2004.

[20] C. Liu, Y. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5), 2011.

[21] R. Martin-Brualla, D. Gallup, and S. M. Seitz. Time-lapse Mining from Internet Photos. *Proc. of SIGGRAPH*, 2015.

[22] A. Oliva. Gist of the scene. *Neurobiology of attention*, 696:251, 2005.

[23] F. Palermo, J. Hays, and A. Efros. Dating Historical Color Images. *Proceedings of ECCV*, 2012.

[24] T. Pouli and E. Reinhard. Progressive color transfer for images of arbitrary dynamic range. *Computers & Graphics*, 35(1):6780, 2011.

[25] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *Computer Graphics and Application*, 21(5):3441, 2001.

[26] G. Schindler and F. Dellaert. Inferring Temporal Order of Images From 3D Structure. *Proceedings of CVPR*, 2007.

[27] G. Schindler and F. Dellaert. Probabilistic Temporal Inference on Reconstructed 3D Scenes. *Proceedings of CVPR*, 2010.

[28] J. Schnberger, F. Radenovic, O. Chum, and J.-M. Frahm. From Single Image Query to Detailed 3D Reconstruction). *Proceedings of CVPR*, 2015.

[29] S. Seitz and J. Kim. The Space of All Stereo Images. *IJCV, Marr Prize Special Issue*, 48:21, 2002.

[30] Y. Shih, S. Paris, F. Durand, and W. Freeman. Data-driven Hallucination for Different Times of Day from a Single Outdoor Photo. *ACM Transactions on Graphics (TOG)*, 32(6), 2013.

[31] N. Snavely, R. Garg, S. Seitz, and R. Szeliski. Finding Paths through the World's Photos. *ACM Transactions on Graphics*, 27(3):11–21, 2008.

[32] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[33] L. Tao, L. Yuan, and J. Sun. SkyFinder: Attribute-based Sky Image Search. *ACM Transactions on Graphics*, 28(4), 2009.

[34] J. Wang, X. Tong, S. Lin, M. Pan, C. Wang, H. Bao, B. Guo, and H. Shum. Appearance Manifolds for Modeling Time-Variant Appearance of Materials. *ACM Transactions on Graphics*, 25(4), 2006.

[35] X. Xu, L. Wan, X. Liu, T. Wong, L. Wang, and C. Leung. Animating Animal Motion from Still. *SIGGRAPH Asia*, 2008.

[36] Y. Yen. An algorithm for finding shortest routes from all source nodes to a given destination in general networks. *Quarterly of Applied Mathematics*, 27:526, 1970.

[37] L. Zelnik-Manor and P. Perona. Automating Joiners. *NPAR*, 2007.

[38] F. Zhang and F. Liu. Parallax-tolerant Image Stitching. *CVPR*, 2014.